SLICK – New Methods for Protein-Carbohydrate Docking

Dissertation
der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. Andreas Kerzmann

Tübingen
2006

Datum des Kolloquiums:	16. Mai 2007
Dekan:	Professor Dr. Michael Diehl
1. Berichterstatter:	Professor Dr. Oliver Kohlbacher, Tübingen
2. Berichterstatter:	Professor Dr. Volkhard Helms, Saarbrücken

Abstract In drug design, the prediction of the binding mode of drug-like substances in the binding site of a receptor is one of the most important tools. Over the last years, many different approaches have been developed and applied successfully to a number of substances. However, some classes of substances are still hard to predict with docking techniques. One very important class of such molecular complexes is the domain of carbohydrates binding to proteins.

This thesis introduces an approach to protein-carbohydrate docking. The method is based on thorough analysis of those interactions, which are important for protein-carbohydrate complexes. Computational models of these effects are employed for creating a scoring and an energy function for this kind of complexes. By integrating these functions into a docking programme, the effectiveness of this new approach is proven.

After a brief introduction into the problem addressed by this theses, the biochemical background of protein-carbohydrate complexation is explained. The main characteristics of such complexes, which are the basis for the development of the functions, are detailed. Furthermore, physical models for the atomic interactions that are deemed important for protein-carbohydrate complexes are explained. Based on these interactions, the SLICK package, consisting of an energy function and a scoring function is developed. The effectiveness of both functions is proven. Eventually, their effectiveness in protein-carbohydrate docking is shown by integrating SLICK into a molecular docking programme.

Zusammenfassung In der Wirkstoffentwicklung ist das so genannte *Docking*, also die Vorhersage des Bindungsmodus wirkstoffähnlicher Moleküle an einen Rezeptor, ein wichtiges Werkzeug. In den letzten Jahren wurde eine Vielzahl von Methoden und Ansätzen entwickelt und für eine Reihe von Wirkstoffklassen erfolgreich eingesetzt. Jedoch gibt es immer noch Bereiche, in denen Dockingverfahren scheitern. Eine wichtige Klasse von Molekülkomplexen, die mit Dockingverfahren immer noch schwer zu handhaben sind, stellen zuckerbindende Proteine dar.

Die vorliegende Arbeit stellt einen Ansatz vor, der das Docking von Zuckern an Proteine ermöglicht. Dieses Verfahren basiert auf einer genauen Analyse der für Protein-Zucker-Komplexe wichtigen Wechselwirkungen und der Modellierung jener Eigenschaften in Form einer Energie- und einer Scoringfunktion. Es wird gezeigt, dass durch die Einbettung dieser beiden Funktionen in ein Dockingprogramm eine automatische Vorhersage des Bindungsmodus von Protein-Zucker-Komplexen möglich wird.

Nach einer knappen Einführung in das zu behandelnde Problem beleuchtet diese Arbeit den biochemischen Hintergrund der Bindung von Zuckern an Proteine. Dabei werden die Hauptmerkmale erläutert, die für die Entwicklung eines Dockingverfahrens von Bedeutung sind. Desweiteren werden physikalische Modelle für die für Protein-Zucker-Komplexe charakteristischen atomaren Wechselwirkungen behandelt. Aufbauend auf diesen Modellen werden Scoring- und Energiefunktion entwickelt, die zusammen das Paket SLICK bilden. Es wird gezeigt, dass die Scoringfunktion in der Lage ist, aus einer großen Menge von Dockingkandidaten viel versprechende Strukturen herauszufiltern. Anschließend wird die Vorhersagequalität der Energiefunktion statistisch anhand verfügbarer experimenteller Daten analysiert. Schließlich wird durch die Integration von SLICK in ein Dockingprogramm nachgewiesen, dass mit Hilfe der in dieser Arbeit entwickelten Verfahren das bisher schwierige Problem des Dockings von Zuckern an Proteine lösbar wird.

Acknowledgements

This thesis would not have been possible without the support and effort of many people. First of all, I would like to thank Prof. Dr. Oliver Kohlbacher, who sparked my interest in structural bioinformatics when he himself was still doing his Ph. D. thesis. His support and amicable guidance made the last six years a very important experience. I also want to thank Dr. Dirk Neumann, who often straightened out my rather large gaps in lectinology and sugar chemistry and provided encouraging support in difficult situations. For friendship and advice I want to thank all the members of our group, especially Muriel Quenzer, Dr. Annette Höglund, Dr. Pierre Dönnes and Emmy.

I also want to thank Prof. Dr. Hans-Peter Lenhof, who made this thesis possible in the first place by giving me the opportunity to start my work in his group under the supervision of Oliver Kohlbacher. For fruitful discussions on electrostatics and table soccer, I want to thank Dr. Andreas Hildebrandt. Special thanks go to Jan Fuhrmann, who provided me with his docking code that represents the basis for my sugar docking programme.

Many students have contributed to this thesis through their efforts in the student projects I supervised. Florian Kirchner implemented the binding site finder; Serhat Saydam confirmed my previous results regarding electrostatics models; Adrian Schröder generated first results in integrating CH$\cdots\pi$ into Jan's docking code; Hedwig Zimmerer implemented a first approach to the CH$\cdots\pi$ model and validated its correctness. Thank you all for providing your work and your opinions.

Finally I want to thank my proof-readers Dr. Pierre Dönnes, Martin Palige, and Nora Toussaint for giving valuable feedback and thereby improving this thesis significantly.

Contents

1. Introduction

Today, progress in drug discovery is almost unthinkable without the help of computational models. Many achievements of the last years would not have been possible without sophisticated computer programmes, assisting pharmacists and chemists in the laboratories when designing new drugs and improving existing drugs.

There are many problems to be addressed when developing or improving drugs. The active agents have to cross biological barriers, get to the location where their effect is needed and then interact strongly with a target molecule. At the same time, drugs have to be cheap and achieve all this without harmful side-effects. Thus, the development of drugs is not a trivial challenge.

Drug discovery always requires experiments. With constantly improving experimental methods, the amount of data on molecular processes and their effects on living organisms is perpetually increasing, giving new starting points for drugs development. Computational methods can facilitate the development of drugs by providing means of simulating the real processes, analysing very large amounts of data in short time and predicting the properties of new chemical entities. But computational models cannot be devised without knowledge gained from experiments, which represent he basis for understanding how drugs interact with other molecules.

Drugs take effect on the patient's metabolism by binding tightly to a target molecule, often referred to as a *receptor*. In drug design, the first step is to find a target molecule which is responsible for an illness. As soon as the target is defined, many different chemicals are tested against that receptor, *i. e.* the binding affinity of these molecules to the receptor is determined. If a molecule binds well, it qualifies for further investigation. However, a binding molecule, a so-called *lead*, is not necessarily a potent drug. In fact, most of the substances that are found to be binding will do so rather weakly. Some might even be toxic compounds and therefore cannot be used as a drug at all. Consequently, the goal is to pick a few promising binding substances and use them as starting points for developing a real drug by optimising their pharmacological properties. During this optimisation process, compounds derived from the lead molecule have to be synthesised and tested over and over again, until a molecule is developed that can be tried in animals and eventually in humans.

However, synthesising and testing new compounds in the laboratory is a very time-consuming and expensive task. For that reason, scientists try to use computational methods for predicting the binding affinity of a certain compound to a target molecule without actually synthesising it. The goal is to reduce the necessary laboratory work and focus on promising compounds. Doing so will dramatically reduce time and money needed during drug development. There are several ways of predicting binding affinities. Some methods focus on chemical or physical properties of the molecules, *e. g.* the number of hydrogen bond donors and acceptors available. These features can *e. g.* be used in machine learning approaches which might give reasonable results with respect to predicting the binding affinity. But apart from the binding affinity, the geometry of binding, *i. e.* the actual three-dimensional structure of the molecules in the bound state, is crucial for further developing the lead compound into a working drug. In this context, the geometry of a

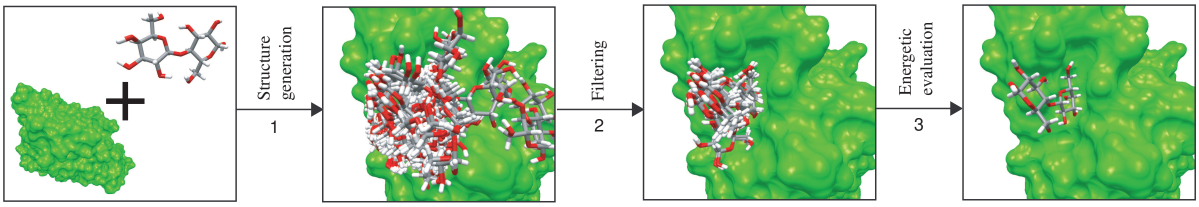


Figure 1.1.: The abstract docking scheme: Start from the spatial structures of receptor and ligand, (1) create a large number of putative docking candidates, (2) filter out bad approximations and (3) evaluate the remaining structures energetically.

molecule is called *conformation*.

The process of computationally predicting the binding conformation of two molecules and their binding affinity is called *docking*. In simple terms, the docking problem is the following:

> Given two molecules A and B, which are known to form a complex AB, and their conformations in the unbound state, compute the binding conformation and binding affinity of the complex.

How can this problem be solved? The *binding free energy* of a molecular system is the amount of energy that is liberated when two molecules form a complex. Nature always tries to reach states of minimal energy. Consequently, two molecules will only form a complex if it is energetically favourable to do so, *i. e.* if there is no additional energy necessary to form the complex and the binding free energy is actually liberated during the process. Thus, if we know how to calculate accurate energies of the system and use this knowledge for searching energetically minimal states, we will be able to predict approximations of the real complex. This idea is the basis for computational molecular docking. Docking programmes are roughly based on the following scheme (see Fig. 1.1):

Structure Generation Create a large number of putative complex structures, referred to as *candidate structures* or simply *candidates*.

Filtering Filter out candidates that are probably bad approximations of the real complex using a fast *scoring function*. Such a function connects three-dimensional structure with a number, the *score*. The better the score of a candidate is, the more probable it is a binding conformation. Such a function should be computationally inexpensive in order to score many conformations in little time.

Energetic evaluation Evaluate the remaining candidates energetically, *i. e.* compute the binding free energy of these candidates by means of an *energy function*. Such a function also connects structure with a number, but in this case, the actual binding free energy is calculated. As there are only few putative complexes left after filtering, this function can be computationally more intensive than a scoring function.

However, this scheme is only a very coarse view on what docking programmes really do. In most cases, during the process of structure generation, bad candidates will be filtered out immediately, depending on the strategy used to generate candidates.

Docking programmes can be classified by their purpose. There are *e. g.* docking programmes for protein-protein complexes or for small ligands binding to proteins. Additionally, docking programmes are often distinguished by their structure generating strategy. There are many different ways of creating putative complex structures, some of which will be discussed in this work. However, the important classification for the pharmacist is most probably the range of applicability of a docking programme in terms of chemical compounds predictable with a certain method. Protein-ligand docking programmes usually are designed for arbitrary ligands. While this might sound reasonable as it simplifies application of a method, it certainly is a very strong claim. As a matter of fact, docking programmes do have a range of ligands for which they produce satisfactory results. For others, they fail.

A prominent example of complexes that are notoriously hard to predict with general docking programmes is the realm of protein-carbohydrate complexes. Such complexes are very interesting from a pharmaceutical point of view. Protein-carbohydrate interactions are known to influence many biologically important processes. They are crucial to pathogen recognition, play various important roles in our immune system, and are directly connected to cancer diagnosis.

This thesis will introduce a docking method for protein-carbohydrate complexes. Up to now, docking carbohydrates to protein receptors is only possible in a very limited range of complexes. Thorough analysis of the literature available on docking of protein-carbohydrate complexes revealed that there exists no systematic docking method for this type of complexes except for the approaches by Coutinho and coworkers [1, 2, 3, 4] and the subsequent adaptation of the AutoDock method [5] presented by Laederach in 2003 [6]. All other attempts at docking carbohydrates to proteins were of very limited success. In most cases, the resulting structures had to be subjected to additional computationally intensive treatment like molecular mechanics simulations or other optimisation techniques in order to gain acceptable complex structures. Thus, the advantages of a docking method – speed and low cost – vanished with the increased amount of work and computing time necessary for gaining reasonable results.

The work of Coutinho and coworkers presented a system for docking carbohydrate oligomers into protein binding sites that was based on AutoDock. This system used a two-stage approach. First, the binding position of one sugar ring of the oligomer was spatially fixed based on prior knowledge. Second, the oligosugar was docked with the previously fixed ring being kept rigid throughout the docking process. Although the results were encouraging, the approach has an obvious disadvantage – prior knowledge is required. In drug design application, this knowledge is not necessarily given. Laederach improved Coutinho's work by introducing a recalibrated version of AutoDock [6] for docking carbohydrates into protein binding sites, which is not dependent on prior knowledge of the complex under consideration. However, Laederach's work does not take the rather specific nature of protein-carbohydrate interactions into account. The CH$\cdots\pi$ interaction, which is known to influence binding through ring-stacking, is completely missing and solvation effects can only be covered as far as AutoDock already permits it. A thorough treatment of polar and nonpolar solvation contributions is not possible.

Additionally, there are empirical methods designed for the optimisation of the molecular geometries of carbohydrates, but these approaches are hardly applicable for the calculation of binding energies, let alone binding geometries, because they also neglect interactions which have proven important in protein-sugar complexation. Clearly, a new method for docking protein-carbohydrate complexes is necessary if these complexes are to be examined for drug design purposes.

The goal of this work is to develop and validate a scoring function and an energy function that can be used in automated docking of protein-carbohydrate complexes. These functions will be specifically designed for this problem, based on thorough analysis of structural data from publicly available databases and energetic properties that were reported in literature. Moreover, these functions will be incorporated into a docking method that will be applied to a rather large set of known protein-carbohydrate complexes in order to prove their effectiveness. The resulting system will comprise the first design of a docking method for protein-carbohydrate complexes that is not based on mere reparameterisation of an existing docking method.

The method introduced in this thesis is called SLICK, which is an acronym for **S**ugar-**L**ectin **I**nteractions and Do**CK**ing. It is a package consisting of SLICK/score, a scoring function for docking purposes, and SLICK/energy, an empirical energy function. SLICK introduces a new term for so-called CH$\cdots\pi$ interactions, which are not covered by energy functions used in docking programmes so far. These interactions, which will be discussed in Chapter 2, play an important role in protein-carbohydrate binding and are very important in docking. Additionally, SLICK/energy considers so-called solvation effects with state-of-the-art computational models, which is also uncommon in energy functions of docking programmes. Solvation effects arise from interactions of the molecules with their surrounding solvent, in which biological processes take place.

Another very important part of SLICK is the consideration of hydrogen bonds. These bonds strongly influence protein-carbohydrate binding. Van der Waals interactions are calculated with a softened form of the well-known Lennard-Jones potential. Softening this potential makes calculations much less susceptible to inaccuracies in the structural data. Such inaccuracies are often introduced in docking experiments.

SLICK/energy is an energy function that was designed to predict interaction energies. This goal is achieved by including two different types of computational approaches. On the one hand, SLICK/energy employs models for calculating energies of whole molecular systems and computes interaction energies by subtracting two different states of a system. For example, the van der Waals component calculates the energy of the system in the bound state and in the unbound state. The energy difference is then the interaction energy. On the other hand, SLICK embraces models that calculate interactions directly. An example is the CH$\cdots\pi$ contribution, which is based on geometric considerations and does not yield interaction energy but a score that discriminates good from bad interactions. Merging these two approaches together results in a function that is able to predict intermolecular interactions very well but cannot be used in optimisation or molecular dynamics calculations. Although this is a drawback compared to energy functions of molecular mechanics force fields, the model provides large flexibility in the choice of energy contributions for a specific problem and is thus not limited to one domain.

Another important difference between force fields and SLICK is the parameterisation flexibility of the model. Usually, an energy function has its own parameterisation that is based on fitting. In the case of SLICK, another approach was chosen in order to avoid a full reparameterisation of all involved models and to gain accuracy. In principle, every energy contribution in SLICK can use its own set of parameters. For example, the solvation component needs two different sets of radii, one for the polar part and one for the calculation of nonpolar effects. Using one parameter set for both contributions reduces the prediction quality significantly, which is a result of the very different theoretical background of both models. Thus, the flexibility of the SLICK approach permits the highest possible accuracy in each involved contribution without the need of merging different and possibly incompatible parameter sets into one.

The robustness and prediction quality of SLICK/energy is thoroughly evaluated by means of statistical assessment. Both SLICK/score and SLICK/energy are incorporated into the docking programme BALLDock [7], yielding a protein-carbohydrate docking tool called BALLDock/SLICK. This tool is validated on an extensive set of high-quality experimental data and the results are compared to existing docking methods.

The results of this thesis are very encouraging. With SLICK/score, 13 top-scored structures out of 20 complexes in the validation set are binding conformations. In four other cases a correct pose was under the ten highest scoring structures. In only three cases, the binding conformation was not determined correctly. SLICK/energy achieves very low error rates in statistical assessment of the calibration with an average absolute error of as low as 1.3 kJ/mol and a maximum absolute error of only 3.1 kJ/mol. In validation, it achieves an average absolute error of 2.1 kJ/mol. These numbers clearly indicate that SLICK is fit for being used in protein-carbohydrate docking, although further improvement is most probably possible. By integrating SLICK into BALLDock, high docking accuracy could be achieved. On the calibration set, BALLDock/SLICK outperformed two existing general docking methods in both ranking of docking candidates and accuracy of the energy predictions. On a larger set of known protein-carbohydrate complexes, BALLDock/SLICK also produced better results than the general docking programme it was compared with.

Developing the method for protein-carbohydrate docking presented in this work requires expertise from many different fields, like molecular biology, physics and bioinformatics. The necessary chemical and biological background and some additional information on the role of lectins and sugars in pharmacy will be discussed in Chapter 2. Most importantly, this chapter will detail the peculiarities of the lectin-sugar binding process. Knowing these features that make protein-carbohydrate complexes special is the basis for finding computational models necessary for calculating energies and predicting binding geometries. These models will be explained in Chapter 3, along with basic principles of the underlying physics and the algorithmic approach to molecular docking. However, the models alone are not sufficient. Predicting binding energies with an empirical method like the one described in this thesis is depending on experimental data on which functions are calibrated. The data sets and details on the preparation of available data are presented in Chapter 4.

Having computational models and experimental data ready, the lectin-sugar docking method can be assembled. Chapter 5 will explain the approaches to scoring and energetic evaluation of protein-carbohydrate complexes, which were developed and implemented for this work. This chapter also covers the integration of the resulting functions into a programme for molecular docking and the analysis of the results gained from validating the method. Chapter 6 will then review and critically assess the findings and conclude this thesis with a glimpse at the future of protein-sugar docking based on SLICK.

2. Biochemical Background

Sugar-binding proteins and carbohydrates have been getting more and more attention in pharmaceutical research over the last few years. One of the most prominent examples of sugar based drugs being in headline news is the publication of the Seeberger group in 2005, in which they describe the synthesis of an oligomeric sugar which could be the basis for a vaccine against Anthrax [8]. Protein-carbohydrate interactions are of great importance for many biological processes and thus are very interesting from a pharmaceutical point of view. Among general sugar-binding proteins, there is one class of special interest, the so-called *lectins*. These proteins bind sugars without changing them, *i. e.* they have no enzymatic ability. Lectins have many important properties and functions. Therefore, this study focuses on lectins.

This chapter gives an overview of existing approaches exploiting protein-carbohydrate interactions in pharmaceutical applications and models. Moreover, a short introduction to the chemistry and structural features of carbohydrates and lectins will be given. More importantly, the peculiarities of binding interactions in protein-carbohydrate complexes are reported in detail, which will give the basis for understanding the strategy chosen in creating the functions of the SLICK package.

2.1. Carbohydrates

Sugars are very important biomolecules. Their function as energy storage and structural element *e. g.* in cellulose has been known for years, while their contribution to biology as active components has been neglected for a long time. Fortunately, glycobiology gained more and more attention over the past years [9], thus giving the opportunity to improve our knowledge about sugars and their role in biological processes. Carbohydrates take part in cell recognition, apoptosis, fertilisation, growth control, tumor spread and many more biologically relevant processes. Consequently, sugars are very interesting from a pharmaceutical point of view because they can be utilised as drug targets and drugs alike. Besides their aforementioned possible use as vaccines against Anthrax [8], there are already a number of applications based on sugars. In mouse models, Anti-tumor treatment is enhanced [10] by sugars and a method for enhancing cancer treatment by coating particles with sugar was just recently reported [11]. Other fields of application include treatment of Gaucher's disease or the development of antibiotics, for which sugar-based drugs are already on the market [12].

Carbohydrates are not only pharmaceutically relevant, they are are also information carriers with huge capacity. Sugar polymers are non-linear compounds in contrast to proteins and DNA, which both have a secondary and tertiary structure, but first of all are linear chains of amino acids or nucleotides, respectively. With sugars it is possible to create large branched molecules from a rather small set of building blocks. These large polycarbohydrates, which can *e. g.* be found on the surface of cells, are called *glycans* and have been identified in numerous articles as

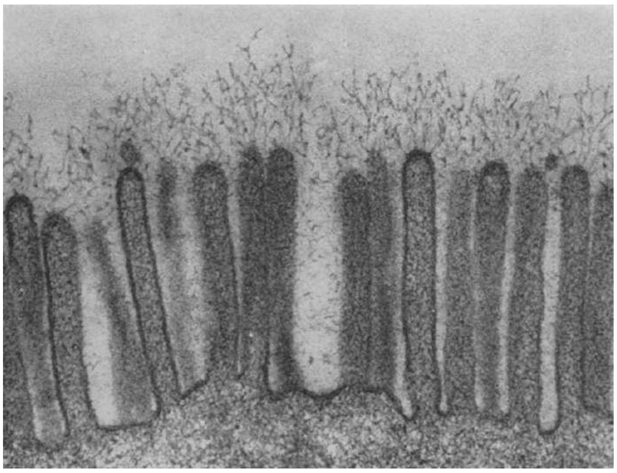


Figure 2.1.: The glycocalyx of a cell [15].

the third large group of biomolecules carrying information [13, 14]. This led to coining the term *glycome* as the third large source of information in molecular biology.

The glycans on the cell surface build the so-called *glycocalyx* (see Fig. 2.1) coating the whole cell with a very specific composition of sugars. It is known that this composition of sugars can be seen as a kind of fingerprint discriminating cell types, which is a basis for cell recognition and thus can be exploited for targeting specific cells.

In order to exploit sugars in pharmaceutical applications, it is necessary to understand the structure of polycarbohydrates. From a simplistic point of view, sugar monomers are aliphatic carbon rings. Each ring carbon is supporting a hydrogen and a hydroxyl group. The rings derive from linear carbohydrate chains of four or more carbons, one of which is supporting a functional group (aldehyde or ketone). By connecting the functional group with an OH group, the linear chain spontaneously forms rings in solution by building hemiacetals and hemiketals. Figure 2.2 shows an example for a sugar ring evolving from its linear form. Most simple sugars are *pentoses* (five carbons) and *hexoses* (six carbons). They are further distinguished by the number of ring carbons. Rings of five carbons and one oxygen are called *pyranoses*. Rings with only four carbons and an oxygen are called *furanoses*. While the basic structure of sugars is relatively simple, the number of isomers stemming from one configuration is rather large. The main reason for that is the large number of asymmetric carbons in the molecule leading to enantiomerism and diastereoisomerism. Additionally, the conformation of the sugar ring leads to two different isomeric forms, one of which is energetically more stable.

Sugar monomers can contain substituents other than hydroxyl groups, like acetyl groups or aromatic rings, linked to hydroxyl oxygens or the aliphatic ring carbons, which makes the chemistry very complicated. However, the sugar ligands investigated in this study support only methyl and N-acetyl groups, if any. Sugars with aromatic ring systems would demand an additional model covering aromatic ring interactions, which is not available in SLICK so far.

Sugar molecules are not limited to monomers consisting of only one sugar ring. Dimers, *i. e.* molecules that are built from two sugar rings, can be created by building acetal bonds between two hydroxyl groups of two different monomers. The two rings are then linked over one oxygen which has a freely rotatable bond to each of the rings (see Fig. 2.3). The torsion angles of these

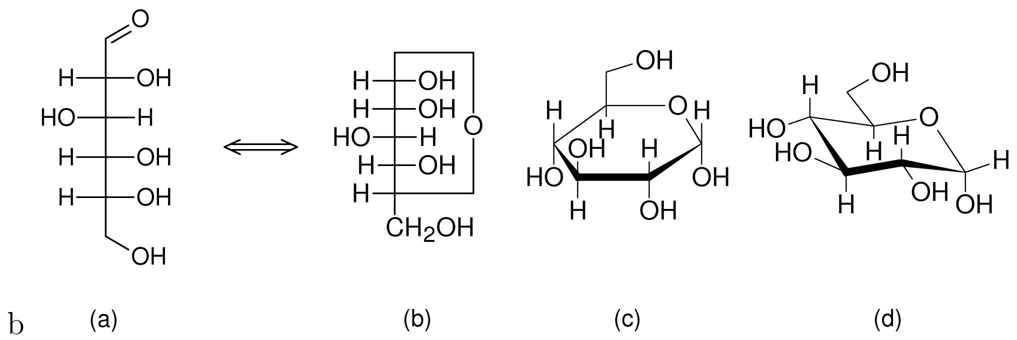


Figure 2.2.: Linear (a) and pyranosylic form (b) of a pentose (glucose). The ring (c) usually occurs in conformation (d). Note the large number of hydroxyl groups, which are anchor points for dimerisation.

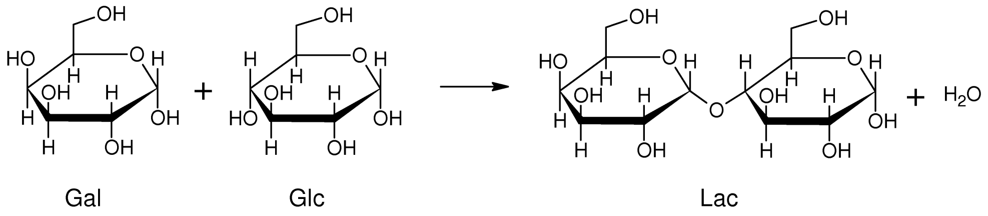


Figure 2.3.: Building a glycosidic bond between two carbohydrate monomers.

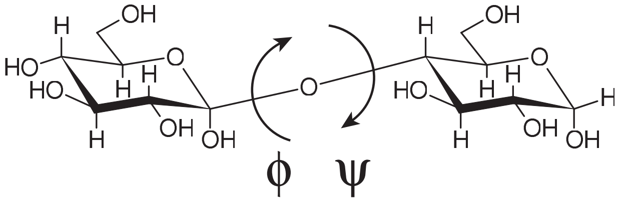


Figure 2.4.: Torsion angles ϕ and ψ of the glycosidic bond.

two rotatable bonds are usually denoted with the greek letters ϕ and ψ (cf. Fig. 2.4). The link connecting two sugar monomers over one oxygen is also called the *glycosidic bond* and is very flexible.

The extent of this flexibility was analysed computationally by scanning the conformational space of a sugar dimer in solution and calculating energies [16] for every investigated combination of torsion angles. The resulting Ramachandran plot is given in Fig. 2.5. Furthermore, experimental NMR studies have been conducted [17] in order to analyse the flexibility of such dimers. These studies lead to the conclusion that contrast to *e. g.* peptides the minima of the energy plot are rather shallow and very wide. This means that the energy of the sugar does not change much if the combination of ϕ and ψ remains within the energy valley. If energy differences are small, all conformations within the energy valley are approximately equally probable. Thus there is no single favourite conformation of the sugar dimer, as would be in the peptide case, where energy valleys tend to be much narrower and deeper. In the context of molecular docking, this behaviour leads to a very large conformational space that has to be traversed during a docking experiment,

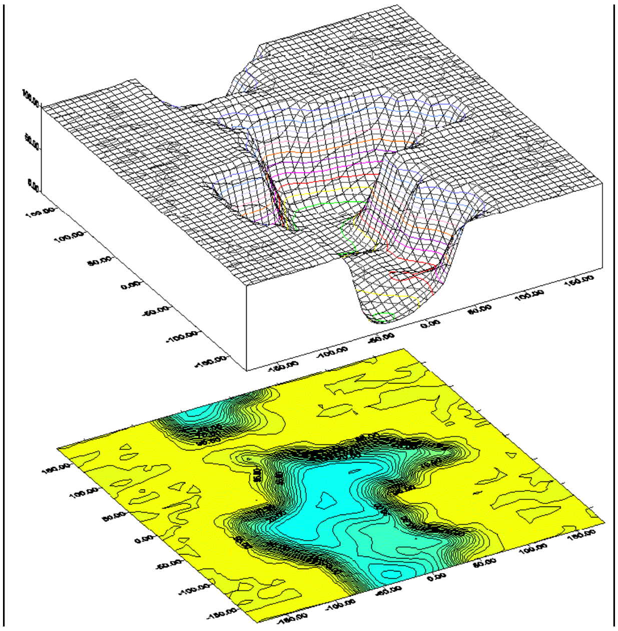


Figure 2.5.: Torsion angle energy surface of a carbohydrate dimer [16]. The energy is drawn as a surface with respect to the ϕ and ψ torsion angles. Note the rather shallow energy valley, indicating that carbohydrates tend to rotate freely in the unbound state.

because one cannot reduce the set of possible angles to a small number of energetic favourable ones.

Since a sugar monomer supports at least two hydroxyl groups, long chains of sugar can be built. If the number of building blocks is rather small, such chains are called *oligosugars* or *oligosaccharides*. If many single sugar rings build one molecule, the resulting compound is a *polysaccharide*. Oligosaccharides are not necessarily linear chains. As there are so many hydroxyl groups in one sugar monomer the number of possible bond partners is rather large. In pyranoses derived from hexoses theoretically there are five positions for glycosidic bonds possible, although sterical hindrance will not allow all of these hydroxyl groups to be linked to other monomers at the same time. Nevertheless there is the possibility to create branched oligomers by binding two sugar monomers to two different hydroxyl groups of the terminal ring of a linear oligomer (see Fig. 2.6).

This non-linearity of carbohydrate polymers gives rise to a huge number of possible combinations of monomers in one molecule. When comparing the three big classes of biomolecules carrying information, the information content of sugar polymers clearly outperforms the possible content of DNA strands or proteins [18] (see Tab. 2.1).

From the physicochemical perspective, sugars are also remarkable molecules. Carbohydrates have a polar surface because of the polar nature of the many hydroxyl groups. This leads to high hydrophilicity. Furthermore, the polar surface of carbohydrates causes an anisotropic rearrangement of water molecules in solution [19] which in turn leads to entropic and enthalpic contributions to the solvation energy of carbohydrates. This behaviour makes the prediction of solvation effects very difficult because most approaches for determining these effects treat the

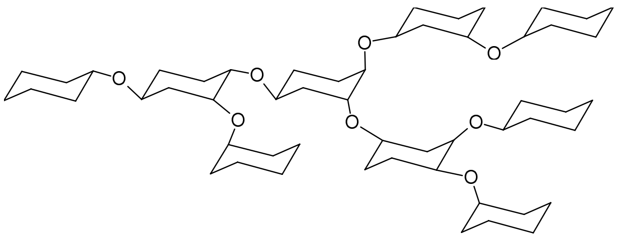


Figure 2.6.: Building branched structures from several monomeric sugars. In principle, every sugar ring can bind up to five other sugar rings.

Biomolecule	Alphabet size	Number of possible hexamers
DNA	4	$4 \cdot 10^3$
Proteins	20	$6 \cdot 10^7$
Carbohydrates	≈ 20	$\approx 10^{15}$

Table 2.1.: Comparison of possible information content of different kinds of biomolecules [18].

solvent as a continuum, which clearly does not work for an anisotropic solvent structure. But solvent effects are known to influence the binding behaviour. Therefore, a sophisticated treatment of solvation seems necessary. Additionally, polar groups imply strong electrostatic interactions with a putative binding partner. Hence, the calculation of the electrostatic interactions are very important. Moreover, hydroxyl groups are known hydrogen bond donors and hydrogen bonds are known to influence binding strongly. Because the hydroxyl groups of sugars are freely rotatable, a sugar can "adapt" its hydrogen bond donors to the structural requirements of a binding site if there are acceptors present. Consequently, hydrogen bonding needs special attention, too.

In summary, carbohydrates have many interesting properties. Sugar oligomers are information carriers with a very high information density, although the monomeric building blocks are rather simple in structure. The flexibility of the bond connecting two monomers leads to a huge conformational space making sugars a rather hard problem for molecular docking. Additionally, the highly polar surface leads to strong intermolecular interactions when binding to another molecule and to complicated solvation effects that might influence the binding behaviour. A prediction model will have to address these interactions along with hydrogen bonding, which also influences the binding noticeably.

2.2. Lectins

Like carbohydrates, lectins are also important biomolecules and can be found in virtually every organism. Lectins have been found in many plants but also in animals and humans. In animal metabolism and immune systems they drive many important biological processes like cell aggregation [20] and cell differentiation. Lectins are also involved in pathogen recognition [21] and other immunologically important tasks like inducing maturation of dendritic cells [22]. One group of lectins, the so-called galectins, have been used as helping agents in tumor suppression [23] and are being investigated as a diagnostic element in detection of breast cancer [24, 25].

Their importance and their functions are based on the fact that lectins bind carbohydrates very

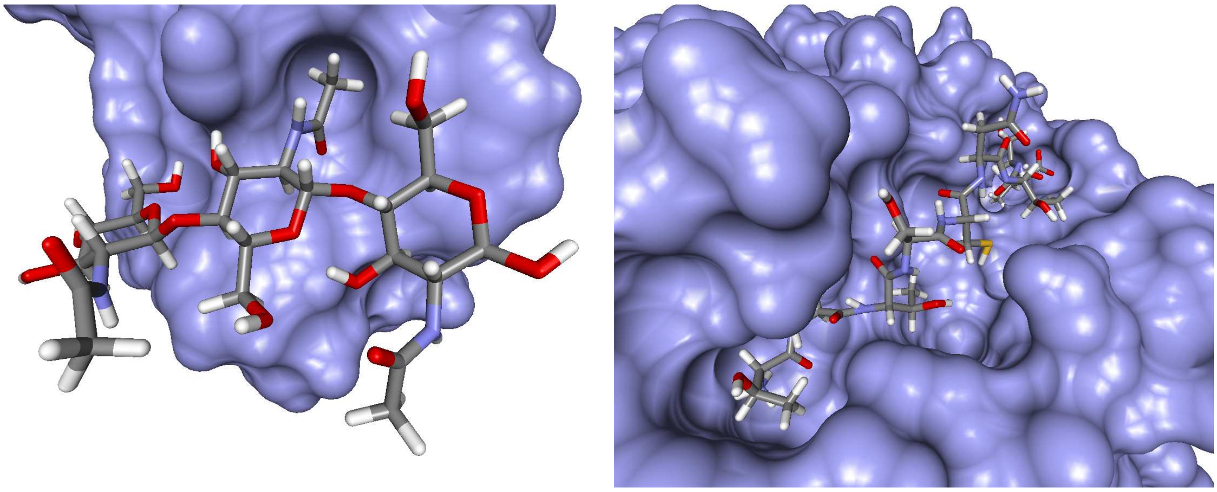


Figure 2.7.: Left: binding site of UDA. Note that the ligand is lying on the surface rather than binding into a binding groove. Right: binding site of an MHC complex. The bound peptide is deeply buried in the binding site.

specifically [26]. This specificity turns lectins into decoding engines for the information stored in large glycans and thus into very interesting targeting devices when it comes to drug targeting and drug delivery. They have repeatedly been employed as mediating or functionalising components in drug targeting [27, 28] and drug delivery [29] systems. Lectin conjugates were also used in directed gene transfer [30]. Furthermore, lectins can be drug targets [31] as well.

Lectins are more or less ordinary proteins, linear chains of amino acids, which assume a certain geometry in three-dimensional space. This conformation is crucial for their function. Thus lectins rely on remaining in the same shape resulting in a rather rigid structure. This is common for proteins, which are all functionally dependent on their three-dimensional structure. For a more detailed introduction to proteins and structural bioinformatics in general see *e. g.* [32].

Lectin binding sites differ from most ordinary protein binding sites in shape. While binding sites for peptides or other ligands often are deep binding grooves strictly defining the shape of the binding partner, lectin binding sites are rather shallow [33]. Carbohydrates bind onto a “binding surface” rather than into a binding groove. Figure 2.7 (left) shows the shallow binding site of *Urtica dioica* agglutinin (UDA) binding to a trimer of N-acetyl-glucosamine (GlcNAc)[1]. Apparently the binding mode of the ligand must be guided by other interactions that define the geometry of binding rather than steric hindrance defining the binding shape. Figure 2.7 (right) shows an example of a nonapeptide binding to a human major histocompatibility complex (MHC) molecule. Obviously the shape of the binding pocket strictly defines the conformation of the binding peptide.

For developing a thorough energy function for lectin-carbohydrate interactions we had to decide on the data basis for calibrating and validating the approach. We chose to focus on plant lectins because these lectins are very well researched, both structurally and energetically. Moreover,

[1] Abbreviations of lectin and carbohydrate names are listed in Appendix D

plant lectins are generally very stable against heat and digestion [34, 35]. These features make plant lectins valuable for pharmaceutical applications where pro-drugs or functionalised drug carrier systems have to survive the gastro-intestinal tract while retaining their function.

Additionally, animal lectins often contain glycosylated or phosphorylated amino acids, which makes parametrisation of energy functions harder if these non-standard side chains are located near the carbohydrate binding site. Furthermore, the binding behaviour of animal lectins, especially of C-type lectins, often depends on interactions mediated by metal ions (Ca, Mn) in the binding site. This type of interactions is hard to model and not covered by our approach so far. There are approaches to metal binding based on linear functions (see *e. g.* [36]) that might broaden the range of addressable complexes, but they were not incorporated into this study.

2.3. Lectin-Carbohydrate Interactions

The chemical and structural properties of carbohydrates and lectins make the complex formation somewhat special compared to protein-peptide and general protein-ligand models. Keeping in mind that lectins bind very specifically to one kind of carbohydrate, it is evidently necessary to identify the underlying interactions responsible for that behaviour in order to create a suitable model for thoroughly predicting lectin-carbohydrate complexes.

Most peculiarities of lectins and sugars were already covered in Sections 2.1 and 2.2. There is the high flexibility of polycarbohydrates, the polar nature of their surface, the many hydroxyl groups and the resulting solvation effects on the one hand. On the other hand it is known that lectin binding sites are very shallow, resulting in ligands rather lying on a binding surface than delving deeply into a binding pocket, leaving steric hindrance out of the question when trying to explain the high specificity of binding. There clearly must exist an interaction defining the binding geometry other than spatial constraints.

Hydrogen bonds have a great influence on protein-carbohydrate binding. Every sugar ring of a polysaccharide carries at least five hydroxyl groups and hence provides a large number of hydrogen bond donors compared to the total number of atoms in the ligand. Additionally these hydroxyl groups are freely rotatable, thus enabling the ligand to adapt itself to the binding site in a multitude of ways.

However, hydrogen bonding is not the only important interaction. Looking at complexes of *Erythrina corallodendron* lectin (ECorL) binding different sugars we find an intriguing structural feature: One sugar ring always seems to be oriented quasi in parallel to the aromatic ring of phenylalanine 131 in the lectin's sugar binding site (see Fig. 2.8). This behaviour is called *ring stacking*. Generally, there are two forms of ring stacking, which are caused either by $\pi\cdots\pi$ or CH$\cdots\pi$ interactions, respectively. The $\pi\cdots\pi$ form is a stacking of aromatic rings, which cannot be observed in complexes with ordinary sugars, because there are no aromatic groups in the sugar structure. The CH$\cdots\pi$ interaction, on the other hand, is the result of several CH groups interacting with π orbitals of aromatic rings, which can be observed in sugar binding sites.

There are studies that indicate that this ring-stacking effect based on CH$\cdots\pi$ interactions guides sugars into their binding mode [13, 37, 38]. This effect, which will be covered in detail in Section 3.4.2, forms a weak hydrogen bond between aliphatic CH groups and aromatic ring systems. As sugar rings are aliphatic carbon rings offering a large number of aliphatic CH groups compared to their size, the impact on the binding mode is obvious.

The existence of ring stacking based on CH$\cdots\pi$ interactions in protein-carbohydrate complexes

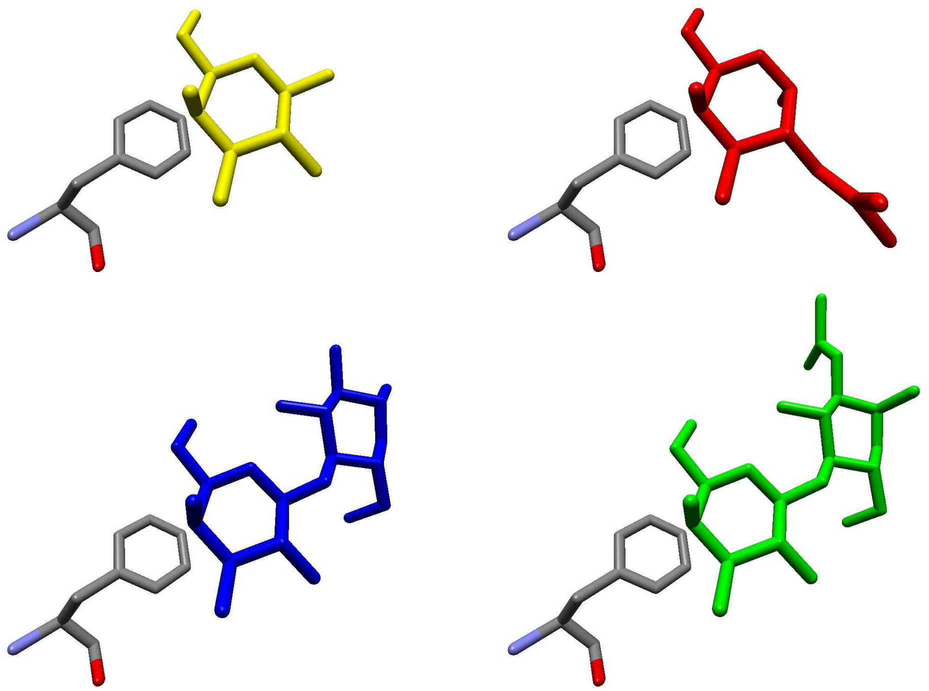


Figure 2.8.: Ring stacking in complexes of ECorL with four different sugars. Clearly, one guiding sugar ring stacks on the benzene ring in PHE 131 of ECorL.

was shown by Muraki in several experimental studies (see *e. g.* [38, 39, 40]), culminating in the conclusion [38] that carbohydrate binding proteins rely on this kind of interaction. Furthermore, Muraki theorises from the bound conformations of different sugars in the WGA binding site, that CH$\cdots\pi$ interactions tend to replace hydrogen bonds, thus making CH$\cdots\pi$ the critical component for shaping the bound ligand conformation.

CH$\cdots\pi$ interactions have been under investigation for quite a while now, particularly by the groups of Nishio (see *e. g.* his book on CH$\cdots\pi$ [41]) and Muraki. While Nishio's research concentrates on general CH$\cdots\pi$ interactions in protein-ligand interactions [42, 43], Muraki focuses on the impact of CH$\cdots\pi$ on protein-carbohydrate complexes. Furthermore, the study of Brandl *et al.* [44] examines the role of CH$\cdots\pi$ for the stability of proteins and reveals that this interaction occurs frequently between different amino acid side chains and even between side chains and protein backbone. They conclude that CH$\cdots\pi$ accounts for a significant part of the structural stability of proteins.

On the energy level, the binding of carbohydrates to lectin binding sites is rather difficult to handle because the energy differences between different ligands binding to the same lectin are very small. Moreover, the energy difference between different lectins binding the same carbohydrate are also very small. This leads to several complexes having more or less the same free energy on binding, causing much trouble for a statistical analysis. Furthermore, it means that the prediction model has to be extremely accurate in order to produce reasonable results.

Because of the polar nature of OH groups and the resulting charged surface the electrostatic solvation effects and electrostatic interaction of the system can not be neglected. Computing a thorough model for electrostatics is a computationally intensive task slowing down the calculation of binding energies drastically. We tested two approaches for these calculations, a finite difference Poisson-Boltzmann solver and the so-called Generalised Born model, the latter being the stronger approximation, but with reasonable results for most test cases. These models will be explained in Section 3.4.4.

2.4. Biochemical Importance of Solvents

Biological processes take place in a physiological environment which primarily means water. Binding processes in general are heavily influenced by the solvent they occur in. Consequently the effects of water on the binding processes have to be considered. From a computational point of view, Water is a difficult solvent. It is highly polar, thus water molecules attract each other very strongly. Additionally water molecules form hydrogen bonds easily leading to cage-like ordered molecules around small hydrophobic solutes [45].

Chervenak and Toone [46] directly measured the enthalpic influence of solvent rearrangements on the binding enthalpy of protein-ligand complexes. They proved that rearrangements account for about 10% of the binding enthalpy of the investigated complexes. Liu and Brady have shown that water molecules do rearrange around solvated sugar molecules in a clearly anisotropic fashion [19] making it almost impossible to predict solvent behaviour with strongly approximating approaches. Consequently, a thorough treatment of solvation effects seems necessary for a decent computational prediction of binding free energies.

3. Computational and Physical Background

Creating a docking programme for protein-carbohydrate complexes requires deep insight into the physical and computational background of the underlying interactions. Additionally, a model describing molecules is necessary, as well as techniques for molecular docking. This chapter will cover the interactions in detail – from the physical and the computational perspective – after introducing the ideas of molecular modelling, which are the basis for the molecular description. Furthermore, docking strategies and some statistical methods will be briefly addressed.

3.1. Basic Principles

The key question to be addressed when searching for new drugs is whether a new or adapted compound will bind to a target molecule. Structural bioinformatics generally tries to predict molecular properties based on physical and/or chemical properties of molecules with known three-dimensional structure at atomic level. Although the computational models in use are usually strong approximations of the real world, the underlying physical laws are always the basis for any prediction.

In this context, the most important fact is that everything is driven by energy. The movement of atoms is a direct consequence of the energy of a system which is the result of interactions between atoms. These interactions can be abstractly formulated as potentials that link atom positions with the energy of the system, although this formulation is in many cases much too complicated to be done analytically. Therefore many interactions are modelled with an approximative numerical method.

Why is the energy of a system so important? Nature always tries to reach a state of minimal energy. If the total energy of a system can be minimised by forming a complex of two molecules, then these molecules will bind with high probability. This is, of course, a very simplified picture of nature's behaviour, but it should clarify that having a method for reliably predicting energies will at the same time reliably predict the physical behaviour of the molecules in question and thus the binding affinity. Any model that aims at energetic accuracy has to embrace the physical context.

This work aims at predicting the *binding energy* ΔE^{bind} of a molecular complex. This energy is the difference between the energy E of the system when molecules A and B form a complex and the energy of the system when these molecules are unbound.

$$\Delta E^{\text{bind}} = E(AB) - (E(A) + E(B)) \tag{3.1}$$

If this difference ΔE^{bind} is negative, then the formation released energy thus minimising the total energy of the molecular system and obeying nature's law. If the difference is positive, then additional energy is necessary to form that complex, which means that this complex will not occur in a natural environment.

Unfortunately, reality is not that simple. What we really want to predict is the *binding free energy* ΔG^{bind} which also accounts for the thermodynamics of the system and incorporates *enthalpy* ΔH, *entropy* ΔS and *temperature* T of the system. The binding free energy is given by

$$\Delta G = \Delta H - T\Delta S \tag{3.2}$$

The methods described in this work do not consider thermodynamic properties explicitly, except where otherwise noted, but assume that the thermodynamics of the system are covered by additional constants. This is of course an approximation but valid under the assumption that the difference in the thermodynamic state variables is small compared to the energetic change on binding.

3.2. Molecular Modelling

In order to address the docking problem a concept for representing atoms, molecules and their interactions is necessary. This section covers these representations and gives a short introduction to molecular mechanics, which is the basis for computational molecular docking.

3.2.1. Modelling Molecules

Computing energies of molecular systems requires a model for the molecules under investigation, *i. e.* a theoretical representation of the molecular features allowing to do computations on the model. Molecules are built of atoms that consist of a nucleus and electrons. Simply speaking, electrons move around the nucleus in a certain space around the nucleus. This space is called electron orbital. These orbitals have particular shapes depending on the electronic configuration of an atom. Bonds are formed by overlapping electron orbitals if merging the orbitals poses an energetic advantage over not forming the bond. The mechanisms responsible for this process are the subject of quantum mechanics and orbital theory.

Taking all these details into account when doing calculations on molecular representations quickly boosts the complexity of the model to a level where computations are not feasible anymore. iTherefore, the model of a molecular system must be an approximation of the real world if reasonable computations shall be practical.

The necessary degree of approximation is clearly a function of the complexity of the systems under consideration. While small systems can still be calculated using small approximations, larger systems can only be calculated by using stronger approximating models of the physical reality. Obviously the accuracy of the calculations is directly connected to the degree of approximation. In practise this means that the bigger the systems are, the stronger the approximations have to be and the less accurate the results are.

When dealing with molecules as big as whole proteins with thousands of atoms and tenthousands of electrons, the approximations needed for calculating molecular features are very strong. One fundamental assumption used in modelling molecules is the Born-Oppenheimer model which assumes that the equations of motion of the nuclei and the electrons of a molecular system can be separated. Calculating these functions for systems with a very small number of electrons (and nuclei) is feasible with pure quantum-mechanical methods. But the computational complexity of these approximations of the quantum-mechanical reality is somewhere between $O(n^4)$ and $O(n^8)$ with n being the number of electrons of the system. It is obvious that proteins with tens of thousands of electrons cannot be calculated with such methods in acceptable time scales.

Using the Born-Oppenheimer assumption, the approximation level can be driven to the point where only the motions of the nuclei are relevant and the electronic structure can be neglected. It is then necessary to define a potential which describes the motion of all nuclei. Finding a model that defines a reasonably accurate potential for the description of the nuclear motions is far from trivial.

This kind of modelling molecules is called *molecular mechanics* (MM). As the name suggests, it describes the dynamic behaviour of a molecular system with the help of classical mechanics, thus reducing the complexity of the molecular description from Schrödinger equations to much simpler Newtonian mechanics. In molecular mechanics, electrons are not explicitly modelled and only the much slower motion of atom nuclei are considered. The atoms themselves are represented by discrete spheres with a certain radius depending on the atom's element and its chemical surrounding. Interactions between atoms are described by potentials from classical mechanics. For example, the oscillation of two bound atoms about their ideal bond length can be approximated by means of harmonic potentials, which are easily and rapidly computable. For a more detailed and thorough treatise on molecular mechanics and force fields, see *e. g.* the book by Leach [32].

3.2.2. Energy Functions

Energy functions are a general concept in structural bioinformatics. They are used for many tasks, such as geometric optimisation of experimental structures (as a part of a force field, see below) and simulation of the dynamics of a molecular system. In molecular mechanics, energy functions are generally *empirical* energy functions. Such functions extrapolate from the information gathered from experimental data, *i. e.* empirical knowledge.

In molecular modelling, energy functions always share the same mathematical structure. Different interactions between molecules result in different energetic contributions to the total energy of a system. Therefore, energy functions consist of several individual terms, each representing one kind of interaction. This formulation as individual terms is only possible under the assumption that these energies are completely separable, *i. e.* truly independent terms. Generally, the model has to make sure that energy contributions are independent. The general form of an empirical energy function E describing the energy of a molecular system m is

$$E(m) = c_0 + \sum_i c_i E_i(m) \tag{3.3}$$

where the E_i are different energy contributions and the c_i are adjustable coefficients of the function. Frequently used energy contributions include the van der Waals energy of molecules, electrostatic interactions, rotational entropy loss, hydrophobic interactions and many more energies known to influence the kind of complex which is under investigation. Thus, designing a energy function requires knowledge about the physical and chemical interactions governing the domain of molecular systems one is interested in.

Empirical methods are usually rather strong approximations. Their coefficients c_i are fitted against a set of experimental data in order to predict energies. The data set, often referred to as *calibration set*, has to be two-fold. On the one hand, the binding free energies such molecular systems is needed. One the other hand, structural data on the molecules is necessary, which basically means that the atom positions, the bonds between atoms, and the element of each atom have to be known. Additional information like atom radii and charges are usually not gained from the experiment and have to parameterised for the calculations. Chapter 4 will cover the experimental side in more detail.

The main advantage of empirical models is speed. Because the terms used in the empirical formulation are chosen to be very simple, the computational effort necessary for calculating the results is very low. In many molecular mechanics applications, speed is critical. However, the speed obviously comes at the price of reduced accuracy.

But there are other problems. One problem associated with empirical energy functions is the questionable transferability of results. Empirical methods are always fitted to a limited set of experimentally accessible data. Thus the parameters obtained by fitting a method against empirical data are in most cases only applicable in a very narrow region of relatively similar problems. For example, if a molecular mechanics model is calibrated with data consisting only of protein data, it will most certainly not be useful for predicting features of DNA strands. Consequently, there are quite many empirical models for different problems available depending on what the purpose of the model is.

3.2.3. Force Fields

The motion of atoms in a molecular system is caused by forces acting on every atom. In order to determine the dynamics of a molecule, these forces must be calculated. Force and energy are closely related. In fact, acting forces are the result of the current energy of a system.

Since nature always tries to reach the energetically minimal state, the forces acting on atoms in a molecule are always directed to the energetic minimum of the system. Therefore, energy and force are directly linked by the negative derivative of the energy function. In three-dimensional space this derivative is represented by the Nabla-operator ∇. Every differentiable energy function can be directly converted into a term defining the forces acting on a system by applying this operator. In the simple case of one particle located in a position dependent energy field $E(\mathbf{r})$, the force $\mathbf{F}(\mathbf{r})$ acting on the particle at location $\mathbf{r}$ is given by

$$\mathbf{F}(\mathbf{r}) = -\nabla E(\mathbf{r}) \tag{3.4}$$

The generalisation of that concept to systems with many particles or bodies is straight-forward.

An energy function hence defines the forces acting on atoms, thus describing a *force field*. In molecular mechanics, the term “force field” does not only denote an energy function but also includes the atomic parameters necessary for calculating the energy function, which are combined to a *parameter set*. For example, the AMBER force field consists of one energy function and a large set of parameters for atoms and their interactions. In fact, there is not only one parameter set for the AMBER energy function. Parameters can be adapted in such a way that the same energy function is applicable to different problems, although the underlying model will only permit a certain range of domains. In this study, the term “force field” will always refer to molecular mechanics force fields.

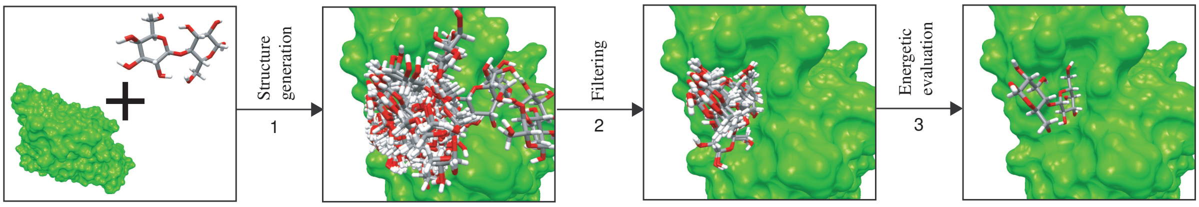


Figure 3.1.: The abstract docking scheme: Start from the spatial structures of receptor and ligand, (1) create a large number of putative docking candidates, (2) filter out bad approximations and (3) evaluate the remaining structures energetically.

3.3. Molecular Docking

The main goal of this work is the creation of a molecular docking programme for the domain of protein-carbohydrate complexes. In general, docking programmes are software tools for the prediction of bound complexes of two chemicals, usually a protein with some sort of ligand. To put it short, the docking problem can be formulated as follows: Given two molecules A and B, which are known to form a complex, compute the complex structure AB and its binding affinity.

Roughly, docking programmes work by creating a rather large number of tentative complex structures, filtering these tentative structures by certain, selectable features and evaluating the remaining complex structures energetically, *i. e.* calculating ΔG_{bind}. The complex structure with the best free energy on binding should be a good approximation of the real complex. These programmes need some ingredients for working properly: the spatial structure of both receptor and ligand, a strategy for creating putative complex conformations, a filter for sorting out bad putative complexes and an energy function for deciding how well a tentative complex approximates a natural complex.

There are diverse approaches to molecular docking, differing in how complex creation, filtering and evaluation are done, ordered or combined. Moreover, computer docking programmes are divided into different domains of application. There are tools for docking proteins to proteins, others for docking small ligands to protein binding sites which are further distinguished by the type of ligand to be docked, *e. g.* peptides or drugs.

Energy function, filtering and strategy for complex creation define the major part of a docking programme. In most cases, the latter is fixed, *i. e.* a docking programme may be adapted to different domains by exchanging the energy function with a more suitable one, but the structure creation strategy stays the same. Generally, in protein-ligand docking the ligand is considered as flexible molecule, *i. e.* groups can rotate around bonds, thus changing the conformation of the molecule. The receptor, however, stays fixed. This model is justified by the fact that in many cases proteins do not drastically change their conformation on binding, while the binding ligands have to change their conformation dramatically in order to build the correct interactions or fit into the binding pocket. The complex creation step thus only searches the so-called *conformational space* of the ligand and not of the protein.

This model clearly is an approximation, because amino-acid side chains in the binding site of a protein will probably rearrange on binding, at least to some extent. Furthermore, there are cases where the whole protein will change its entire conformation when binding a ligand. However, including receptor flexibility into the computational model will dramatically increase the computational power required for a docking run. In most practically relevant cases the results of docking to a rigid receptor seem to justify the approximation.

3.3.1. Scoring Functions

For filtering tentative complex structures generated during the first phase of docking experiments, docking programmes frequently use a so-called scoring function. A scoring function is in principle an energy function calculating approximative binding affinities of a complex. In some cases, energy functions are used for scoring tentative complexes, but several docking programmes distinguish between scoring and energetic evaluation. The discrepancy between scoring and energy function then lies in computational demand and accuracy. While energy functions are designed to calculate most accurate binding free energies and do so by employing rather complex and timely computations, scoring functions have to be very fast in order to filter very many conformations according to their putative binding affinity. Thus, the scores calculated by a scoring function cannot be treated as real binding free energies.

Scoring functions are in most cases empirically determined functions consisting of several additive contributions that determine the potency of two molecules to form a complex. The score S of a complex is then the weighted sum of all contributions S_i for a complex conformation m.

$$S(m) = s_0 + \sum_i s_i S_i(m) \tag{3.5}$$

As in the case of energy functions, the choice of the individual contributions S_i is determined by the domain of complexes the scoring function has to work upon.

3.3.2. Strategies for Structure Generation

Molecular docking finds binding conformations by creating a large set of tentative complex structures and determining their binding affinity. Numerous approaches to creating such putative complex structures have been developed. There are constructive approaches [47], building the ligand up from small fragments or structure generators searching the conformational space based on genetic algorithms [5, 48]. Some approaches use probabilistic methods or screen the conformational space based on geometric complementarity [49]. Many more different ways of generating structures are existent and under development. For this study, such a strategy has to be chosen.

The choice of search strategy is usually based on speed and accuracy considerations. As docking programmes are usually fixed in terms of energy function *and* search strategy, a direct comparison between search strategies is not easily possible. Although there are many studies which try to compare different docking programmes (see [50, 51, 52, 53, 54, 55] and many more) the only real result gained from these studies is that docking programmes cannot be compared. When trying to compare the setups and results of the different studies, the first striking observation is that most setups are not comparable. This is clearly a consequence of the focus of the respective studies regarding the choice of molecular complexes under investigation. Even with similar sets of docking programmes and molecules, the results of some studies are contradictory. A programme that performs well in one study can easily be found defeated in another. So the question for the best docking programme is still not answered and neither is the question for the best search strategy.

In order to find a reasonable search strategy, a comparison should use several strategies but only one energy function. Vieth and coworkers [51] compared different strategies which were implemented in their group and adapted to all use the same energy function. They also included the AutoDock programme [5] in their study. The docking schemes under investigation were a genetic algorithm (GA), a simulated annealing approach (SA) and a Monte-Carlo strategy (MC). AutoDock itself is based on a genetic algorithm variant. Additionally, the methods were assessed with regard to the size of the set of molecules. Although the authors claim that their GA method outperforms the other approaches, an analysis of their results leads to the conclusion that the performance of the different strategies actually does not differ very much. Even AutoDock, which in the authors' opinion is inferior to the other programmes, performs reasonably well when comparing RMSD values, energy gap and running time.

It seems that the choice of search strategy is more or less a matter of preference unless there are clear structural indications why in an individual *e. g.* an incremental search should case be better than a Monte-Carlo search. Considering this, the choice of search strategy for this study was made in favour of an AutoDock-like method for practical reasons, because there was already an implementation by Jan Fuhrmann [7] at hand, which is named BALLDock. This programme employs a genetic algorithm for structure generation. The ideas behind docking with genetic algorithms will be explained in the following section.

Docking with Genetic Algorithms

In nature, organisms adapt constantly to their environment in order to better cope with their environmental conditions. In some sense, this evolutionary adaptation can be seen as solving an optimisation problem. The idea of survival of the fittest combined with mutation and selection build the basic principle behind genetic algorithms.

Generally, genetic algorithms are optimisation heuristics that employ ideas of evolutionary processes in order to find a global extremum. The variables of the problem are encoded in *chromosomes* with one *gene* for each variable. Such a chromosome represents a putative solution for the problem to be solved. The fitness of a chromosome x is evaluated by the so-called *fitness function* $f(x)$. It must ensure that fitter chromosomes represent better solutions. Chromosomes with higher fitness values are more likely to be *selected* for reproduction. When offspring is generated, it inherits genes from both parents. The *crossover* operation decides which genes are inherited from which parent. After crossover, offspring genes are *mutated* with a certain fixed mutation probability and placed in the current *population* of individual chromosomes.

A rough outline of a genetic algorithm looks as follows:

1. Randomly create a fixed number of chromosomes, each representing a putative solution to the optimisation problem (*initial population*).

2. Compute the fitness of each chromosome of the population.

3. If the predefined exit condition is not reached, go on, otherwise exit with the current chromosome as solution of the problem.

4. Remove bad chromosome and leave only a fixed number of fit chromosomes in the population (*survival of the fittest*).

5. Create a new population out of the remaining chromosomes by repeating the following steps until the population is complete.

    a) Select two parents according to their fitness value.

    b) Create offspring using the crossover operation.

    c) Mutate the offspring.

    d) Include the offspring in the new population.

6. Jump to step 2 with the new population.

If fitness function, crossover and mutation rate are reasonably defined, a genetic algorithm will converge to an acceptable solution. While crossover will provide better solutions by incorporating parts of the best solutions of the last generations, mutation will provide protection against sticking to local minima of the fitness function. This scheme is only a very general form of a genetic algorithm and many variants are in use and in development. For deeper insights into the inner workings of genetic algorithms and their convergence behaviour, please see a text book on this matter.

In flexible ligand docking, genetic algorithms can be used for traversing the search space, *i. e.* the conformational space of the ligand. The conformational state of the ligand must be defined by chromosomes and the fitness function has to evaluate a conformation in terms of putative binding affinity. Obviously a scoring or an energy function can be used for estimating the fitness of a conformation defined by a chromosome.

The method implemented by Fuhrmann employs ideas from the AutoDock programme. AutoDock is based on a so-called Lamarckian genetic algorithm (LGA) which uses Lamarck's idea that acquired phenotypic characteristics become inheritable genetic information. Although this theory was proven wrong in biology, the idea is still existent in Lamarckian genetic algorithms, because in optimisation, the local search can improve results significantly. In AutoDock, the state variables encoded in genes and collected in chromosomes form the *genotype* of an individual. From this genotype the *phenotype* of that individual can be computed. The fitness function then evaluates the fitness of the phenotype instead of the genotype. Based on Lamarck's idea that adaptations of an individual during its lifetime will be incorporated into the genetic code, a local optimisation is conducted in the phenotypic space. The result of that optimisation is then re-translated into genotypic information which in turn can be inherited by offspring of that individual chromosome. BALLDock is based on AutoDock's ideas, but differs in several ways, which will be detailed in Section 5.5.1.

3.4. Interactions

Protein-carbohydrate binding is presumably driven by a number of molecular interactions. Chapter 2 already introduced the peculiarities of sugar binding, but did not give a thorough computational basis for exploiting these interactions in predictive calculations. The following sections deal with all relevant interactions and their computational description in detail. Only with this sound computational basis, the development of SLICK is possible.

3.4.1. Hydrogen Bonds

Hydrogen bonds are formed between so-called hydrogen bond donors and hydrogen bond acceptors. Both, donor and acceptor, are highly electronegative elements. While the donor has to be bound to a hydrogen, the acceptor must possess a lone pair of electrons. A hydrogen bond then has the form D—H· · ·A with D being the donor and A the acceptor. It is an example for delocalised orbital formation in which all three participating atoms provide one atomic orbital.

Because of the orbital geometry, hydrogen bonds always follow a certain ideal geometry which defines bond length as well as bond angle. This behaviour distinguishes hydrogen bonds from undirected interactions like Van der Waals or Coulomb interactions.

Hydrogen bonds are considerably weaker than covalent bonds but nevertheless play an important role in ligand binding. Their strength can reach up to 20 kJ/mol depending on the donor and acceptor atoms involved. Thus hydrogen bonds are stronger than other intermolecular interactions and because of their specific binding characteristics can strongly influence or even dominate the binding conformation of a ligand. In protein-ligand binding, hydrogen bonds occur almost exclusively between oxygen and nitrogen atoms.

A computational model must reproduce the ideal geometry in order to model these interactions correctly. There are several models for hydrogen bonds in use, ranging from undirected potentials to purely geometric models based on the well-defined geometry of a hydrogen bond. The model presented here is based on the one introduced by Böhm [56, 57], which is part of many other energy functions like ChemScore [36] or Fresno [58].

The Böhm hydrogen bonding term is a linearly formulated model scoring putative hydrogen bonds according to their deviation from the ideal geometry. The better a hydrogen bond reproduces the ideal geometry, the higher the score of this individual bond is. The maximum score a hydrogen bond can achieve is 1. The analytical form is defined as

$$S_{\mathrm{hb}} = f(\Delta r, a^r_{\mathrm{hb}}, b^r_{\mathrm{hb}}) f(\Delta\alpha, a^\alpha_{\mathrm{hb}}, b^\alpha_{\mathrm{hb}}) \tag{3.6}$$

with Δr being the deviation of the putative hydrogen bond from the ideal length and $\Delta\alpha$ being the deviation from the ideal angle, as illustrated in Fig. 3.2.

The function $f(x, a, b)$ is a function switching linearly from 0 to 1 or vice versa. The values a and b define the limits of the transition. This function will be called *base function.* Its analytical form is

$$f(x,a,b) = \left\{ \begin{array}{lcl} \left\{ \begin{array}{ccc} 1 & \Longleftrightarrow & x \leq a \\ 1 - \frac{x-a}{b-a} & \Longleftrightarrow & a < x \leq b \\ 0 & \Longleftrightarrow & x > b \end{array} \right\} & \Longleftrightarrow & a < b \\ \left\{ \begin{array}{ccc} 0 & \Longleftrightarrow & x \leq b \\ \frac{x-b}{a-b} & \Longleftrightarrow & b < x \leq a \\ 1 & \Longleftrightarrow & x > a \end{array} \right\} & \Longleftrightarrow & b < a \end{array} \right. \tag{3.7}$$

and defines transition from 0 to 1 and vice versa depending on the relative size of the constants a and b. If a is lower than b, the function will decline from 1 to 0. Otherwise the function will rise from 0 to 1. This form was chosen to have *one* function in contrast to the two functions defined by Böhm.

This linear base function is simple in calculation but has at least two drawbacks. First, the function is not differentiable at two points, which makes this form virtually useless for force field application. Second, the rigorous transition will make the results susceptible to small errors in

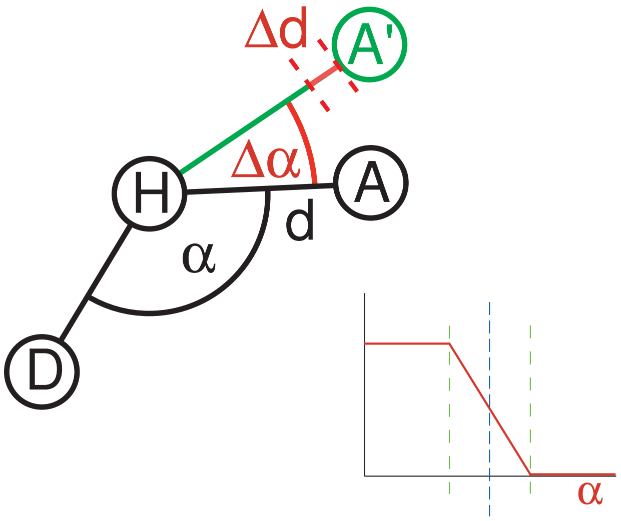


Figure 3.2.: The Böhm model for scoring hydrogen bonds and a linear switching function. D is the donor atom, A the acceptor in ideal position and A' the actual acceptor position. The angle α and the length d denote the ideal geometry values. The Böhm function scores deviations $\Delta\alpha$ and Δd from these ideal values.

the structure. Consequently a form of the base function that is differentiable and smoother in transition is desirable.

In order to soften the transition, the base function for the Böhm model was exchanged with a function based on the Fermi function

$$F(x, a', b') = \frac{1}{1 + \exp(-a'x + b')} \tag{3.8}$$

This sigmoid function defines a smooth transition from 0 to 1 (or vice versa), which we use for scoring each value. Obviously, the coefficients a' and b' have to be deducted from the original linear limits a and b. Because a and b completely define the slope of the linear function, the Fermi coefficients are calculated easily . The limits a and b define the interval of transition. Using the constraint that the slope of $F(x, a', b')$ has to match the slope of the linear function at the centre of the interval, the coefficients a' and b' can be calculated from the original linear parameters a and b with

$$a' = \frac{4}{a - b} \tag{3.9}$$

$$b' = a'(a + \frac{1}{2}(b - a)) \tag{3.10}$$

The derivation of (3.9) and (3.10) is given in appendix C.1.

Comparing the different scoring schemes as shown in Fig. 3.3 demonstrates the advantages of sigmoid scoring. In situations where the geometry is just slightly off the limits given by the Böhm model, the linear term ignores the possible interaction. In the sigmoid form, the score is just very low, which might prove useful when data is not very accurate. Consequently, interactions that might get missed in the linear form can still be detected with the sigmoid scoring, albeit with a low score.

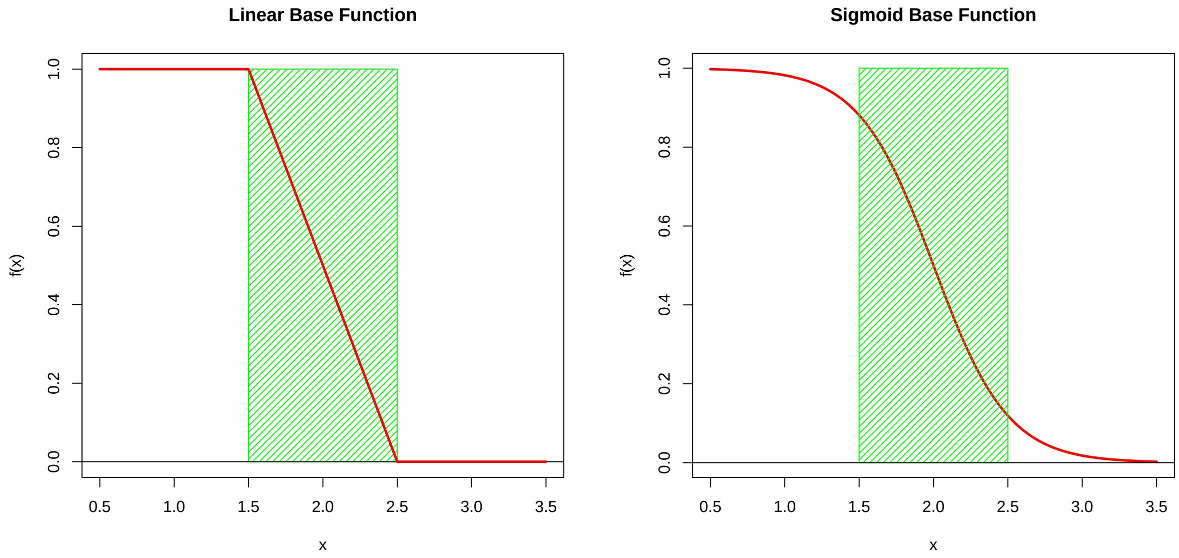


Figure 3.3.: The different base functions compared. Left: the linear transition defined by Böhm. The green area denotes the range defined by the function's limits. Right: the sigmoid base function parameterised for the same limits range (see text).

3.4.2. CH$\cdots\pi$ Interactions

CH$\cdots\pi$ interactions are weak hydrogen bonds between aliphatic CH groups and delocalised π orbitals of aromatic systems. The aromatic system acts thereby as hydrogen bond acceptor while the CH group contributes the hydrogen for the bond, although the electronegativity of carbon is rather weak. These interactions also appear between NH groups and π orbitals and were first described for NH interacting with aromatic rings by Levitt and Perutz in 1988 [37]. In protein-sugar complexes, these interactions are responsible for the characteristic aromatic-aliphatic ring stacking [38]. The computational model has to be able to detect these situations.

There are several models available for the calculation of CH$\cdots\pi$ interactions. Here, a simple geometric model by Brandl and coworkers [44] was chosen, which is very close to the known geometric formulation of ordinary hydrogen bonds. Each of the quantities of this model is scored with a sigmoid scoring function based on the parameters given in [44] and our own observations from crystallographic data.

The model consists of three "observables", shown in Fig. 3.4. It considers the distance $d_{\mathtt{CX}}$ between the carbon atom C and the centre of the aromatic ring, denoted with X. It also scores the angle $\alpha_{\mathtt{CHX}}$ between the CH bond and the connection between hydrogen atom and ring centre. The third value is the distance $d_{\mathtt{H_pX}}$ between ring centre X and the hydrogen atom projected into the plane defined by the aromatic ring.

Brandl *et al.* only give limits for the geometry of a CH$\cdots\pi$ interaction. They state that if $d_{\mathtt{CX}} < 4.5$Å, $\alpha_{\mathtt{CHX}} > 120°$ and 1.0Å $< d_{\mathtt{H_pX}} < 1.2$Å, an interaction is found. Starting from these four limits, it seems hard to define ideal geometries similar to the case of hydrogen bonds. Thus, the new computational model does not score deviations from ideal geometry but accordance with

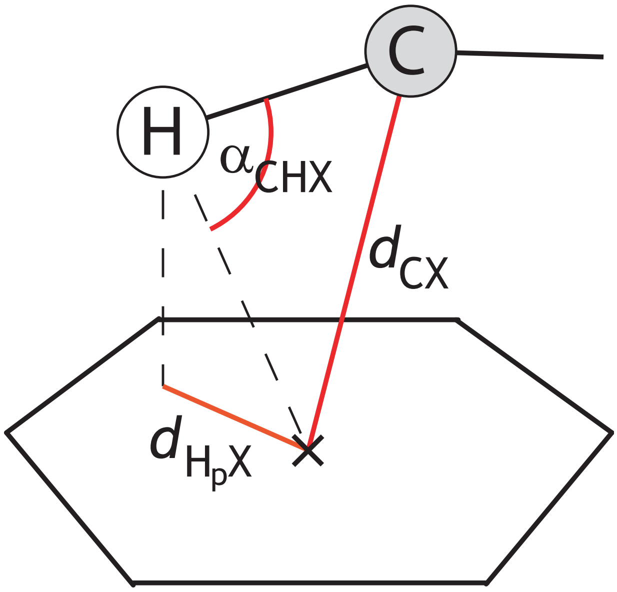


Figure 3.4.: The geometry of the CH···π interaction: $d_{\mathtt{CX}}$ is the distance between aliphatic carbon and ring centre, $\alpha_{\mathtt{CHX}}$ denotes the angle between CH bond and the connection between the hydrogen atom and the ring centre. $d_{\mathtt{H_pX}}$ is the distance of H to the ring centre projected into the ring plane.

the defined limits.

Parameter	Original	Modified
$p_{\mathtt{CX}}$	4.5Å	4.5Å
$p_{\mathtt{CHX}}$	120°	110°
$p^{l}_{\mathtt{H_pX}}$	1.0Å	0.7Å
$p^{u}_{\mathtt{H_pX}}$	1.2Å	1.7Å

Table 3.1.: Parameters for the CH···π interaction: $p_{\mathtt{CX}}$ is the upper limit for the CX distance, $p_{\mathtt{CHX}}$ defines the lower limit on CHX angles, and $p^{l}_{\mathtt{H_pX}}$, $p^{u}_{\mathtt{H_pX}}$ denote the lower and upper limit for the projected HX distance, respectively.

The limits found by Brandl are the results from analysing intramolecular CH···π interactions in proteins. Because it was not clear if these parameters are also applicable to the case of protein-carbohydrate interactions, a number of lectin-sugar complexes from PDB was structurally analysed and compared to the interactions found in the original article. Based on these investigations, the parameters were slightly modified to reflect the intermolecular protein-carbohydrate case. In Tab. 3.1, the original and the modified parameters are listed.

Two approaches for calculating CH···π interactions were implemented and tested. The first version only decided whether all observables were in the range defined by the Brandl parameters and then returned the number of identified CH···π contacts for a complex. This number was

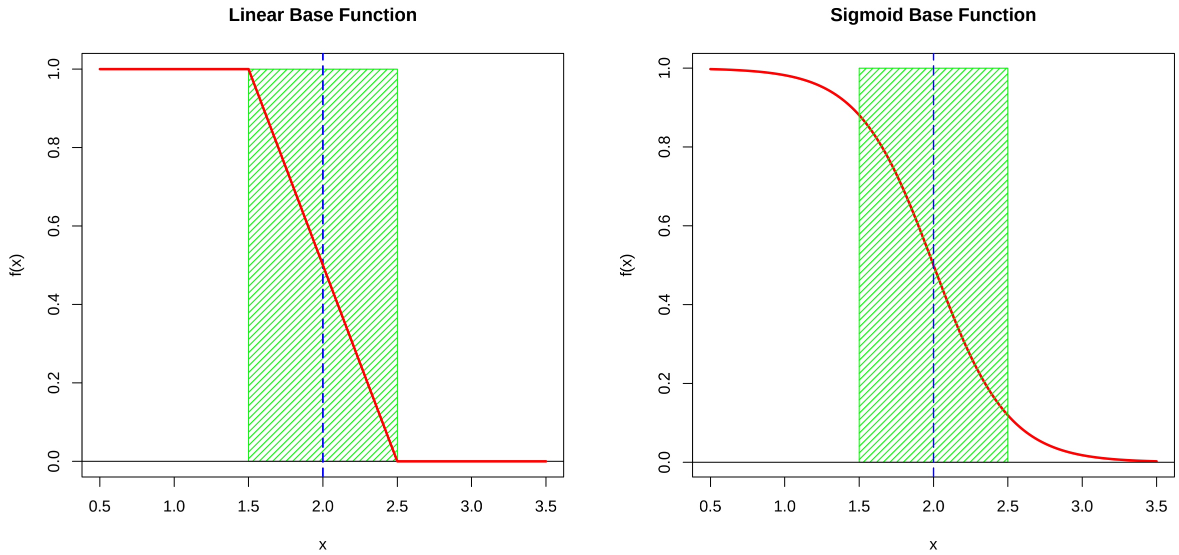


Figure 3.5.: Linear and sigmoid base function for CH$\cdots\pi$. The blue dashed line denotes the limit of the model. The transition interval is denoted by the green area.

used as a score for the complex in question. Although this simple approach improved results, it is very sensitive against small changes in ligand conformations.

Consequently, the CH$\cdots\pi$ term was improved in order to gain a score telling more about the actual interaction. Therefore, the simple yes/no scheme was successively replaced by the base functions introduced for the hydrogen bonding term. Because the parameters provided by Brandl *et al.* do not define intervals for the transition from 0 to 1, these intervals were defined by adding and subtracting a constant ϵ to/from the Brandl parameters. The value of $\epsilon = 0.25$Å was defined such that transition intervals are 0.5 Å in width (see Fig 3.5). This arbitrary definition is based on the experiences gained from the hydrogen bonding term and proved reasonable in the later calculations. The results of distinct calculations with linear and sigmoid base function suggested that choosing the sigmoid form performs better than the linear one. Therefore, the sigmoid base function was chosen for further calculations.

With the sigmoid base function F, the parameters from Tab. 3.1, and the definition of the transition intervals in place, the analytical form of the improved CH$\cdots\pi$ interaction scoring becomes

$$\begin{aligned} s_{\mathrm{CH}\pi} &= \frac{1}{3}(s_{\mathrm{CX}} + s_{\mathrm{CHX}} + s_{\mathrm{H_pX}}) \qquad (3.11) \\ \text{with} & \\ s_{\mathrm{CX}} &= F(d_{\mathtt{CX}}, p_{\mathtt{CX}} - \epsilon, p_{\mathtt{CX}} + \epsilon) \\ s_{\mathrm{CHX}} &= F(\alpha_{\mathtt{CHX}}, p_{\mathtt{CHX}} - \epsilon, p_{\mathtt{CHX}} + \epsilon) \\ s_{\mathrm{H_pX}} &= F(d_{\mathtt{H_pX}}, p^{l}_{\mathtt{H_pX}} - \epsilon, p^{l}_{\mathtt{H_pX}} + \epsilon) f(d_{\mathtt{H_pX}}, p^{u}_{\mathtt{H_pX}} - \epsilon, p^{u}_{\mathtt{H_pX}} + \epsilon) \end{aligned}$$

The values of $d_{\mathtt{CX}}$, $\alpha_{\mathtt{CHX}}$ and $d_{\mathtt{H_pX}}$ can be calculated from the atom positions using simple vector geometry.

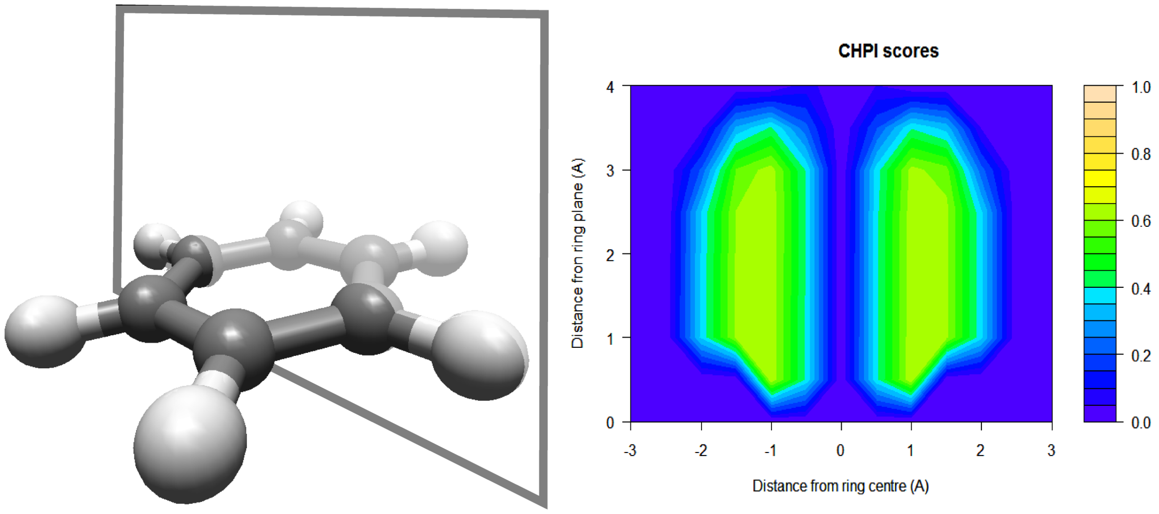


Figure 3.6.: Cross Section of CH···π over the benzene ring of PHE. The right figure displays CH···π scores calculated for the plane given as clipping plane in the left figure. Only the scores above the ring plane are shown. The scores below the ring plane are symmetric.

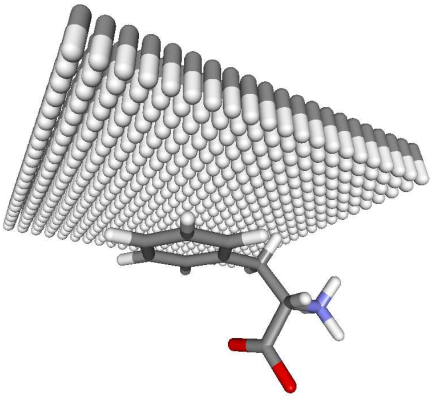


Figure 3.7.: A layer of probe groups for screening space and calculating the CH···π section presented in Fig. 3.6.

Figure 3.6 shows a contour plot of the CH···π interaction created by placing CH groups above the benzene ring of a phenylalanine amino acid. Because of the symmetry of the system, the scores below the ring are just the mirrored case and not shown in the plot. The probes were coordinated perpendicular to the plane defined by the aromatic ring and the space above the ring was screened with many such groups. Fig. 3.7 displays one layer of probe groups.

The region of high scoring is clearly visible as a somewhat distorted torus above the ring. Note that scores are still high in very close vicinity of the ring atoms, which is not realistic because atoms would overlap in this region. An energy function must take care of this behaviour

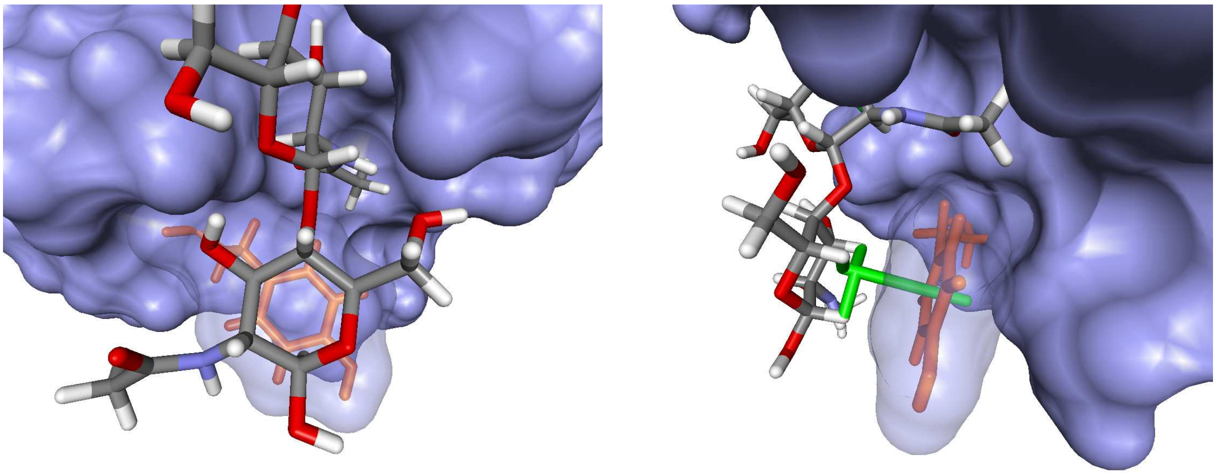


Figure 3.8.: Perfect ring stacking in 1K7U. Left: The aliphatic ring of the sugar (normal colours) is perfectly stacked onto the aromatic ring in the binding site (orange). Right: three CH···π interactions (green sticks) coordinate the stacking. All CH···π interactions that are presumably responsible for the binding mode are identified by the computational model.

and avoid atoms that overlap. Including Van der Waals interactions (see next section) provides reliable prevention of overlapping atoms.

Using this simple geometric model, CH···π interactions are identified reliably. Assessment of this term was done on intramolecular CH···π bonds in structures considered in the Brandl article, where CH···π bonds are listed explicitly, allowing for a direct comparison. Furthermore, CH···π bonds in protein-carbohydrate complexes were analysed.

Ring stacking identified by this model in protein-carbohydrate complexes is exemplary illustrated in Fig. 3.8. It shows wheat germ agglutinin (WGA) binding a GlcNAc dimer. One of the sugar rings stacks perfectly on the aromatic ring of TYR 64 in chain B of the protein. Three CH···π interactions dominate this stacking. They are identified correctly by the CH···π term developed for this work.

3.4.3. Van der Waals Interactions

Van der Waals interactions are interatomic forces. They were discovered in the late 19th century by J. D. van der Waals when he was investigating the differences between ideal and real gases. From the pressure difference between ideal and real gases he concluded that there have to be interatomic forces even between the atoms of uncharged inert gases that are at least in part attractive. This attractive part is based on induced dipoles in the electron hulls of atoms which are not dipoles themselves.

The energy of this attractive interaction can be described with the approximation by London. It states that the so-called London energy, which is the attractive part of the van der Waals energy, can be written as

$$E_{\mathrm{L}} = -\frac{3\alpha^4\hbar\omega}{4(4\pi\varepsilon_0)^2}\frac{1}{r^6} \tag{3.12}$$

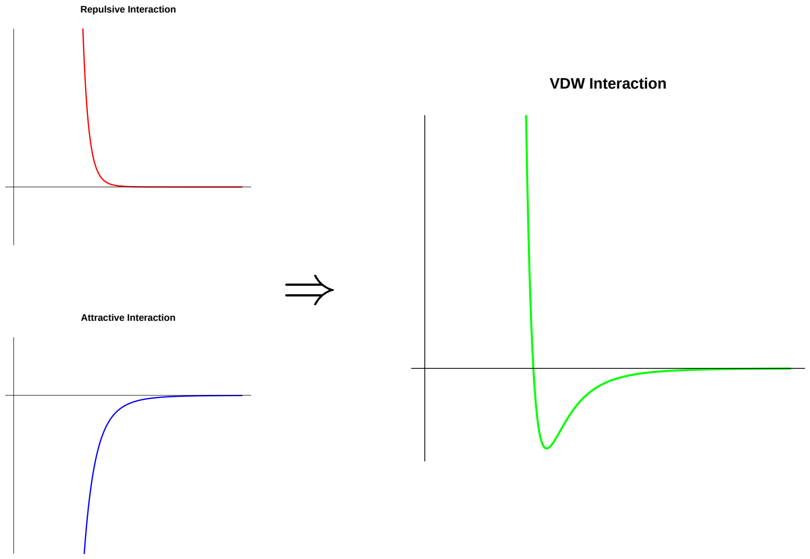


Figure 3.9.: The van der Waals potential (green) is composed of a repulsive (red) and an attractive (blue) potential.

with r being the distance between the centres of two atoms. Without going into the details of the London energy, the important information of equation (3.12) is that the attractive part of van der Waals is a term of magnitude $-\frac{1}{r^6}$.

Van der Waals forces also have a repulsive part which is based on the repulsion of charges with the same sign and indirectly on the Pauli principle. Nuclei approaching each other will repel mutually if they come too close. There are several approaches to modelling the repulsive part of van der Waals interactions, *e. g.* the Heitler-London formulation, but they all share the same form: a hard repulsive potential with very high energies for too close atoms.

Usually, van der Waals interactions are modelled with the Lennard-Jones potential illustrated in Fig. 3.9. For every pair of atoms i, j of a molecular system, the van der Waals energy is computed as

$$E_{\text{vdw}} = \sum_{i,j} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{3.13}$$

where r_{ij} is the distance between atoms i and j. The parameters A_{ij} and B_{ij} denote atom type dependent constants accounting for the van der Waals radii of the atoms.

Ferrari and coworkers have shown in [59] that in protein-ligand docking softening the potential may account for conformational changes in the receptor binding site, thus generally increasing the quality of docking results. The method can be seen as an approximation of the computation of the side chain flexibility in the receptor binding site, based on the assumption that conformational changes on binding are small. Moreover, a softened van der Waals potential will be less susceptible to errors in the experimental data or the complex structures created by structure generators of docking programmes because of diminished penalising of too close contacts.

Ferrari *et al.* used the approach of softening the van der Waals potential for enriching the

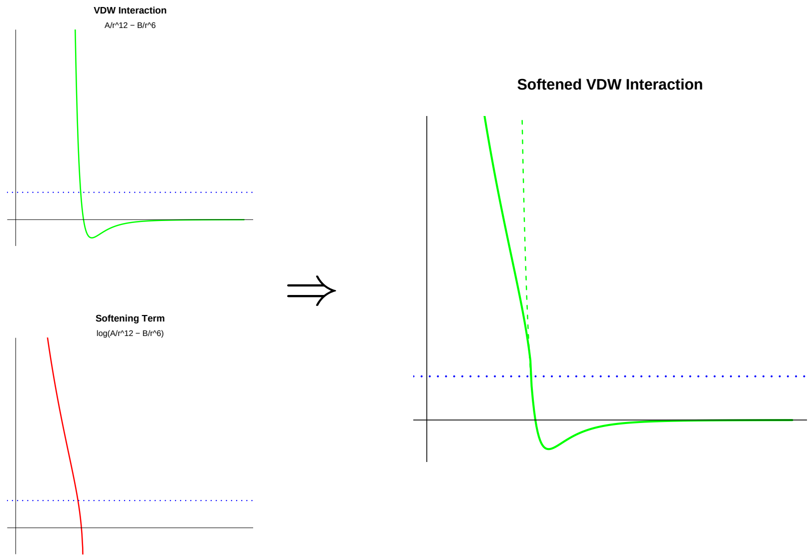


Figure 3.10.: The softened form of the van der Waals potential. Values above a limit l_{vdw} (blue dashed line) are reduced by applying the logarithm. The graph on the right hand side shows the form of the softened potential. The original van der Waals potential is shown as dashed green line.

list of putative complex conformations in the structure generation process and filtering out false complexes in a second re-evaluation step with the hard potential in place again. As a result the overall performance of the docking increased with a drawback in a somewhat decreased conformational accuracy.

In this study, a simple approach to softening the van der Waals term was chosen. The repulsive part of the potential is softened by applying the natural logarithm to values above a predefined upper bound l_{vdw} (Fig. 3.10). The energy is then given by

$$e_{ij} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{3.14}$$

$$E_{\mathrm{vdw}} = \sum_{i,j} \begin{cases} e_{ij} & \text{if} \quad e_{ij} \leq l_{\mathrm{vdw}} \\ l_{\mathrm{vdw}} + \log(e_{ij}) & \text{if} \quad e_{ij} > l_{\mathrm{vdw}} \end{cases} \tag{3.15}$$

With this approach, the van der Waals model does not need reparameterisation, but the potential is clearly not differentiable anymore. Nevertheless, results proved that softening increases docking quality. There are other, more sophisticated softened van der Waals models in use. Ferrari *et al.* suggest a 6-9 potential instead of the commonly used 6-12 potential [59]. However, such a model would need complete reparameterisation and Ferrari's work relies on a two-step algorithm doubling computational effort, which was the reason for introducing this simpler approach.

3.4.4. Electrostatics

Electrostatic interactions are probably the most important interactions. Most intra- and intermolecular interactions are caused by electrostatics. All interactions covered so far, *i. e.* hydrogen bonds, CH$\cdots\pi$ and van der Waals interactions are electrostatic interactions in nature but were modelled on a higher, approximative level, leaving the underlying interactions untouched. The simplifying model of molecular mechanics, treating atoms as spheres and considering them without electrons, does not provide enough modelling accuracy for these effects to be calculated exactly. This, of course, is an accepted reality of the approximation level adopted in these calculations.

This section will give a very brief introduction on the underlying physical principles. Furthermore, the relevance of electrostatic interactions to forming protein-carbohydrate complexes will be explained in more detail.

Basics

When considering electrostatic interactions in molecular mechanics, atoms are at first treated as point charges positioned at the atom centres. This simple model allows for rapid calculations of interactions by employing the Coulomb law, which connects the electrostatic energy E_{es} of a cloud of independent point charges q_i at their respective positions $\mathbf{r}_i$ in vacuum as follows:

$$E_{\text{es}} = \frac{1}{4\pi\varepsilon_0} \sum_{i<j} \frac{q_i q_j}{r_{ij}} \tag{3.16}$$

In this equation, ε_0 is the permittivity of vacuum, a natural constant, and r_{ij} is the distance between point charges q_i and q_j. This simple equation yields the total electrostatic energy of a system of point charges in vacuum.

The electrostatic contribution to the binding free energy $\Delta G_{\text{es}}^{\text{bind}}$ of a molecular complex is simply the difference between the electrostatic energy of the complex AB and the sum of the electrostatic energy of the individual molecules A and B.

$$\Delta E_{\text{es}}^{\text{bind}} = E_{\text{es}}(AB) - (E_{\text{es}}(A) + E_{\text{es}}(B)) \tag{3.17}$$

Thus, knowing the charges carried by individual atoms and the positions of these atoms is sufficient to calculate this important contribution to the binding free energy. For readability reasons, ΔE_{es} will from now on denote this binding contribution.

One important issue when calculating electrostatic interactions are the charges carried by the atoms of the system. In theory, charges are indivisible quantities with their value being a multiple of the unit charge e_0. Physics tells us that charges are carried by electrons, which carry a charge of $-e_0$, and protons, which carry a charge of $+e_0$. Atom nuclei are built from positively charged protons and neutrons, which as the name suggests are neutral and do not carry charges. In their ground state, atoms are always uncharged, which means that the number of positive and negative charges are balanced and sum up to 0. Atoms can be charged by removing or adding electrons. If an electron is missing, the charge balance of the atom is disturbed and the charge of the atom becomes -1 which is short for $-1 \cdot e_0$. Thus, if a molecule is charged, the amount of missing or excess charge is always a multiple of e_0 because the only way to charge it is removing or adding charge carriers.

Unfortunately, this static picture does not reproduce the natural behaviour of molecules too well. In a simplified view, electrons can move freely in their orbital space. The probability to find an electron at a certain point in space is defined by a *probability density.* This probability density changes if there is a charge nearby. In that case, charges tend to be drawn into a certain direction, disturbing the charge balance of an orbital. This phenomenon is called *charge displacement.*

Because of charge displacement, atoms in a molecular system do not seem to carry charges of unit e_0 when observed over a prolonged period, although the charge carriers still do. That is because of the permanent movement of the electrons in the orbitals under influence of external fields changing the probability of observing electrons at a certain point in space. This behaviour is accounted for in molecular mechanics by assigning so-called *partial charges*, which in some sense represent a temporal mean value for the charge observed on an atom.

Obtaining values for these partial charges, which can be used in computing electrostatic energies, is difficult. There is the possibility to calculate these charges from quantum mechanical principles, which is again a very complicated and time-consuming procedure. The usual approach is the already introduced usage of standardised parameter sets, this time containing partial charges which can be assigned to atoms of a molecule. There are several parameter sets. In most cases, these charge sets are part of a force field parameter set like AMBER or CHARMM. These charge sets are well-established and tested but apply best to the domain and energy functions defined in the particular force field. Thus, a certain amount of pre-fitting is introduced into the calculations which might diminish the generality of computed energies. See Section 3.4.10 for details on the parameter set chosen in this thesis.

Importance for Protein-Carbohydrate Complexes

As mentioned before, electrostatic interactions are very important for all kinds of molecular complexes. There are two main reasons: First, the energy contributed by electrostatic interactions is quite large compared to the energies of other intermolecular interactions. Second, electrostatics are rather long-ranged interactions in contrast to *e. g.* van der Waals. Electrostatics impact atom movements over large distances, because the potential is reciprocal in the distance of two interacting atoms. Potentials like van der Waals diminish much quicker.

In the simple view of Coulomb's law, which assumes vacuum conditions and does not include so-called non-classical effects, the range of influence of the electric fields resulting from atom charges is already rather long-ranged. When including *e. g.* non-local effects, the potential can even be more influencing on atoms further away [60]. As pointed out in Section 2.3, carbohydrates carry many freely rotatable hydroxyl groups. These groups consist of an oxygen and a hydrogen. Oxygens are far more electronegative than hydrogens, which means that the electrons of the O-H bond are drawn into the direction of the oxygen atom. Thus, the partial charges of the oxygen and the hydrogen of the group have different signs resulting in a so-called *polar group.* Obviously, a polar group will develop much stronger electrostatic interactions than non-polar groups where charge carriers are balanced and therefore lead to much smaller partial charges. Having in mind that hexoses carry five hydroxyl groups, the strong impact of electrostatics on the binding behaviour of carbohydrates becomes evident.

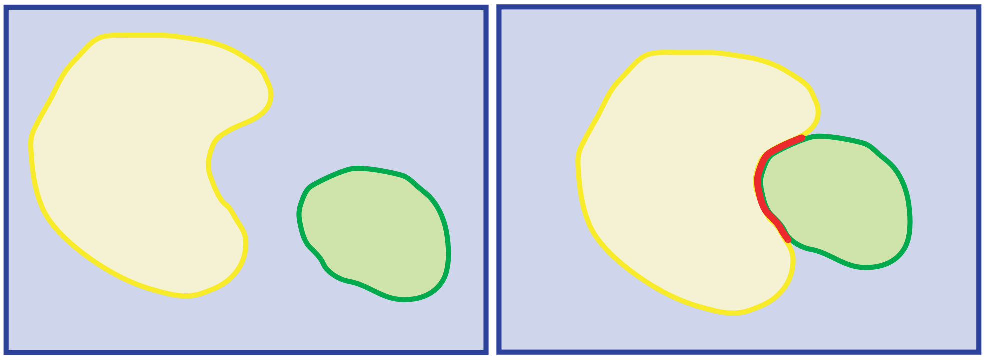


Figure 3.11.: Desolvation at the molecular interface surfaces. Left hand side: Two molecules completely solvated in water. The whole surface of both molecules is surrounded by water, hence *solvated*. Right hand side: The molecules have bound. The binding surface, depicted by a red line, is no longer covered by water. It has been *desolvated*.

3.4.5. Solvation

Biological processes take place inside living organisms. Hence, the systems we look at, *i. e.* biomolecular complexes, will never be found in vacuum but in physiological milieu, which basically means water. The process of bringing molecules into a solvent is called *solvation*. This process is characterised energetically by the *solvation free energy* ΔG^{solv} which for any water-soluble compound is negative because energy will be liberated during that process.

Solvation influences the binding process of molecules. In the unbound state, both molecules are completely surrounded by the solvent. Their entire molecular surfaces are in contact with solvent molecules. On binding, parts of these surfaces lose their contacts with the solvent and build up contacts with the surface of the respective binding partner (see Fig. 3.11). These interfacing surface areas become *desolvated*, which is the reverse process of solvation and thus contributes to the binding free energy.

How can these influences be modelled? As the solvent molecules interact with the solvated molecules in more or less the same way as the binding partners do, the change in interactions on binding can be computed by placing water molecules around the complex under investigation and compute all the interactions explicitly. The problem is that there are so many water molecules to be considered. When computing the interactions between binding partners and solvent molecules, one has to ensure that there is a sufficient amount of water around the complex in order to satisfy long-range interactions like electrostatics. For an average protein-carbohydrate complex, the number of necessary water molecules easily reaches 10^5 and more. This approach needs too many resources.

Implicit solvent models are a way out of this computational dilemma. They do not consider the explicit water molecules but so-called *bulk properties* of a large amount of water molecules representing the average influence on the solvated molecules. With this approximation, the influence of a large amount of solvent molecules can be computed much more efficiently. The following sections will provide details on solvation theory and the computational methods based on implicit solvent models employed.

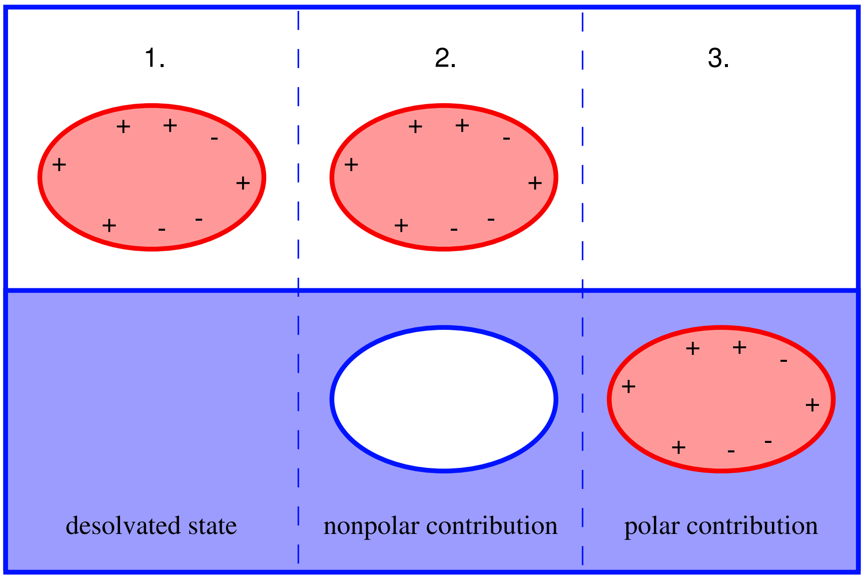


Figure 3.12.: The different phases of the solvation model: 1. Before solvation, the molecule is in vacuum. 2. A cavity of the size and shape is opened in the solvent and the molecule is transfered into this cavity without its charges, yielding the nonpolar contribution. 3. The molecule's charges are transfered into the molecule, leading to electrostatic interactions between molecule and solvent.

Solvation Theory

The *solvation free energy* ΔG^{solv} is the difference between the energy of a molecule in the ideal gas phase and in solvated or ideal liquid phase.

$$\Delta G^{\text{solv}} = G^{\text{solvated phase}} - G^{\text{gas phase}} \tag{3.18}$$

When complexes bind there is a change in solvation free energy because parts of the molecules' surfaces become desolvated at the interface of these bound molecules. Let A and B be the binding molecules and AB their complex, then the change in solvation free energy on binding is

$$\Delta\Delta G^{\text{solv}} = \Delta G^{\text{solv}}(AB) - (\Delta G^{\text{solv}}(A) + \Delta G^{\text{solv}}(B)) \tag{3.19}$$

Solvation theory divides the process of solvating a molecule into a solvent into two parts, illustrated in Fig. 3.12. The *nonpolar* part $\Delta G^{\text{solv}}_{\text{np}}$ of the solvation free energy is caused by the process of bringing a molecule into a solvent without considering electrostatic interactions between solvent and solute. After that, charges are transfered into the molecule, generating in the second part of ΔG^{solv}, the *polar* or electrostatic part $\Delta G^{\text{solv}}_{\text{es}}$. It results from electrostatic interactions between solvent and solute. The sum of these two contributions is the total solvation free energy of a molecule

$$\Delta G^{\text{solv}} = \Delta G^{\text{solv}}_{\text{np}} + \Delta G^{\text{solv}}_{\text{es}} \tag{3.20}$$

Solvation effects are of enthalpic and entropic nature. For this reason, the discussion of solvation effects will include some references to enthalpic and entropic effects.

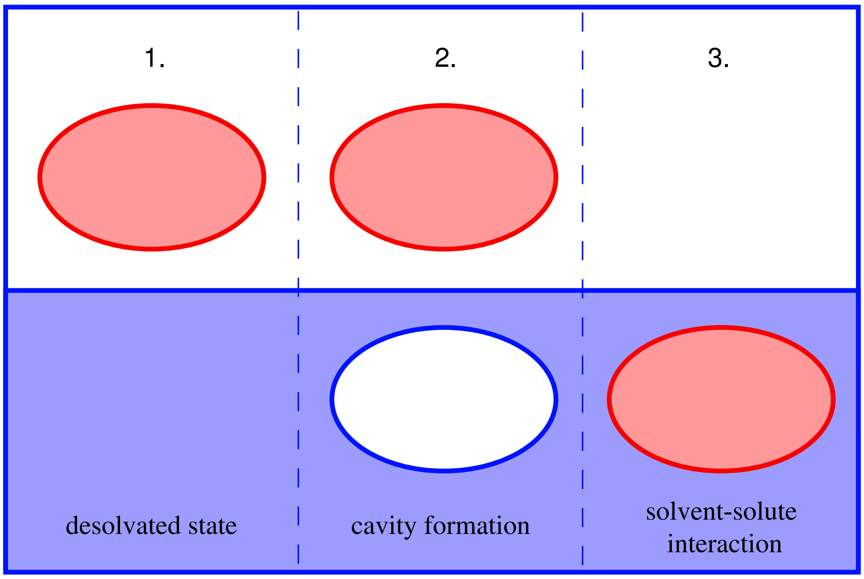


Figure 3.13.: The different phases of the nonpolar solvation model: 1. The molecule is in gas phase. 2. A cavity of size and shape of the molecule is opened in the solvent. 3. The molecule is transfered into the cavity. Molecule and solvent can interact.

For simplicity reasons, the change of solvation free energy on binding $\Delta\Delta G^{\text{solv}}$ will often be referred to as solvation free energy or solvation energy in this chapter. It should always be clear from context and formulae which solvation free energy is under consideration.

3.4.6. The Theory Behind Nonpolar Solvation Effects

Current theory divides nonpolar solvation energies into different contributions based on the picture that nonpolar solvation can be separated into independent steps. The first step is to create space in the bulk solvent which is exactly as large as the molecule which will be transferred into the solvent. From a theoretical point of view, in this phase infinitesimally small "holes" are brought into the solvent, which in a manner of speaking are then blown up to the size of the molecule that is to be solvated. Building this *cavity* obviously consumes energy. Additionally, it creates a "forbidden space" for the solvent, thus contributing to a change in system entropy. During this phase, there are no interactions to be considered because the molecule is still not in the solvent.

The second step of nonpolar solvation is the transfer of the molecule into the cavity in the solvent. As soon as the molecule is in place, the interactions between solute and solvent are considered. For the nonpolar part, these interactions are merely van der Waals interactions. In literature these are referred to as dispersive and repulsive interactions and are in some cases also considered separately. The total nonpolar solvation energy sums up to

$$\Delta G^{\text{solv}}_{\text{np}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdw}} = \Delta G_{\text{cav}} + \Delta G_{\text{dis}} + \Delta G_{\text{rep}} \tag{3.21}$$

There are many approaches to calculating nonpolar solvation energies. The next few sections will introduce the most promising ones which were also implemented and tested during the development of this work.

3.4.7. Computational Models

Surface Tension

The simplest model for the nonpolar contribution is based on surface tensions. It was found by Uhlig [61] and used in several contexts. Some models use the Uhlig model as a mere cavitational contribution while others compute the whole nonpolar part with an Uhlig-like term.

The model of Uhlig is based on a solubility expression derived directly from the Clausius-Clapeyron equation

$$\frac{dp}{dT} = \frac{\Delta S}{\Delta V} = \frac{\Delta H}{T \Delta V} \tag{3.22}$$

where p is the pressure, V is the volume, T is the temperature, S is the entropy and H denotes the Helmholtz energy (or enthalpy) of the system. This equation describes phase transitions of substances and thus is a basis for calculating the transition from the gas-phase to the solvated (or liquid) phase.

Uhlig's model describes the solubility of an ideal gas in an ideal liquid. For solvating a spherical gas molecule in a liquid a cavity has to be created in the liquid. The amount of work necessary for creating this cavity is assumed to be given by the increase of cavity surface area multiplied by the surface tension of the solvent. In addition to the cavity creation, interactions between solvent and solute contribute to the solubility. This model uses a macroscopic concept in molecular dimensions, an approach disregarding all effects that arise on atomic length scales. Using the nomenclature of Uhlig's article [61], the energy difference Δu between solvated and low-pressure gas phase of a spherical substance with radius r is

$$\Delta u = 4\pi r^2 \sigma - E \tag{3.23}$$

with σ being the molecular solvent surface tension and E being the interaction energy of the solute with the solvent. Uhlig argues that variations in E in going from one solvent to another are very small and that E will also be rather dispensable for ideal gases which only interact very weakly with their surrounding. Thus the transfer energy and the solubility will mainly depend on the term $4\pi r^2 \sigma$ which is the surface area A of a sphere of radius r multiplied with the surface tension σ of the solvent. Thus the transfer energy can be written as

$$\Delta u = \sigma \cdot A \tag{3.24}$$

Translated in our nomenclature and arbitrary surfaces rather than only spherical ones, expression (3.24) becomes

$$\Delta G_{\mathrm{np}}^{\mathrm{solv}} = \gamma \cdot A \tag{3.25}$$

with γ being a generalised surface tension and A being the surface area of the solute. This generalisation is valid under the assumption that the surface tension approach is applicable to surfaces other than spherical ones. Most sources use the letter γ instead of σ, mainly to discriminate a pure surface tension from an approximative solubility coefficient.

The surface area A in equation (3.25) depends on the model used for calculating surface terms. In most cases the solvent-accessible surface (SAS, [62]) is used. The SAS describes the surface accessible by solvent atoms and thus represents the boundary for solvent-solute interactions. In this work molecular surfaces are always calculated with the SAS method.

Scaled Particle Theory

Uhlig's approach is very simple, but shows acceptable accuracy for small molecules. Nevertheless, it is frequently used on larger molecules, as well, although it often needs reparameterisation. In the 1960's, Reiss, Frisch, Lebowitz and coworkers [63, 64, 65, 66, 67, 68] developed a more elaborate statistical mechanical theory of fluids based on radial distribution functions. Their so-called *scaled particle theory* (SPT) describes solvation thermodynamics of dilute solutions. In their work, Reiss *et al.* only consider interactions that are electrostatic, hence, van der Waals interactions between solvent and solute are not considered in their theory for nonpolar solvation effects. Consequently, the Reiss term should be regarded as a mere cavitational term.

In classical thermodynamics, the reversible work $E(r)$ required for inserting a spherical molecule of radius r into the solvent is defined as

$$E(r) = \frac{4}{3}\pi r^3 P + 4\pi r^2 \gamma \left(1 - \frac{4\delta}{r}\right) \tag{3.26}$$

where P is the pressure, γ denotes the surface tension and δ is a constant describing surface curvature. In this equation, the first summand is the volume work and the second describes the surface work with a corrective for surface curvature denoted by δ.

In the original formulation, SPT considers solvent and solute molecules as hard spheres with different radii. From statistical mechanical considerations, Reiss *et al.* deduct an approximation for $E(r)$ of the form

$$E(r) = K_0 + K_1 r + K_2 r^2 + K_3 r^3 \tag{3.27}$$

which obviously has the the same form as the law of classical thermodynamics. The question remains how the K_i have to be defined. By comparison with the classical law, Reiss *et al.* determine the constants K_i to be

$$K_0 = kT\left(-ln(1-y) + \frac{9}{2}\left(\frac{y}{1-y}\right)^2\right) - \frac{1}{6}\pi P r_S^3 \tag{3.28}$$

$$K_1 = -\frac{kT}{r_S}\left(6\,\frac{y}{1-y} + 18\left(\frac{y}{1-y}\right)^2\right) + \pi P r_S^2 \tag{3.29}$$

$$K_2 = -\frac{kT}{r_S^2}\left(12\,\frac{y}{1-y} + 18\left(\frac{y}{1-y}\right)^2\right) + 2\pi P r_S \tag{3.30}$$

$$K_3 = \frac{4}{3}\pi P \tag{3.31}$$

with $y = \frac{1}{6}\pi r_S^3 \rho_S$. Here, r_S is the radius of the spherical solvent molecules, $\rho_S = \frac{N}{V}$ is the number density of the solvent with N being the number of solvent molecules and V being the volume of the system.

Reiss *et al.* legitimate the occurrence of a constant term K_0 by arguing that equation (3.26) is a macroscopic formula while equation (3.27) will be used on microscopic length scales Therefore a correction term for very small effects that do not affect the classical case is needed in the microscopic case.

The above form of the coefficients K_i is quite unhandy. Pierotti reports in his review [69] the following, more concise form. Let $\mathbf{R} = \frac{r}{r_S}$ be the ratio of the hard sphere radii. Then the work for creating a cavity of radius r in a hard-sphere fluid with number density ρ_S is

$$\frac{E(\mathbf{R}, \rho_S)}{kT} = -\ln(1-y) + \left(\frac{3y}{1-y}\right)\mathbf{R} + \left(\frac{3y}{1-y} + \frac{9}{2}\left(\frac{y}{1-y}\right)^2\right)\mathbf{R}^2 + \frac{yP}{\rho_S kT}\mathbf{R}^3 \tag{3.32}$$

where $y = \frac{1}{6}\pi\rho_S r_S^3$ is the reduced number density of the solvent.

Equation (3.32) constitutes a theoretically well-defined expression for the calculation of the work for cavity formation if the cavity is spherical. Unfortunately, proteins are scarcely spherical in shape. Therefore, following the suggestion of Langlet and coworkers [70], the Reiss equation is taken as an atomic contribution summed up over all atoms of a protein surface. The cavitational energy is then

$$\Delta G_{\text{cav}} = \sum_i \frac{A_i}{4\pi r_i^2} E(\mathbf{R}_i, \rho_S) \tag{3.33}$$

where i is the index of the i-th atom of the solvated molecule, r_i is the radius of that atom, r_S is the radius of the solvent molecules, $\mathbf{R}_i = \frac{r_i}{r_S}$, ρ_S is again the number density of the solvent, $E(\mathbf{R}_i, \rho_S)$ is the work for creating a cavity for that atom in the solvent and A_i is the surface area occupied by that atom on the solvent accessible surface of the protein. With this equation, every atom of the solvated molecule contributes to the energy depending on its portion of the molecular surface.

Like in the Uhlig case, this generalisation is valid under the assumption that the term for surface work is applicable to surfaces that are not spherical. Nevertheless, volume work might be underestimated or simply neglected in this formulation. However, in this study the application of this term will be limited to calculating binding energies, *i. e.* differences between energies. Assuming that the protein is rigid and thus does not change its volume and that the ligand is so small that there are no atoms buried so deeply that they do not contribute to the surface of the molecule, the difference between bound and unbound state will almost exclusively affect the change in surface. Thus, in the difference, the volume work will annihilate.

Van der Waals Interaction between Solute and Solvent

Intending to use SPT as a term for calculating the cavitational energy implies that we still need a term for solvent-solute interactions. When considering the nonpolar part of ΔG^{solv}, solvent molecules do interact with the solute via van der Waals interactions. This bulk water interaction can likewise be formulated as an interaction with a continuum. Huron and Claverie [71] developed a theory for calculating the interactions of a molecule with its whole surrounding. The approach is simple in idea but rather complicated in the mathematical formulation.

Their model starts from defining a potential for the interactions of solvent and solute molecules by considering the atomic interactions. The potential is basically the van der Waals potential introduced earlier, but it defines a different form for the repulsive part. Instead of using a r^{-12} term, this potential, introduced by Kitaygorodski [72], employs an $\exp(-r)$ type function. Let R_i and R_j be the van der Waals radii of atoms i and j, r_{ij} the distance between these two atoms and α a potential parameter, then the energy contribution from these two atoms is defined as

$$e_{ij} = e_{ij}^{\mathrm{dis}} + e_{ij}^{\mathrm{rep}} = -C_{\mathrm{dis}}(4R_iR_j)^3\frac{1}{r_{ij}^6} + C_{\mathrm{rep}}\exp\left(-\frac{\alpha}{\sqrt{4R_iR_j}}r_{ij}\right) \tag{3.34}$$

with C_{dis} and C_{rep} being adjustable coefficients of the model. At this point, the energy of the whole system could be calculated by summing up over all atoms of the system including each atom of every existing solvent molecule.

The potential (3.34) is a so-called 6-exp potential. In this work, however, van der Waals energies are calculated with the help of the Lennard-Jones potential from the AMBER implementation in BALL and its parameters. In order to avoid using an additional set of parameters introducing another source of error and additional effort in implementing the energy calculations, the potential used for nonpolar solvation effects had to be adapted to the form used for the Lennard-Jones potential. Therefore, guided by the calculations of Huron and Claverie, the model was recalculated using the Lennard-Jones potential.

The total van der Waals interaction E^{vdw} between solute and solvent can be defined as a sum of all pairwise van der Waals interactions of the atoms s of all solvent molecules and the atoms m of the solute molecule:

$$E_{M,S}^{\mathrm{vdw}} = \sum_{m\in M}\sum_{s\in S} E^{\mathrm{vdw}}(m,s) \tag{3.35}$$

Using the number density ρ_S of solvent molecules this can be rewritten as a volume integral over the volume V_S occupied by the solvent. In this formulation only the occurring atom types s' are considered. Let S' be the set of all occurring atom types in the solvent, then the energy can be written as

$$E_{M,S}^{\mathrm{vdw}} = \sum_{m\in M}\sum_{s'\in S'} \iiint_{V_S} \rho_S E(m,s')\,dv \tag{3.36}$$

The pairwise interaction energy $E(m,s)$ usually depends on the distance of the two atoms m and s, thus possesses a radial symmetry. With this starting point and the interaction potential

$$E^{\mathrm{vdw}} = \frac{A}{r^{12}} - \frac{B}{r^6}, \tag{3.37}$$

the solute-solvent interaction energy can be calculated by transforming the volume integral into a surface integral. The details of this transformation are given in Appendix C.2.

The complete interaction energy of the solute and the solvent $E_{M,S}^{\mathrm{vdw}}$, can be obtained by summing the individual interaction contributions over all atoms of the solvent and the solute. Let $\mathbf{n}(S_M)$ be the normal of the molecular surface S_M at $\mathbf{r}$, then the energy is given by

$$E_{M,S}^{\mathrm{vdw}} = \sum_{m\in M}\sum_{s'\in S'} \rho_S \iint_{S_M}\left(\frac{A_{m,s'}}{9r^{12}} - \frac{B_{m,s'}}{3r^6}\right)\cdot\mathbf{r}\cdot\mathbf{n}(S_M)\,ds \tag{3.38}$$

The surface integral $\iint_{S_M} ds$ can be approximated easily by summing up the surface areas of all involved atoms. Let A_M be the set containing the normalised surface areas a_m for all atoms $m \in M$. Then the energy is given by

$$E_{M,S}^{\mathrm{vdw}} = \rho_S \sum_{m\in M}\sum_{s'\in S'}\sum_{a_m\in A_M}\left(\frac{A_{m,s'}}{9r^{12}} - \frac{B_{m,s'}}{3r^6}\right)a_m \tag{3.39}$$

This triple sum can be calculated efficiently .

3.4.8. The Theory Behind Polar Solvation Effects

The polar part $\Delta G_{\mathrm{es}}^{\mathrm{solv}}$ of the solvation free energy in equation (3.20) is caused by electrostatic interactions between the atoms of the solute molecule and the atoms of all solvent molecules. In principle, the starting point for a computational formulation is again the Coulomb law introduced in Section 3.4.4. While the Coulomb law was introduced for calculating electrostatic energies in vacuum, from now on the treatment will include effects from a medium, *e. g.* a solvent.

Generally, matter which is permeated by an electrical field will react to it. By induction, the field causes a charge shift, which will lead to the creation of a second field within the medium. This *response* field or *reaction* field is of opposite sign. Because electrical fields are additive, this induced field will dampen the strength of the original field. When molecules are solvated, the strength of electrostatic interactions with surrounding molecules will hence be altered by the reaction field in the solvent.

In order to compute these effects, some theory of electrical fields and potential is necessary. Without going into the details of advanced electrostatics, the equations needed for the calculation of electrostatic effects are introduced. A thorough treatment of this matter is beyond the scope of this work. Every physics text book on electrostatics and electrodynamics, *e. g.* [73], will give more insights into theory and its derivation.

A charge distribution $\rho(\mathbf{r})$, *e. g.* the partial charges of atoms in a molecule, creates the electrical potential $\phi(\mathbf{r})$. Potential and charge distribution describe the electrostatic system. This is formulated in the Poisson equation

$$-\nabla \cdot \nabla \phi(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\varepsilon_0 \varepsilon_r} \tag{3.40}$$

with ε_0 being the vacuum permittivity. The existence of a medium, also called *dielectric*, is captured by the *relative dielectric constant* ε_r which is a material constant describing the dampening effect caused by that material when an electrostatic field is permeating it. This constant is derived from the electric flux density under the assumption that the medium in question is an isotropic, homogeneous dielectric.

Potential and electrical field $\mathbf{E}(\mathbf{r})$ are connected directly via

$$\mathbf{E}(\mathbf{r}) = -\nabla \phi(\mathbf{r}) \tag{3.41}$$

With equations (3.40) and (3.41), every electrostatic quantity of a system can be calculated as long as the dielectric constant is a real constant, *i. e.* the dielectric constant is the same at every point in space. Based on these equations, there are several methods for calculating solvation free energies.

3.4.9. Computational Models

In this study, two elements comprise the calculation of the solvation free energy of a molecule. First, there is the model of Jackson and Sternberg [74], which allows for calculating the influence of solvation effects of the binding of two molecules very accurately by circumventing certain numerical problems. This model relies on methods for doing the actual electrostatics computations. Such a method represents the second element of our solvation model. In this study, two different approaches to calculating electrostatic solvation interactions were investigated: a finite-difference

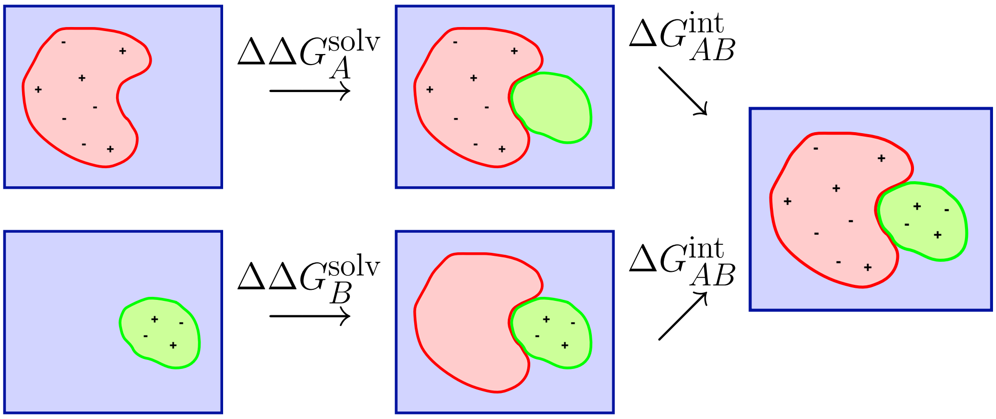

$\Delta\Delta G_A^{\text{solv}}$
$\Delta\Delta G_B^{\text{solv}}$
$\Delta G_{AB}^{\text{int}}$
$\Delta G_{AB}^{\text{int}}$
+
-

Figure 3.14.: The different phases of the binding model of Jackson and Sternberg for calculating the electrostatic free energy on binding. First a molecule is placed in the solvent and its ΔG^{solv} is calculated. Then a "ghost" molecule of the binding partner is inserted that does not carry any charges. Calculating the difference between this system and the one without the uncharged partner results in $\Delta\Delta G_A^{\text{solv}}$ and $\Delta\Delta G_B^{\text{solv}}$, resp. Inserting the charges into the uncharged molecule allows the calculation of the interaction energy $\Delta G_{AB}^{\text{int}}$. The sum of $\Delta\Delta G_A^{\text{solv}}$, $\Delta\Delta G_B^{\text{solv}}$ and $\Delta G_{AB}^{\text{int}}$ comprises the total electrostatic free energy on binding.

solver for the Poisson-Boltzmann equation and the so-called Generalised Born approach. The next section will explain the Jackson-Sternberg framework, followed by a section for each of the electrostatics models.

The Jackson-Sternberg Model

In their study [74], Jackson and Sternberg developed and applied a continuum model for protein-protein interactions. Within the scope of their model, they describe a method for calculating the electrostatic free energy on binding including both solvation and interaction contributions. Like many other models, their method is based on the assumption that the processes of solvating a molecule can be separated into several independent steps. They generalise this idea to the case of binding processes in the solvated phase. The main advantage of the Jackson-Sternberg model is that it is very robust. Additionally, the model achieves high accuracy in calculating electrostatic energies.

Similar to the nonpolar solvation model introduced in Section 3.4.6, the electrostatic free energy on binding $\Delta G_{\text{es}}^{\text{bind}}$ is separated into three independent components. It consists of the change in electrostatic solvation free energy on binding $\Delta\Delta G_A^{\text{solv}}$ and $\Delta\Delta G_B^{\text{solv}}$ of the two participating molecules A and B and the electrostatic interaction energy $\Delta G_{AB}^{\text{int}}$ between A and B. The electrostatic free energy on binding $\Delta G_{\text{es}}^{\text{bind}}$ can then be written as

$$\Delta G_{\text{es}}^{\text{bind}} = \Delta\Delta G_A^{\text{solv}} + \Delta\Delta G_B^{\text{solv}} + \Delta G_{AB}^{\text{int}} \tag{3.42}$$

Figure 3.14 illustrates the different phases. The change in solvation free energy of a molecule is caused by desolvation of parts of the molecule on binding. In the computational model desolvation of A is achieved by introducing a region of the size of the binding partner B into the solvent.

This region is uncharged but contains the dielectric constant of the molecule instead of that of the solvent. Creating this uncharged “ghost” molecule corresponds to the loss of solute-solvent interactions of the first molecule when binding the second. After calculating this energy difference, the charges of molecule B are transfered into the dielectric cavity and the electrostatic interaction energy $\Delta G_{AB}^{\text{int}}$ between the two binding partners can be calculated which is given by

$$\Delta G_{AB}^{\text{int}} = \sum_i q_i \phi_i \tag{3.43}$$

where q_i is a newly transfered charge in the molecular cavity of B and ϕ_i is the potential at the position of q_i generated by the charges of molecule A. The change in solvation energy must be calculated for both molecules because both are partially desolvated. The interaction energy can be calculated by considering the effect of the potential of A on the charges of B or vice versa. Both interaction energy quantities should be of the same value.

Finite-Difference Poisson-Boltzmann Solvers

The Poisson equation (3.40) is a partial differential equation which can be solved with numerical methods. There exist many different implementations of such solvers. Usually, these programmes try to solve the even harder Poisson-Boltzmann equation

$$\nabla \varepsilon \nabla \phi \bar{\kappa}^2 \sinh\left(\frac{e\phi}{kT}\right) = -4\pi\rho \tag{3.44}$$

that includes additional effects of solvated ions (*salt effects*). With a solution to this equation, electrostatic energies are directly computable. Unfortunately, equation (3.44) cannot be solved analytically for molecular systems of the size of proteins, so numerical approaches have to be employed.

Being able to calculate the electrostatics of a system enables us to calculate solvation effects that are caused by electrostatic interactions. As stated in Section 3.4.5, the solvation free energy is

$$\Delta G^{\text{solv}} = G^{\text{solvated phase}} - G^{\text{gas phase}}$$

In order to determine electrostatic solvation free energy, it is necessary to compute both phases. $\Delta G_{\text{es}}^{\text{solv}}$ is the difference of the electrostatic energy G_{es} in gas phase and solvated phase:

$$\Delta G_{\text{es}}^{\text{solv}} = G_{\text{es}}^{\text{solvated phase}} - G_{\text{es}}^{\text{gas phase}} \tag{3.45}$$

Consequently, computational effort doubles for these calculations.

There are several approaches to solving such a differential equation, ranging from boundary element methods over finite elements to finite-difference approaches. The most successful and wide-spread method is the finite-difference Poisson-Boltzmann (FDPB) solver. BALL provides such a FDPB solver which is mainly based on the work by Zhou *et al.* [75] and Bruccoleri *et al.* [76]. The following paragraphs will introduce the methodology to some extent.

Finite-difference methods solve the Poisson-Boltzmann equation through transforming it into a set of difference equations by discretising space into an equally spaced three-dimensional grid, illustrated in Fig. 3.15. Every grid point represents the physical properties of that point in space, such as charge distribution and dielectric constant. These properties have to be mapped from the

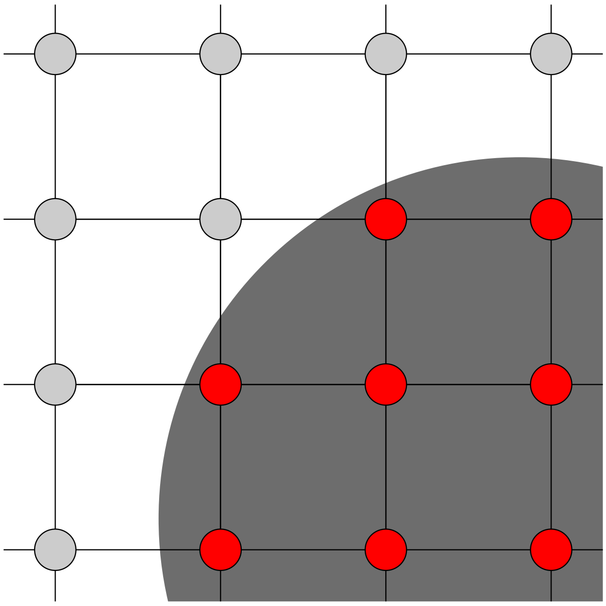


Figure 3.15.: Discretisation of space for FDPB calculations. A cubic grid represents space, depicted here by black lines. The intersection points carry the charge for that spatial grid point (circles). If a grid point is within the van der Waals radius of an atom (dark grey circle), the atom's charge is assigned to that point (red circles).

spatial information given by the molecular data. The solution of the difference equations yield the electrostatic potential. From the potential, field strength and energy can be calculated.

The accuracy of a FDPB solver is tightly connected with the grid spacing chosen for the calculation. The smaller this spacing is, the more accurate is the computation. But there remain some numerical problems. The system contains a certain amount of energy called the *self-energy*. The solvation energy is the energy difference between two states of the system. Consequently, the self energy should vanish in the difference and the solvation contributions should remain. Unfortunately, the self-energy of the system tends to be rather large compared to solvation contributions so efficient elimination of the self-energy is imperative as differences between values of very different magnitude tend to become very inaccurate. The problem of eliminating self-energies was addressed by Zhou *et al.* in [75] and is integrated into the BALL FDPB solver.

Besides numerical problems, modelling physical properties accurately and efficiently is a non-trivial task. The accuracy of the finite-difference approach is greatly influenced by the method of mapping and representing the charge distribution and the (local) dielectric constant onto the grid. Bruccoleri *et al.* proposed in [76] a model employing harmonic smoothing of the dielectric properties of space and antialiasing of the charge distribution. This approach improves the independence of the calculation against changes in the grid positioning and reduces numerical problems introduced by the discretisation.

The basic idea is to "smear" charges across several discretisation points. Instead of assigning a charge to a single grid point, a fraction of a charge is assigned to a grid point. This fraction is determined by the amount of space surrounding that grid point that is occupied by an atom as illustrated in Fig. 3.17. For this determination, space around the grid point is subdivided into

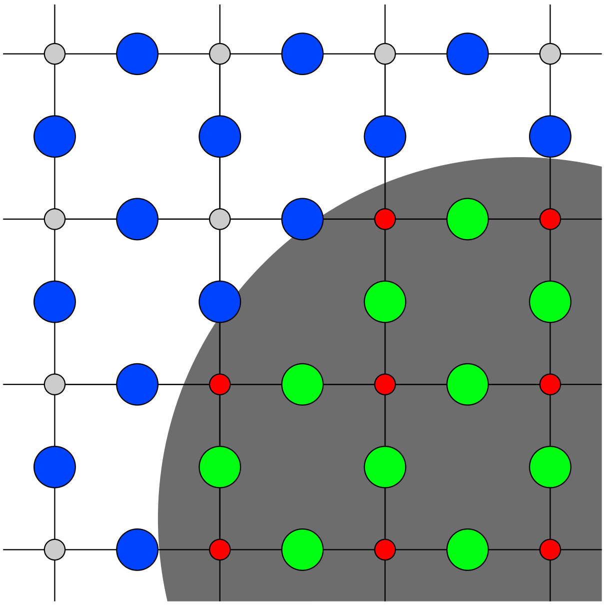


Figure 3.16.: Representation of the dielectric in the spatial grid: The values of the dielectric are given at points (big circles) on the connecting lines between single charge grid points (small circles). The dielectric grid points are assigned with a constant representing the internal molecular environment (green circles) or the surrounding solvent (blue circles) according to their position relative to atom centres.

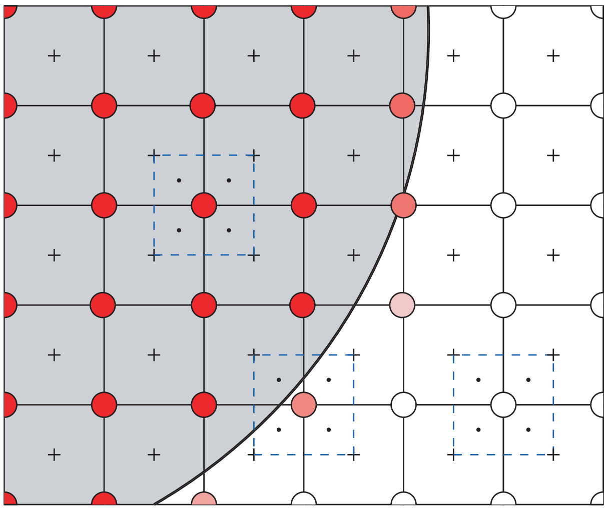


Figure 3.17.: Charge antialiasing: Space around the grid point is further divided into small boxes (blue dashed lines). If the centre of such a small box (small black dots) is within the van der Waals radius of an atom (large grey circle), it is counted as occupied. The number of occupied small boxes around a grid point designates the charge assigned to the grid point.

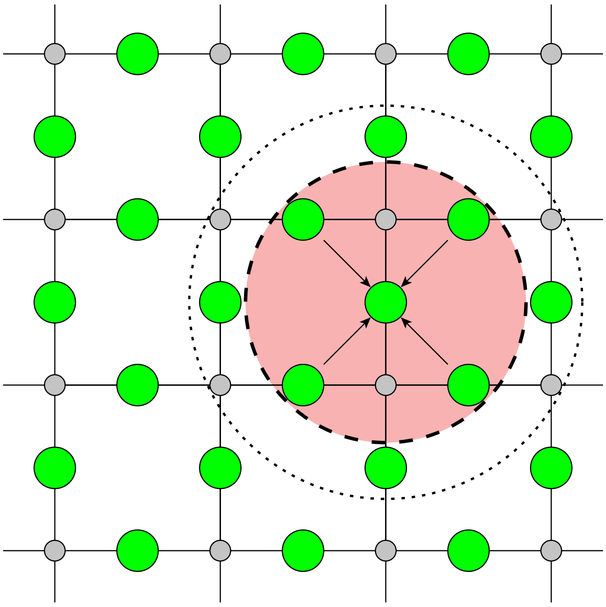


Figure 3.18.: Dielectric smoothing: After assigning the dielectric values, for every grid point, the dielectric value is assigned by computing the mean of the original value of all neighbouring grid points within a predefined radius (dashed red circle).

small rectangular boxes. If the centre of such a small box is within the radius of an atom, it is counted as occupied. The more boxes are occupied, the larger is the charge fraction.

With such a smooth representation of charges, the dependence of the calculated potential on the actual positioning of the grid points relative to the atom centres will be decreased. Smoothing also improves the accuracy of the finite difference method if applied to the dielectric properties of a system. Usually, the dielectric boundaries are clearly defined by the molecular surface. Inside the molecule, the dielectric constant (DC) is ε_i, outside the molecule it is ε_o, which in most cases will be the DC of pure water. The dielectric constant of one grid point can be defined using a volume filtering approach averaging the dielectric constant of sufficiently many neighbouring grid points, depicted in Fig. 3.18. The main idea of this method is to compute average values of the DC around a grid point and to use this mean value in the actual calculation.

FDPB solvers require much time for calculating the electrostatics of systems as large as proteins. In order to reduce the computation time, most models already include cut-off values limiting the number of atom pairs considered interacting. The BALL FDPB implementation already provides such means of acceleration. Nevertheless, some functions still have to traverse large sets of atoms in order to find interacting pairs. Consequently, reducing the size of the actual system speeds up the computations significantly.

For our calculations, further acceleration was desired. A simple way of doing so is the reduction of the molecular system to areas of interest, which basically is the protein binding site. This method is frequently used in docking calculations (cf. FlexX [47]). When cutting out interesting portions of the molecular system, one has to make sure that the remaining part is large enough to cover all relevant interactions. The idea realised here and depicted in Fig. 3.19 works as follows:

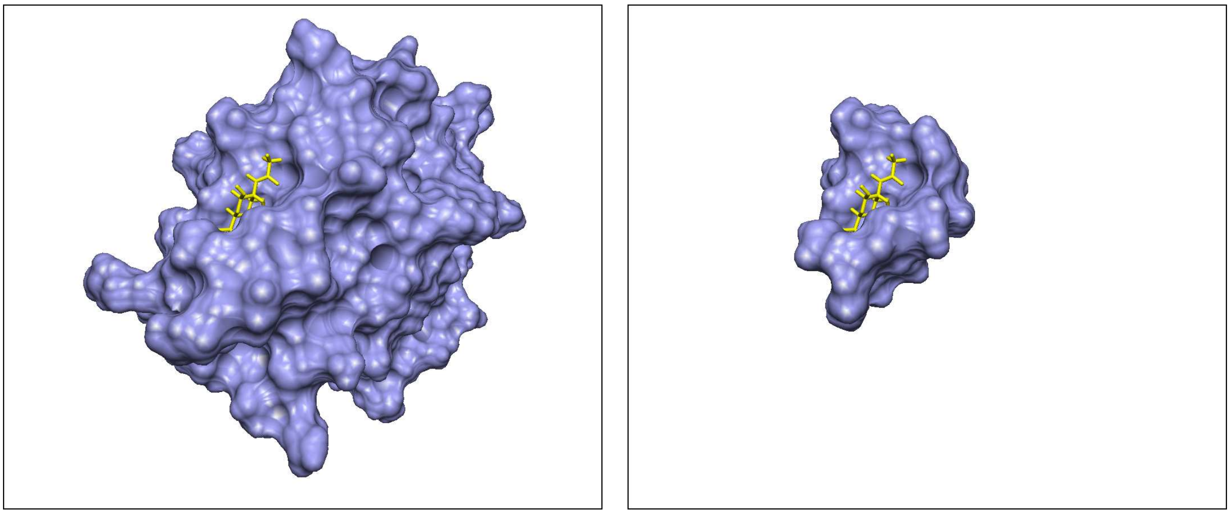


Figure 3.19.: An original system (left, PDB-ID 1AX0) in comparison to the cut-out system (right). Note the significant reduction in size.

1. Calculate a bounding box around the ligand.
2. Extend the ligand bounding box evenly into every direction thus defining the "cut-out" box.
3. Create a new system which only consists of residues which have atoms within the extended bounding box.
4. Cap the residue chains which are not terminated correctly with ACE and NME caps.

The last step is necessary in order to balance charges of the discontinuous backbone. Test calculations showed only small energy deviation of such a cut-out binding site compared to the whole system while decreasing computation times significantly.

The FDPB calculations used in this work were done using harmonic antialiasing of the charges, a trilinear filtering approach for smoothing the dielectric and electrostatic focusing for determining the boundary values of the grid. Spacing of the grid was set to 0.5Å. The extension of the cut-out box was 8Å into every direction in space.

3.4.10. Generalised Born

Besides FDPB approaches, methods based on the Generalised Born (GB) formulation become more popular in molecular mechanics. For this study, several GB models were analysed and the most promising one was implemented. This section covers GB in some detail and introduces into the methods chosen for our calculations.

Generalised-Born models are computational methods that are based on the formulation of the electrostatics of a single spherical solvated ion, which was first published by Max Born in 1920 [77]. In his model, the solvation free energy of the solvated ion is

$$\Delta G_{\mathrm{es}}^{\mathrm{solv}} = -\frac{q^2}{2R}\left(1 - \frac{1}{\varepsilon_w}\right) \tag{3.46}$$

where q is the charge of the ion, R denotes its radius and ε_w is the dielectric constant of the solvent.

The Born approximation yields good agreement with experimentally determined solvation free energies of spherical ions, but for larger non-spherical molecules, a new formulation has to be found. Originally, the generalisation of the Born model was formulated as

$$\Delta G_{\text{es}}^{\text{solv}} = \frac{1}{4\pi\varepsilon_0}\left(1 - \frac{1}{\varepsilon_0}\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{q_i q_j}{r_{ij}} - \frac{1}{8\pi\varepsilon_0}\left(1 - \frac{1}{\varepsilon_s}\right)\sum_{i=1}^{n}\frac{q_i^2}{\alpha_i} \tag{3.47}$$

where the α_i are the so-called *generalised Born radii* of the atoms, which roughly represent the distance of an atom centre to the dielectric surface. These Born radii are crucial for the success of a calculation and can be estimated on the basis of the Coulomb field approximation (see below). Still and coworkers introduced in [78] a more concise formulation of the GB term

$$\Delta G_{\text{es}}^{\text{solv}} = -\frac{1}{8\pi\varepsilon_0}\left(1 - \frac{1}{\varepsilon_s}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_i q_j}{f_{ij}^{\text{GB}}(r_{ij})}, \tag{3.48}$$

which is the form now used in GB calculations. The function$f_{ij}^{\text{GB}}(r_{ij})$ was introduced as

$$f_{ij}^{\text{GB}}(r_{ij}) = \sqrt{r_{ij}^2 + \sqrt{\alpha_i\alpha_j}\exp\left(-\frac{r_{ij}^2}{4\alpha_i\alpha_j}\right)} \tag{3.49}$$

and combines the terms of equation (3.47) in a way such that classical electrostatics is reproduced correctly. This model predicts solvation free energies of small solutes very well. Several modifications of the Still formulation of the GB model have since been developed (*e. g.* [79, 80, 81, 82, 83, 84, 85, 86]), which mostly differ in the estimation of the Born radii.

The GB model has some advantages over the FDPB methods in terms of running time. The computationally most expensive part is the estimation of the Born radii. The actual summation only needs very little resources. Additionally, the solvation free energy can be calculated with one single calculation. Using the FDPB method, two calculations have to be performed, because the solvation free energy is the difference between vacuum and the solvated phase. Thus both cases have to be calculated and subtracted. However, FDPB models still represent the benchmark for electrostatics calculations. The accuracy achieved with FDPB methods is generally better than with GB calculations. Nevertheless, GB models are under constant development because of their speed, which makes them applicable in molecular dynamics simulations.

Onufriev, Case, Bashford and coworkers improved the Still formulation [79, 80, 81, 82] with larger molecules like proteins in mind. In [79] their GB model for macromolecules was presented. It accounts for the fact that large molecules cannot be treated without considering their own dielectric properties. Additionally, their model introduces a modification of the Born radii approximation. In [83] salt effects were included into the Born model, which are captured in the Poisson-Boltzmann equation but were lacking in GB methods. These improved methods will be described here as they build the basis for the GB model implemented for this work.

Estimating Born Radii

The Born radii used in the Still formulation are critical for the accuracy of the method. Many methods use the formalism developed by Schaefer and Froemmel [87], which was improved by

Hawkins and coworkers [84, 88]. The expression for calculating Born radii is based on an approximation of the Coulomb integral, which is directly connected to the Born radius α_i of atom i. Let R_i be the intrinsic (van der Waals) radius of atom i and r_{ij} be the distance between atoms i and j, then

$$\frac{1}{\alpha_i} = \frac{1}{R_i} - \sum_j \int_{R_i}^{\infty} \hat{H}_{ij}(r_{ij}, R_j) \frac{1}{r^2} dr \tag{3.50}$$

where $\hat{H}_{ij}$ is the fraction of the area of a sphere of radius r centred at atom i that is shielded by a sphere of scaled radius $s_{ji}R_j$. The scaling factors s_{ji} were introduced to compensate for overlap effects of the individual spheres, which is not directly covered in the original Coulomb integral approximation. The integral part of this equation can be approximated to a level where analytical treatment is possible leading to a function $H_{ij}(r_{ij}, R_j)$ representing the integral $\int \hat{H} r^{-2} dr$:

$$H_{ij}(r_{ij}, R_j) = \frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{r_{ij}}{4}\left(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}\right) + \frac{1}{2r_{ij}} \ln \frac{L_{ij}}{U_{ij}} + \frac{s_{ji}^2 R_j^2}{4r_{ij}}\left(\frac{1}{L_{ij}^2} - \frac{1}{U_{ij}^2}\right) \tag{3.51}$$

$$L_{ij} = \begin{cases} 1 & \text{if} \quad r_{ij} + s_{ji}R_j \leq R_i \\ R_i & \text{if} \quad r_{ij} - s_{ji}R_j \leq R_i < r_{ij} + s_{ji}R_j \\ r_{ij} - s_{ji} & \text{if} \quad R_i \leq r_{ij} - s_{ji}R_j \end{cases} \tag{3.52}$$

$$U_{ij} = \begin{cases} 1 & \text{if} \quad r_{ij} + s_{ji}R_j \leq R_i \\ r_{ij} + s_{ji}R_j & \text{if} \quad R_i < r_{ij} + s_{ji}R_j \end{cases} \tag{3.53}$$

With equations (3.51) – (3.53) the Born radius of atom i can be written as

$$\alpha_i = \left(\frac{1}{R_i} - \frac{1}{2}\sum_j H_{ij}(r_{ij}, R_j)\right)^{-1} \tag{3.54}$$

The Onufriev Model

The modified GB model of Onufriev *et al.* uses the GB formulation of equation (3.48) and introduces additional empirical constants into the Born radius approximation. Instead of using scaling factors s_{ji} depending on both atoms, Onufriev *et al.* use factors s_j, which only depend on atom j as introduced by Srinivasan *et al.* in [83]. Additionally, a factor λ is introduced to compensate for missing volume caused by the spherical approximation of the molecular surface. Because this scaling tends to overcompensate and thus overestimate radii, the generalised Born radii are slightly reduced by subtracting the constant δ. The introduction of δ was done for practical reasons in order to avoid a full reparameterisation of the model. Finally, atomic radii are reduced by R_0, which is an empirically determined offset marking the beginning of the actual solvent (see [78]). The Onufriev model hence calculates Born radii as follows.

$$\alpha_i = \left(\frac{1}{R_i - R_0} - \lambda \sum_j H_{ij}(r_{ij}, s_j (R_j - R_0))\right)^{-1} - \delta \tag{3.55}$$

Srinivasan and coworkers [83] introduced an effective way of including salt effects into the generalised Born model, a feature that Poisson-Boltzmann solvers naturally possess. Salt effects can be achieved by including the Debye-Hückel screening parameter into the term describing the dielectric effects.

$$\left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_s}\right) \longrightarrow \left(\frac{1}{\varepsilon_m} - \frac{\exp(-\kappa f_{ij}^{\mathrm{GB}}(r_{ij})}{\varepsilon_s}\right) \tag{3.56}$$

Salt effects can thus be easily covered by a GB calculator by simply exchanging the dielectric term with a salt-aware form.

Onufriev *et al.* also give an expression for calculating the electrostatic potential at each atom position. This potential is necessary to compute interaction energies in the Jackson-Sternberg model. The potential $\phi(\mathbf{r}_i)$ at the position $\mathbf{r}_i$ of each atom i is defined as

$$\phi(\mathbf{r_i}) = -\left(\frac{1}{\varepsilon_m} - \frac{\exp(-\kappa f_{ij}^{\mathrm{GB}}(r_{ij})}{\varepsilon_s}\right) \sum_j \frac{q_j}{f_{ij}^{\mathrm{GB}}(r_{ij})} + \frac{1}{\varepsilon_m} \sum_{j \neq i} \frac{q_i}{r_{ij}} \tag{3.57}$$

The empirical parameters used in the implementation created for this thesis were adopted from literature. In the Born radii approximation, the constants are $\lambda = 1.33$, $R_0 = 0.09$Å and $\delta = 0.15$Å. The scaling factors s_j were taken from [83] and are listed in Tab. 3.2. Unfortunately, these scaling factors only include seven elements, which poses a certain limitation.

Element	s_j
H	0.85
C	0.72
N	0.79
O	0.85
P	0.86
S	0.96
Fe	0.88

Table 3.2.: Scaling factors for GB models

Parameters for Electrostatic Contributions

Parameter sets are usually created by fitting computed quantities to experimental ones. It is also possible to use the results of calculations at the quantum mechanical level as a calibration source if these calculations are performed with sufficient accuracy. In general, choosing a parameter set is not trivial because every set is calibrated on its own calibration source and thus the comparison of parameter sets without actually using them in calculations is virtually impossible.

In this study, the parameter set chosen for electrostatic contributions to the binding free energy is the PARSE parameter set developed by Sitkoff and coworkers [89]. These parameters were optimised for reproducing solvation free energies of compounds representing amino acid side chains with the FDPB method. Sitkoff *et al.* compared different existing parameter sets that stem from several wide-spread force fields, among them AMBER [90], CHARMM [91] and OPLS [92], with their own parameter set gained from FDPB calculations on a large set of small compounds. Their results suggest that PARSE predicts solvation free energies with high accuracy, outperforming the other parameter sets. The performance of PARSE in predicting solvation free energies in FDPB calculations was verified in the scope of a student research project [93] which yielded reassuring results considering the choice of the parameter set. Additionally, in contrast to many other force field parameters, PARSE rules are not depending on correct naming or typing of atoms in the system. Thus, molecular data which does not comply to PDB, AMBER or CHARMM naming conventions can easily be parameterised as long as the topological information of the molecule is complete and includes bond orders.

Chemical Group	Atom	Charge
–OH	O	-0.49
	H	0.49
$-NH_2$	N	-0.78
	$H_{1,2}$	0.39
–CONH–	C	0.55
	O	-0.55
	N	-0.40
	H	0.40
...	...	...

Table 3.3.: Example rules from the PARSE parameter set

The PARSE parameter set contains partial charges and radii for atoms of biomolecules. They are assigned to a particular atom by means of rules that are based on whether the atom belongs to a certain chemical group. Rule examples are shown in Tab. 3.3. These rules were integrated into BALL by developing a flexible parsing and assigning mechanism that reads rules from a file and applies these rules to a selectable set of atoms in the system under investigation. See Appendix A for details on the implementation of PARSE rules in BALL.

3.5. Performance Measures

Empirical energy functions rely heavily on experimental data sets. The coefficients of an energy function are fitted to the experimental data, which means that having a high quality data set is imperative for creating energy functions with decent accuracy. In this light, statistical analysis becomes important for the critical evaluation of the gained results.

Statistical methods provide valuable tools for analysing and assessing the quality of predictions produced by computational models. Moreover, the calculations of coefficients for individual energy contributions are most effectively done using statistical procedures. This section will briefly cover the methods employed in this work.

3.5.1. Correlation Coefficients and Co.

Computational methods should reproduce experimental values with high accuracy. The deviation of computed values c_i from the experimental ones e_i should be minimal. A first impression of the quality of a prediction can hence easily be obtained by computing both the average absolute error $\epsilon_a = \frac{1}{n}\sum_n |c_i - e_i|$ and the maximum absolute error $\epsilon_m = \max |c_i - e_i|$. The first number gives an impression of the overall performance of the prediction. The second number reveals the existence of largely deviating outliers, which should be investigated.

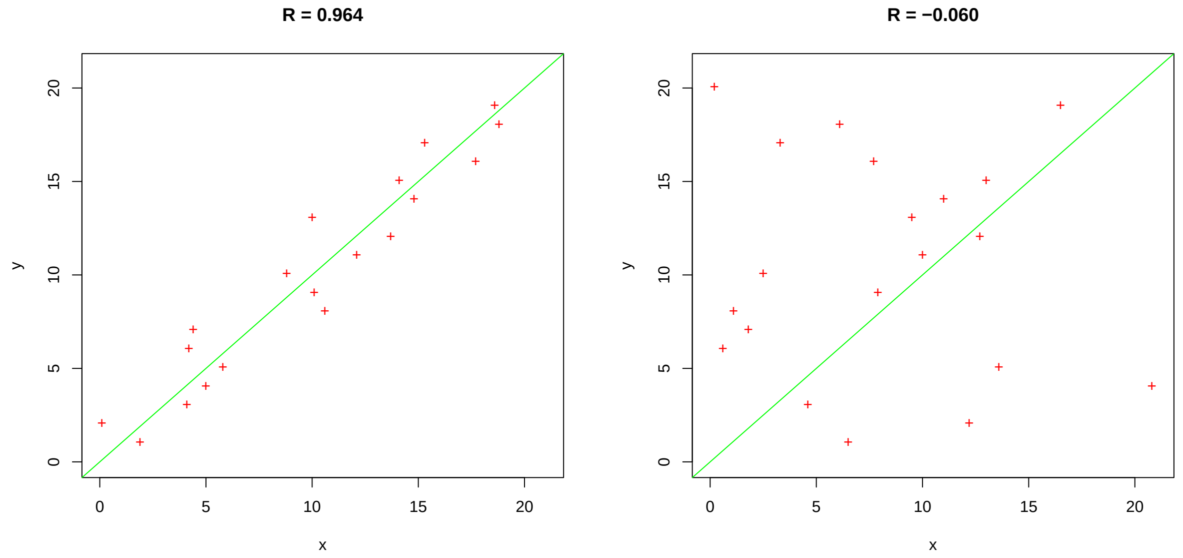


Figure 3.20.: Two exemplary plots illustrating correlation. Left: very good correlation. Right: correlation of a random assignment.

Plotting experimental versus computed values gives a more precise picture. A perfect prediction would yield a graph with all points on the bisector in the first quadrant. An approximative computational model will of course not produce perfect results. Hence a number denominating how well both data sets are correlated would be desirable. A quantity generally considered for this task is the *correlation coefficient* R which is a measure for the linear correlation between two data sets x_i and y_i with N elements and is defined as

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.58}$$

with $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. The values of R lie in the interval $[-1, 1]$. If two data sets are not correlated, R is zero. The more two data sets are linearly correlated, the more the absolute value of R approaches 1. Fig. 3.20 illustrates this behaviour. In the context of energy predictions, a prediction method with high accuracy should yield a correlation coefficient above 0.8.

3.5.2. Multiple Linear Regression

Empirical energy functions like SLICK/energy have to be calibrated with experimental data in order to obtain values for the coefficients of the contributing terms. This calibration is done by fitting. There are several techniques for fitting functions against data sets. Since we assume a linear correlation between predicted and experimental values, the method chosen was multiple linear regression (MLR).

MLR is a generalisation of linear regression analysis that starts from a linear model $y_i = \alpha + \beta x_i + \epsilon_i$ with y_i being the independent variable, which in our case are the experimental values and x_i the dependent variable, the predicted values. The ϵ_i denote the error for each estimate, the so-called *residues*. Linear regression computes the linear best fit by minimising these ϵ_i values. Generally the method used for minimising the residues is *least squares fit*. The optimisation problem to be solved is

$$\min \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - (a + bx_i))^2 \tag{3.59}$$

In the case of ordinary linear regression the coefficients a and b can be calculated directly:

$$b = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{3.60}$$

$$a = \bar{y} - b\bar{x} \tag{3.61}$$

In multiple linear regression the linear model contains M dependent variables per independent variable. Thus, the model has $N \cdot M$ model variables x_{ij} with coefficients a_j and N independent variables y_i. The model then becomes

$$y_i = a_0 \sum_{j=1}^{n} a_j x_{ij} + \epsilon_i \tag{3.62}$$

The method of resolution is basically the same. The residues of the individual estimates have to be minimised, this time by solving a linear equation system. Details on solving MLR can be found in any text book on descriptive statistics.

3.5.3. Cross-Validation

Because the calibration of a prediction method is strongly influenced by the data set, it is necessary to assess the robustness of the method against changes in the calibration data, especially if data sets are rather small. Cross-validation provides a thorough assessment of the susceptibility of the energy function to changes in the calibration data set. The idea is simple: Split the data set into two sets, calibrate it on one set and predict the energies of the second set. From the prediction errors it is possible to gain information about the robustness of the function under investigation.

There are several ways of performing cross-validation. It was chosen here to do full leave-one-out and randomised 5-fold cross-validation. Comparing the results of both methods will give a better basis for assessing the analysis on such a small dataset than relying on just one number. In leave-one-out, the data set is simply reduced by one member of the calibration set. After that, the energy function is calibrated on the remaining data points and the previously removed data point is predicted from the reduced set. This is done for every single data point.

In five-fold cross-validation, the data set is split into five equally large data sets. Every set is then predicted with the energy function calibrated on the data of the four remaining sets and the procedure is repeated for all combinations of data sets (see Fig. 3.21). The data points can be selected exhaustively or randomised, the latter meaning that the composition of the subsets is

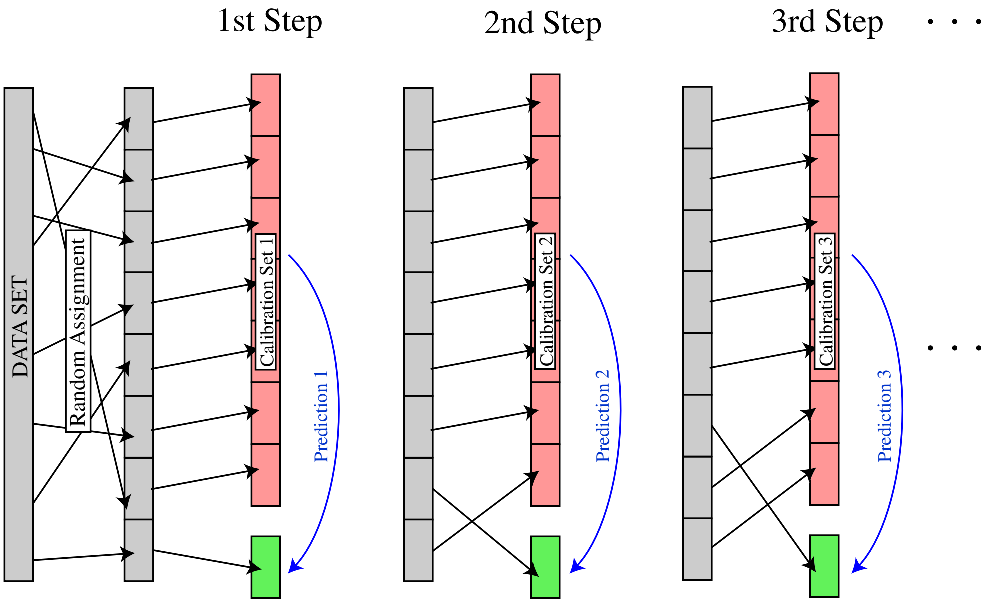


Figure 3.21.: Cross-validation scheme: First, the data set is randomly divided into N equally large data sets. Then, for every subset a prediction is made, which was calibrated on the $N-1$ remaining subsets. The red boxes represent the calibration set. The green boxes are predicted.

chosen randomly. From a statistical point of view, a sufficiently large number of randomly chosen subsets will give satisfactory information on the robustness of the function.

4. Experimental Data

Empirical energy functions have to be calibrated on experimentally measured data in order to produce predictions. In the case of an additive function consisting of a weighted sum of individual terms, the weights of these terms have to be fitted to experimental data. Because an energy function connects the structure of a complex with its binding free energy, the necessary data is two-fold. One the one hand, data on the spatial structure of the complex and on the other hand data on the binding free energy of these complexes is necessary.

Structural data most often stems from X-ray crystallography. For this technique, crystals are grown from solutions of the molecules or complexes in question and then examined with X-ray light. The resulting X-ray spectra, illustrated in Fig. 4.1, can be used to calculate electron densities in space from which the spatial structure of the molecules in the crystal can be concluded. This method has been in use for several decades and yields high quality structures of large biomolecules. The resolution of these structures is usually in the range of several Ångström (Å). Good structures should have resolutions below three Ångström.

Although the quality of X-ray data is very high, there are drawbacks. One drawback of X-ray structures is that hydrogen atoms cannot easily be "seen" in the experiment. Therefore, X-ray structures usually do not contain the hydrogen positions. Besides X-ray crystallography there are other methods for measuring the structure of a biomolecule, *e. g.* Nucleic Magnetic Resonance (NMR), which does provide hydrogen positions but is limited to molecules with molecular weight of less than 40 kDa. The data used here for calibration is purely X-ray data.

Binding free energies of biomolecular complexes are usually measured with isothermal titration

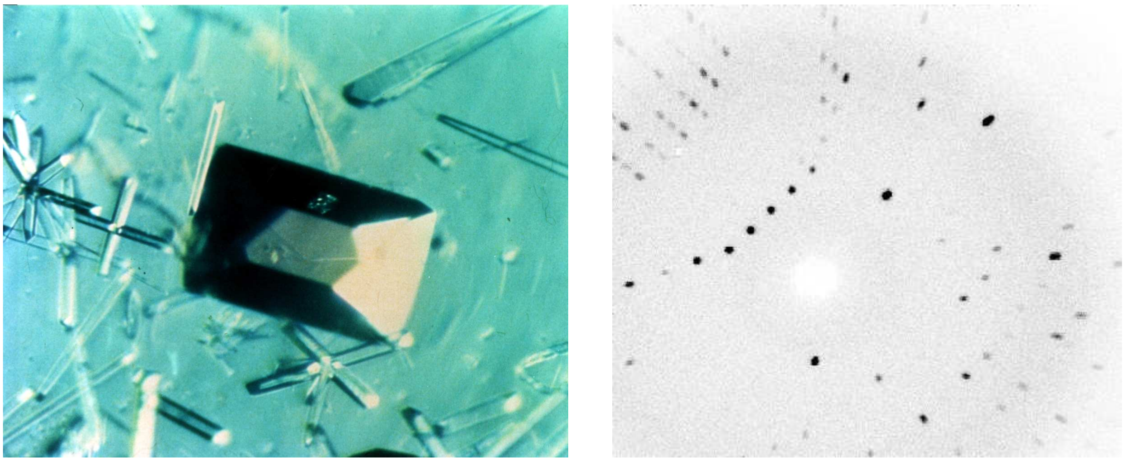


Figure 4.1.: Left: protein crystal for use in X-ray crystallography [94]. Right: X-ray diffraction pattern of myoglobin [95].

calorimetry (ITC). The basic idea is to measure the heat that is absorbed or released during a reaction. From this quantity and the known amount of receptor and added ligand in the experiments, enthalpy H and association constant K_A can be calculated. These can be used to compute the binding free energy ΔG_{bind} of a complex. A review on ITC with protein-carbohydrate complexes is found in [96].

Comparing the experimental environments of the experiments yielding structures on the one hand and binding free energies on the other hand, several discrepancies show up, which have to be kept in mind when analysing the quantities predicted from this data. First of all, the pH of the measurements often differs drastically leading to different protonation states of amino acid side chains. Furthermore, the temperature of the experiments is very different. Generally, X-ray diffraction spectra are generated at very low temperatures while calorimetry usually is conducted at room temperature. And last but not least, X-ray diffraction is done with crystallised proteins while thermodynamics experiments take place in solution.

There are already many lectin-carbohydrate complexes recorded in structural databases, mainly in the Protein Data Bank (PDB) [97]. PDB entries reveal many facts about the experimental setup, like pH, temperature and, most significantly, resolution. Thus, finding high-quality structures is more or less straightforward. Unfortunately this data does not cover binding free energies of the complexes. Consequently, data on the thermodynamics of the binding process from other sources is needed. Such data is not found in databases but has to be obtained from literature. A thoroughly conducted literature search yielded only 18 high-quality structures of plant lectins binding carbohydrates of which binding free energies are of comparable quality.

Only structures of resolution of 3.0 Å and better were accepted. Actual measurement resolutions of our data range between 1.9 and 2.9 Å. Additionally, structures with glycosilations or phosphorylations near the carbohydrate binding site were discarded because of missing parameters for the preparation of crystal structures and the fact that we want to predict lectin-carbohydrate interactions, not carbohydrate-carbohydrate ones. If glycosilations are far away from the carbohydrate binding site, these additional sugars can be removed without risking a distortion of the results.

Complexes which are directly coordinated by metal ions were also discarded, again because of missing parameters as well as uncertainty about the right model for coordinating metal ions in this domain. Nevertheless, there are metal ions present in our data ($\mathrm{Ca}^{2+}, \mathrm{Mn}^{2+}$). These ions play a structurally stabilising role for the protein binding site but do not directly interact with the ligand. Rather exotic ligands like thio-sugars or phosphates were also ignored because of their peculiar chemistry and the resulting parameterisation problems.

The energetic data available was also investigated regarding the data quality. As stated earlier, the difference between binding energies of different lectin-carbohydrate complexes is rather small which made it imperative to particularly analyse the quality of the thermodynamic data that was found in literature.

4.1. Preparation of the Structures

Before being able to use experimentally determined structures, the data has to be checked for errors and then refined for the calculation. Error checking includes searching for disrupted backbones, missing side chain atoms and even completely missing side chains. If there are alternate locations of atoms specified in the structure, one has to decide which location will be used in

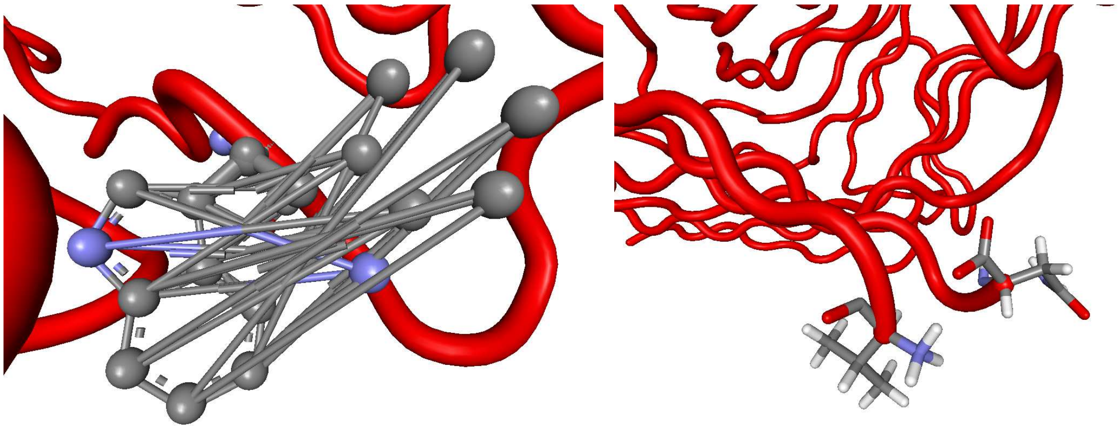


Figure 4.2.: Incorrect structures. Left: incorrectly specified alternate locations in 1DGL. Right: Missing residues in 1BQP. The gap is far away from the binding site. End caps are attached to the backbone ends.

the calculation. This can usually be decided easily by looking at the measured occupancy of the atom. Sometimes, alternate locations are not specified correctly leading to malformed structures as illustrated in Fig. 4.2.

If atom positions are missing in the data, one has to decide whether that data is useful at all. Structures may be acceptable for the calibration set in spite of missing atoms if these errors are far from the protein binding site, because in that case their influence on binding is very small. If missing atoms are very close to the binding sites, the quality of prediction will most probably decrease, even if these missing atoms are added and optimised before doing the calculations. Such structures have to be discarded. However, missing atoms from distantly located residues can be added and optimised without a great danger of decreasing prediction quality because most interactions are short-ranged and even the long-range interactions like electrostatics decline rather sharply. In this context, residues located more than 8 Å away from the binding site are considered distant.

If there are complete amino acids missing (which was encountered twice, see Fig.4.2) that are far enough away from the binding site, there are two options: rebuild the backbone and residues computationally or attach end caps to the open connections of the protein backbone and ignore the gap in the amino acid sequence. Both approaches were tested. The differences were negligible, so the computationally less demanding option can be chosen safely.

Since the structures stem from X-ray crystallography, they lack hydrogen atoms which have to be added to the molecule using computational methods. Adding hydrogens involves optimising the hydrogen positions of the resulting structure, which in turn needs a force field that is able to find energetically favourable positions of these atoms. Therefore a force field is necessary that can cope with carbohydrates and which is also known to create reasonably good results. Since in the subsequent energy calculations parts of the interactions are computed with AMBER terms and parameters, the decision was made in favour of the Glycam parameter set [98], version 2001a for the AMBER94 [90] force field. Glycam is known to reproduce sugar conformations reasonably

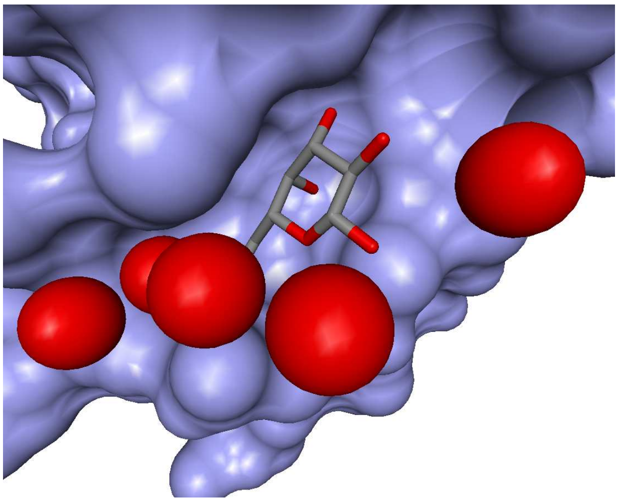


Figure 4.3.: Water molecules in the binding site of ECorL. The molecule in the cleft on the right hand side will be displaced by larger ligands.

well.

Optimising structures with force fields introduces an additional parameterisation and, in the worst case, another source of error. Energy contributions that rely on hydrogen positions like CH$\cdots\pi$ and hydrogen bonding, are clearly directly affected by the results of the optimisation. At the time of preparation, there was no force field available that included CH$\cdots\pi$ terms. Thus the optimisation of the hydrogen positions is possibly not optimal for this investigation.

X-ray structures also include water molecules which sometimes are located in the binding site even within complexes. Most visible water molecules do not participate in binding the ligand to the protein. However, water molecules in close vicinity to the ligand in the active site of the protein are often needed by the complex as a stabilising agent mediating hydrogen bonds between ligand and protein. Although these special water molecules might play an important role in the binding process, all water molecules were discarded from the crystal structures. The decision for doing so was made because water-mediated hydrogen bonds pose a non-trivial problem for docking methods [99] and while there are strategies under investigation [100, 101], there is still no recipe for handling water molecules in docking simulations at hand. Finding such a strategy is beyond the scope of this work.

Some lectins contain glycosilated amino acid side chains far away from the active site as depicted in Fig. 4.4. These glycosilations are removed from the structures. There were also cases where proteins contained oxidised cysteine side chains. These were transformed into standard cysteines. UDA and WGA contain PCA side chains, which do not belong to the standard set of proteinogenous amino acids. Parameters for these side chains were developed and included into the BALL set of AMBER parameters.

Hydrogen atoms are added separately to crystal structures of the protein and the carbohydrate ligand, respectively. The protein's hydrogens are taken from a template library that extrapolates the atom positions from the geometry of an amino acid side chain. This library is part of the BALL framework [102, 103], which will be covered in more detail together with important issues

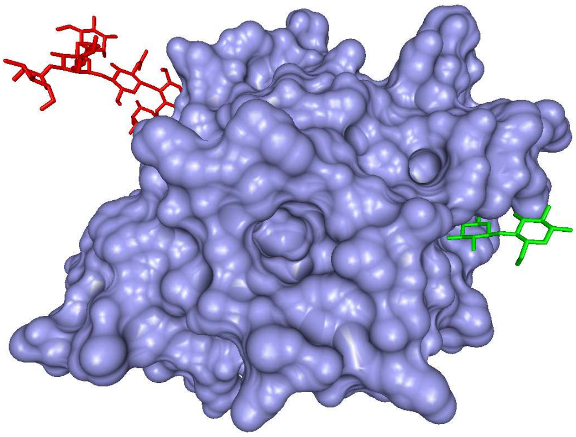


Figure 4.4.: Glycosilations of *Erythrina corallodendron* lectin. The covalently bound glycan (red) is on the far side of the binding site and will not impact the binding mode of the lactose (green). Hence, it can be removed.

of the implementation in Appendix A. Since the templates currently available in BALL only cover proteins and DNA, the carbohydrate hydrogens had to be added with another programme. The Molecular Operating Environment (MOE) [104] of the Chemical Computing Group was chosen for this task. Subsequently, the hydrogen positions were optimised while the heavy atoms of the complex were kept fixed.

Having optimised the missing hydrogen atoms the structures still lack radius and charge parameters for the energy calculations. There is a large number of parameter sets, each customised to reproduce experimental data in different chemical domains. Choosing a parameter set is not a trivial task. In electrostatics calculations, the PARSE parameter set was chosen for complex radii and protein charges (cf. Section 3.4.10 for details). The ligand charges were computed *ab-initio*. The nonpolar part of the solvation contribution was calculated with Bondi [105] radii. Van der Waals calculations were done with Glycam2000a [98] parameters.

For calculating the ligand charges *ab-initio* all ligand structures were rebuilt from scratch using the MOE molecular modelling software instead of using the bound conformation. The geometries of the ligand molecules were then optimised *in vacuo*with the GAMESS programme [106]. Vacuum conformations were used instead of bound conformations in order to avoid a bias introduced by the bound conformation into the data. Charges were then calculated using *ab-initio* methods at the HF6-31G* level, which provides highly accurate results. These calculations were also done with GAMESS.

4.2. Data Sets

There are three different data sets used in this study. The *calibration set* is used for calibrating scoring and energy function. The calibration set consists of 18 structures of plant lectin complexes

Lectin (abbreviation)	Ligand	PDB ID	Main Ref.	ΔG [kJ/mol]	Ref. for ΔG
Artocarpus integrifolia aggl. (AIA)	Me-α-D-Man	1J4U	[107]	-18.24	[108]
Concanavalin A (ConA)	Me-α-D-Man	5CNA	[109]	-22.18	[110]
	Me-α-D-Glc	1GIC	[111]	-19.25	[110]
	Me-3-O-(α-D-Man)-α-D-Man	1QDO	[112]	-28.45	[110]
	Me-6-O-(α-D-Man)-α-D-Man	1QDC	[112]	-22.18	[110]
	Me-3,6-di-O-(α-D-Man)-α-D-Man	1ONA	[113]	-30.96	[110]
Dioclea grandiflora lectin (DGL)	Me-3,6-di-O-(α-D-Man)-α-D-Man	1DGL	[114]	-34.31	[110]
Erythrina corallodendron lectin (ECorL)	Gal	1AXZ	[115]	-18.2	[116]
	GalNAc	1AX0	[115]	-17.9	[116]
	Lac	1AX1	[115]	-18.8	[116]
	LacNAc	1AX2	[115]	-22.7	[116]
Pisum sativum lectin (PSL)	D-Glc	2BQP	[117]	-14.0	[118]
	D-Man	1BQP	[119]	-16.6	[118]
Peanut agglutinin (PNA)	Me-β-D-Gal	1QF3	[120]	-16.96	[121]
	Lac	2PEL	[122]	-17.76	[121]
Urtica dioica agglutinin (UDA)	$(\mathrm{GlcNAc})_3$	1EHH	[40]	-21.34	[123]
	$(\mathrm{GlcNAc})_4$	1EN2	[124]	-23.43	[123]
Wheat germ agglutinin (WGA)	$(\mathrm{GlcNAc})_2$	1K7U	[39]	-21.34	[125]

Table 4.1.: Data set for calibrating and validating the energy function.

Lectin (abbreviation)	Ligand	PDB ID	Main Ref.	ΔG [kJ/mol]	Ref. for ΔG
Human galectin-7 (hGal-7)	Lac	4GAL	[126]	-19.25	[127]
	LacNAc-II	5GAL	[126]	-18.41	[127]

Table 4.2.: Galectin data set for validating the energy function.

with different carbohydrates listed in Tab. 4.1. This set contains structures and energies. For validating purposes, a small *galectin set* was defined. It comprises two complexes of human galectin-7 (cf. Tab. 4.2). This set also contains structures and energies. Finally, a *docking set* was created. The complexes in this set are used in the validation of the docking programme. There are no binding free energies available for these complexes. This set, which is listed in Tab. 4.3, contains 20 plant lectins and non-plant lectins.

For reference, the complete set of receptors is listed by PDB ID in Appendix D. Furthermore, this appendix contains lists with all protein and carbohydrate abbreviations used in this thesis.

Lectin (abbreviation)	Ligand	PDB ID	Main Ref.
Plant lectins			
Allium staving agglutinin (ASA)	α-D-Man	1KJ1	[128]
Artocarpus integrifolia agglutinin (AIA)	Me-Man	1KUJ	[129]
Cratylia mollis lectin	Me-Man	1MVQ	[130]
Psophocarpus tetragonolobus lectin (PTL)	Me-Gal	1WBL	[131]
Robinia pseudoacacia bark lectin (RPbA)	GlcNAc	1FNZ	[132]
Heltuba	Man(1-3)Man	1C3M	[133]
Erythrina crista-galli lectin (ECL)	Lac	1GZC	[134]
Maclura pomifera agglutinin (MPA)	GalNAc-Gal	1JOT	[135]
Viscum album lectin (ML)	Gal	1PUM	[136]
	Lac	1PUU	[136]
Pisum sativum lectin (PSL)	Me-α-D-glucopyranoside	1HKD	[137]
	Man_3	1RIN	[138]
	Sucrose	1OFS	[137]
Non-plant lectins and sugar binding proteins			
Tetanus toxin	Gal	1DIW	[139]
Chemotactic protein receptor	Gal	1GLG	[140]
Anguilla anguilla lectin (AAnA)	Fuc	1K12	[141]
Engineered maltose binding protein	Mal	1NL5	[142]
Human galectin-7 (hGal-7)	Gal	2GAL	[126]
Congerin I	Lac	1C1L	[143]
S-lectin	LacNAc	1SLT	[144]

Table 4.3.: Data set for testing the docking programme.

5. Results

Chapter 3 introduced the computational models necessary for those interactions that are deemed important for protein-carbohydrate interactions, as outlined in Chapter 2. On this basis, the scoring function SLICK/score and the energy function SLICK/energy can now be "assembled" using the different models. This chapter covers the SLICK, its calibration, and its validation. Furthermore, the integration of SLICK into BALLDock and the comparison of the resulting predictions with results from other docking programmes will be shown.

5.1. Scoring vs. Energetic Evaluation

As outlined in the introduction, in many docking methods, two distinct functions are used during the computation, a fast scoring function for filtering putative complex conformations and an energy function for predicting the actual binding free energies of the filtered conformations. Both functions connect a complex conformation with a number. In the first case, this number is a score that indicates how well a certain conformation represents a real binding conformation relative to other putative complex structures. The score alone will not provide enough information about a putative complex without a frame of reference, *i. e.* the other conformations of a docking run. In the second case, this number represents an actual binding free energy, which is an absolute number of unit kJ/mol.

In principle, every energy function should also qualify as a scoring function, because lower binding energies can be seen as better scores. The more energy is liberated during the binding process, the more likely the binding is. But energy functions with the ability of predicting accurate binding free energies often have one critical disadvantage when it comes to filtering a large amount of putative complex conformations. They are simply too slow to be applied to the vast amount of structures. The obvious reason for this lies in the grade of approximation introduced into the function. The stronger the approximations are, the worse are the predictions. But at the same time, stronger approximations can drastically cut the necessary computational effort.

The developers of existing docking programmes use different ways to escape this dilemma. The most trivial approach is to accept long running times and use a high-quality energy function for filtering or scoring purposes. Another solution is to choose an energy function that does not require much computational effort but will not reproduce energies with high accuracy. The widely used FlexX [47] method uses this approach. Its relatively simple energy function, which is based on the Böhm approach [56], produces very reliable filtering results. However, the energies calculated by this function can hardly be taken as real binding free energies. Nevertheless, the docking method achieves good structural results very rapidly.

AutoDock follows a different strategy. They use an energy function for their grid-based approach and precalculate real energies for probe groups on every grid point. During the actual docking, the energy of a ligand is calculated by interpolating between grid points, which needs far less resources than calculating the real function. Again, an approximation is introduced to speed up energy calculations during docking. On an abstract level, this can be seen as transforming the real energy function into a fast and reliable scoring function for filtering out bad candidates. There are numerous other examples for scoring functions, *e. g.* those based on purely geometric considerations in protein-protein docking.

In the case of SLICK, the decision was to create two very similar functions. The scoring function should be able to quickly rank candidates in the correct order. Therefore, computationally intensive models could not be included into the model. The energy function should produce binding energy estimates of high accuracy. Since computing time for the energetic evaluation is not as critical as for filtering, the computational models could be more demanding. The choice of models and the performance of the final functions will be addressed in the next sections.

5.2. SLICK

As mentioned before (Chapter 2), deep insight into the structural interactions is necessary to create a specialised energy or scoring function. Based on the knowledge of the features and peculiarities of protein-carbohydrate binding, two functions were created, one for scoring putative complexes created during the structure generation phase of docking programmes and one for calculating binding free energies of such complexes. The scoring function SLICK/score was designed with efficiency in mind while the energy function SLICK/energy was aimed at reproducing high-accuracy binding free energies. Both functions share the same physical basis, but differ in the accuracy of certain models as well as the treatment of solvation effects. Together they constitute the SLICK package.

As a short reminder: It was pointed out that the complexation of proteins with carbohydrates is presumably driven by the following effects:

- Hydrogen bonds, because of the many freely rotatable hydroxyl groups at each sugar ring
- CH$\cdots\pi$ interactions causing the characteristic ring stacking of sugar rings on aromatic rings in the protein binding site
- Electrostatic interactions between protein and carbohydrate, increased by the many polar hydroxyl groups of the sugar
- Solvation effects, *i. e.* electrostatic and non-polar interactions of the individual molecules and the complex with its surrounding solvent

These different interactions have to be covered at least in part by energy and scoring functions.

Both SLICK functions are empirical functions built from the weighted sum of their respective contributions. Each function has to be calibrated and validated on the domain of molecules they were designed for. In the next sections the functions and their validation will be addressed.

5.3. SLICK/score

The SLICK/score scoring function consists of four terms deemed important for protein-carbohydrate binding and especially for the structural basis of binding modes. It includes hydrogen bonding

(S_{hb}), CH$\cdots\pi$ interactions ($S_{\mathrm{CH}\pi}$), van der Waals energies (ΔG_{vdw}) and electrostatic interaction energies ($\Delta G_{\mathrm{es}}^{\mathrm{int}}$), which are computed with the Coulomb law. The SLICK/score $S(m)$ of a molecular complex m is hence defined as

$$S(m) = s_0 + s_{\mathrm{CH}\pi} S_{\mathrm{CH}\pi}(m) + s_{\mathrm{hb}} S_{\mathrm{hb}}(m) + s_{\mathrm{vdw}} \Delta G_{\mathrm{vdw}}(m) + s_{\mathrm{es}} \Delta G_{\mathrm{es}}^{\mathrm{int}}(m) \tag{5.1}$$

This function was calibrated on a set of docked structures. These docking candidates were obtained by docking all plant lectins from the calibration set defined in Chapter 4. At this stage docking candidates were created with the docking programme AutoDock [5]. AutoDock is a general-purpose flexible ligand docking programme and should provide reasonable candidate structures for protein-carbohydrate complexes, although its energy function is not aware of the peculiarities of these complexes. The incorporation of SLICK/score into a docking programme will be covered later.

Ligands were docked to the proteins using the standard AutoDock 3.0.5 parameters unless otherwise noted. The energy grids were centred on the geometric centre of the respective ligand in the binding site. The grid dimension were 65 x 65 x 65 points with a spacing of 0.375 Å. The energies were scaled employing the free energy model 140n coefficients. Non-polar hydrogens were modelled explicitly. For monomers, the number of individuals and maximum energy evaluations per generation was 60 and 1,800,000, respectively. For dimers, these values were doubled. For each carbohydrate, 200 runs were performed. Each run resulted in a final conformation. All final conformations were employed in the following calculations.

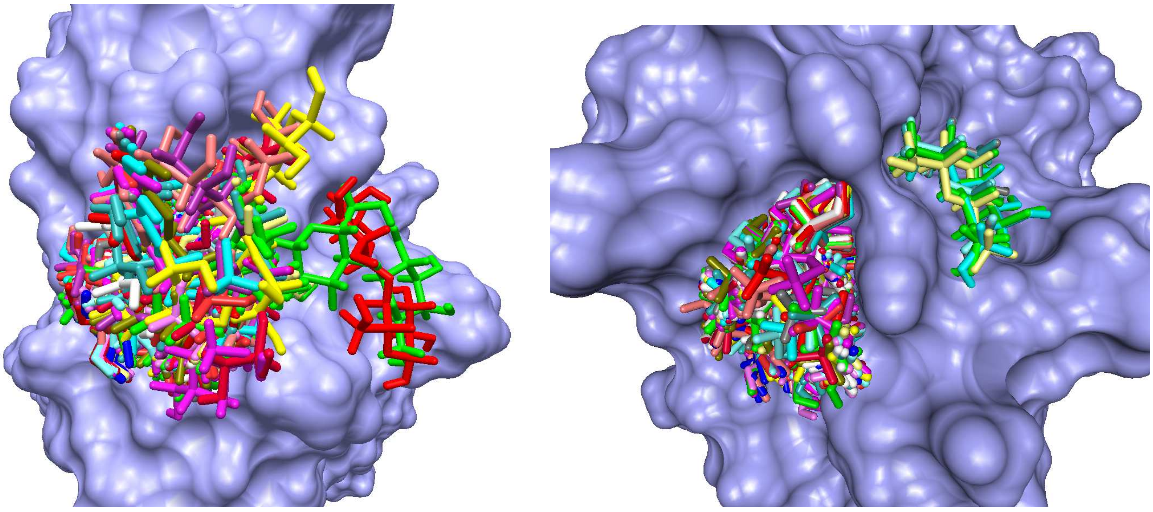


Figure 5.1.: Docking candidates generated by AutoDock covering the binding site of ECorL (left, 1AX1) and PNA (right, 1QF3). There is a large cluster of binding conformations in the binding site, but some structures are misplaced.

The set of candidates created by this procedure covers a sufficient fraction of the conformational space of the ligand in the binding site. Calculation of the RMSD of every individual docking candidate from the crystal structure shows that the range lies between about 0.5 and 13Å. Figure 5.1 shows two examples illustrating the coverage of the binding site. The complete set of RMSD distributions is given in appendix B.1.3.

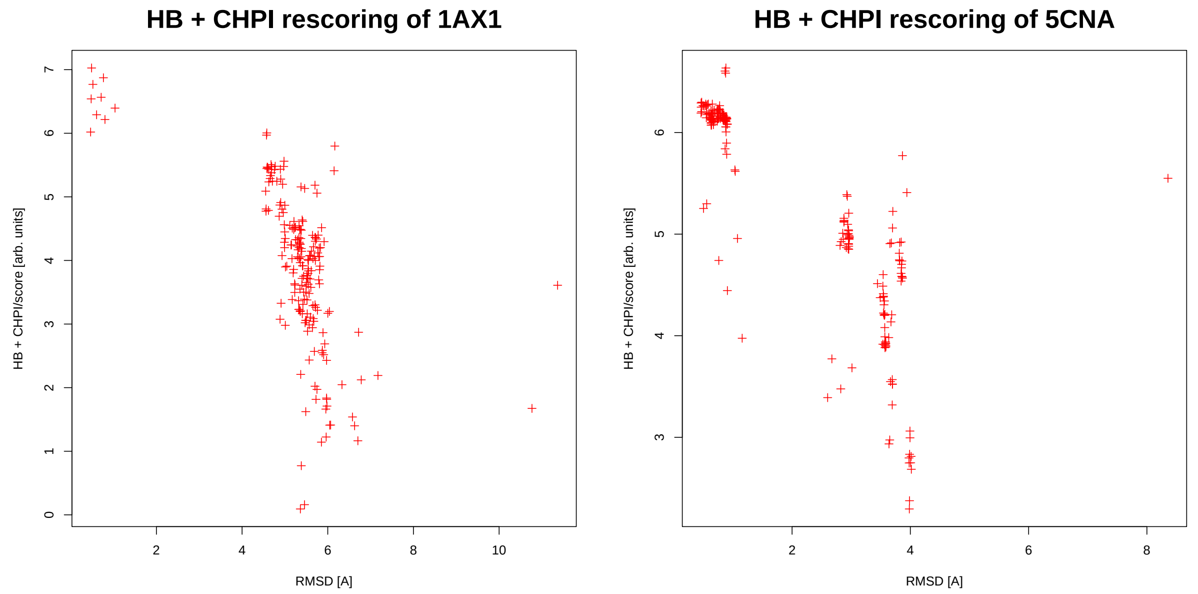


Figure 5.2.: Successful rescoring with CH$\cdots\pi$ and hydrogen bonds only. These graphs show scores (larger values are better).

5.3.1. Training and Validating SLICK/score

The optimisation of SLICK/score coefficients started from the assumption that CH$\cdots\pi$ interactions and hydrogen bonds are very important for the actual binding conformation. By using only these two contributions it was already possible to identify some binding sites of the 18 complexes in the calibration set. Figure 5.2 shows two examples of such successfully identified complexes.

Van der Waals interactions and electrostatics were added and coefficients for these contributions were searched. While the calculations of van der Waals and electrostatic interactions result in energies, the CH$\cdots\pi$ and hydrogen bond terms only give scores that have to be translated into energy-like numbers. Scores are in the range between 0 and 1 depending on how well a found interaction resembles the ideal case. Knowing that hydrogen bonds contribute in the range of 4–20 kJ/mol, the factor -20 was assumed for both contributions. Further investigation showed that the electrostatic interaction seemed a bit overestimated by the Coulomb term included in SLICK/score, which might be a result of the missing dampening effects of the unconsidered solvent. Thus, electrostatic contributions were scaled down. Additionally, all numbers were divided by ten. The final set of coefficients used for rescoring was $s_0 = 0, s_{\text{hb}} = -2, s_{\text{CH}\pi} = -2, s_{\text{vdw}} = 0.1$ and $s_{\text{es}} = 0.08$.

After rescoring the plant lectin complexes with SLICK/score, the root mean square deviation (RMSD) of the ligand's heavy atoms and ranks of the candidates were analysed. Table 5.1 lists the for this analysis. In this context, the most interesting number is the rank of the first true positive (FTP) of the rescored structures, *i. e.* the highest ranked structure with an RMSD. Candidates with an RMSD lower than 1.5 Å are considered as true positives in this analysis. There are two complexes (1QDC, 1K7U) for which AutoDock did not manage to create candidates with an RMSD lower than 1.5 Å. These cannot be included in the analysis of SLICK/score's performance

Complex	R_{ftp}	d_{ftp} [Å]	R_{min}	d_{min} [Å]	n-mer
1J4U	1	1.11	20	0.96	1
5CNA	1	0.46	1	0.46	1
1GIC	1	0.59	1	0.59	1
1QDO	2	1.49	2	1.49	2
1QDC	–	–	4	2.19	2
1ONA	1	1.14	2	1.12	3
1DGL	26	1.04	30	0.62	3
1AXZ	1	0.78	29	0.66	1
1AX0	1	0.89	76	0.70	1
1AX1	1	0.60	19	0.46	2
1AX2	25	1.01	28	0.89	2
2BQP	2	0.45	6	0.42	1
1BQP	1	0.78	76	0.71	1
1QF3	5	0.88	62	0.73	1
2PEL	34	1.06	70	1.01	2
1EHH	30	1.47	30	1.47	3
1EN2	1	1.25	8	0.88	4
1K7U	–	–	8	2.02	2
mean	8.31	0.94	28.75	0.82	–

Table 5.1.: Rescoring AutoDock candidates with SLICK/score: R_{ftp} is the rank of the first true positive, d_{ftp} is its RMSD from the crystal structure, R_{min} is the rank of the candidate with minimal RMSD, d_{min} is its RMSD. The first true positive is the highest ranking structure with an RMSD below 1.5 Å. Docking runs without candidates below this limit are not included in the mean of R_{ftp} and d_{ftp}.

without distorting the results and were therefore discarded.

Table 5.1 lists the scores together with the rank of the candidates with lowest RMSD. In nine cases, the first true positive (FTP) was at the same time the highest ranked structure. In three cases, the FTP was found under the top five ranked structures. In the remaining four complexes, the FTP was scored badly. The mean rank of the FTP is quite high at 8.31, which is caused by the four badly scored complexes. Without these complexes, the mean rank becomes 1.50. The mean RMSD of all FTPs is at 0.94 Å while the mean RMSD of those structures with lowest deviation is only slightly better at 0.82 Å. Unfortunately, some top ranked structures show strong deviations. Fig 5.3 shows four exemplary RMSD plots. All RMSD plots are given in appendix B.1.1.

In order to explain the deviations, the results were investigated structurally and individual contributions to the overall score were analysed. A closer look at Tab. 5.1 reveals a correlation between rescoring quality and size of the ligand. With one exception (1QF3), all monomers were scored very well. Larger sugars seem to be harder to predict in terms of RMSD. Analysing the actual predicted binding poses reveals an important fact. Dimers like Lac, which consists of Glc

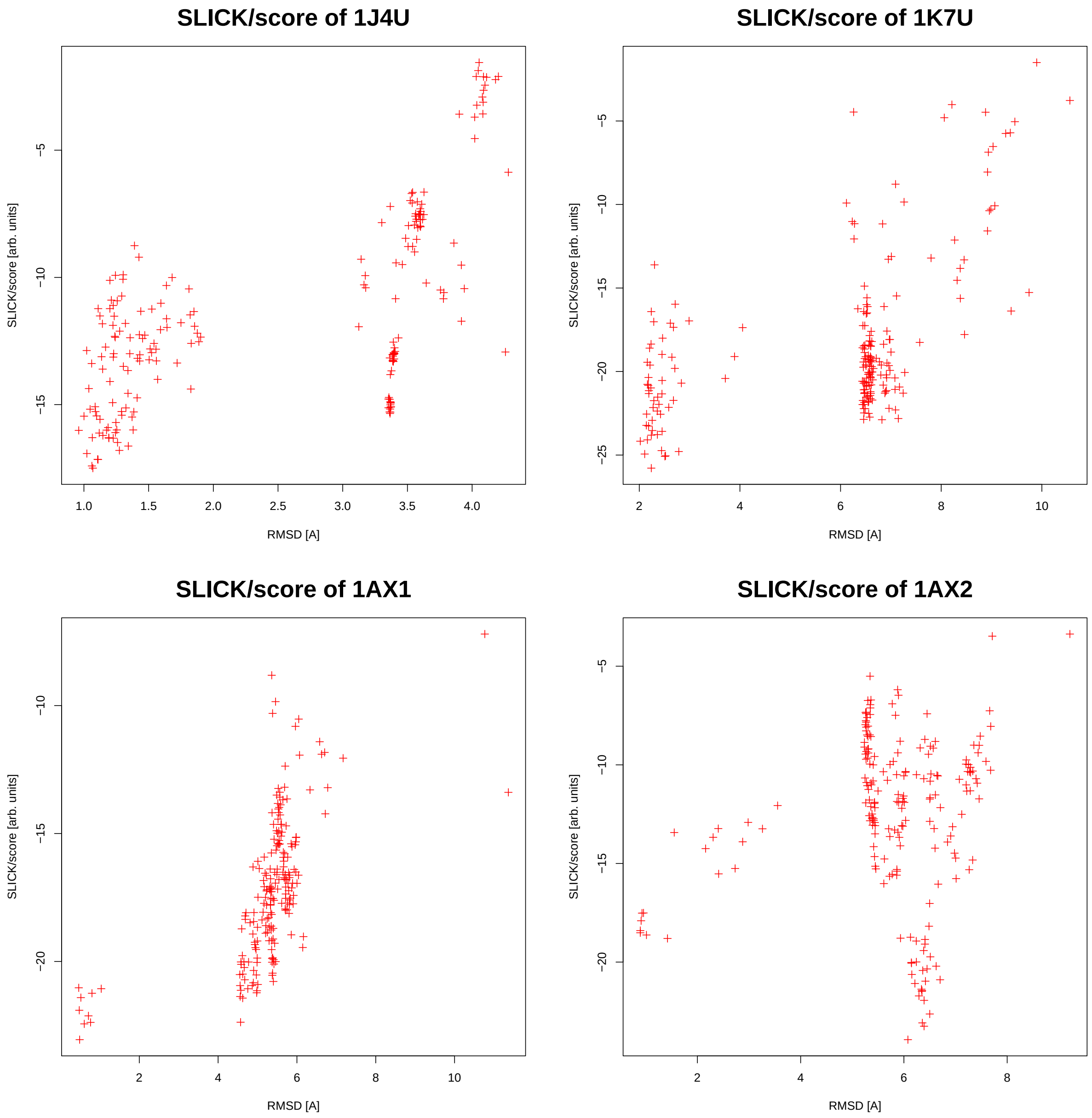


Figure 5.3.: Rescoring results: Exemplary plots of the candidate RMSD versus SLICK/score. For 1J4U and 1K7U the plots indicate SLICK/score's ability to identify good approximations of the real complex. In 1AX1 and 1AX2 the results differ although the ligands are only differing in the NAc substituents

and Gal, tend to bind tightly with one sugar ring into the binding site while the second ring is directed into the solvent (shown exemplary in Fig. 5.4). This behaviour revealing a "pivotal" ring, *i. e.* one guiding sugar ring, is observable in 1AX1, 1AX2, 2PEL, 1QDO and 1QDO. The second ring seems to build hydrogen bonds to water molecules that are located close to the binding

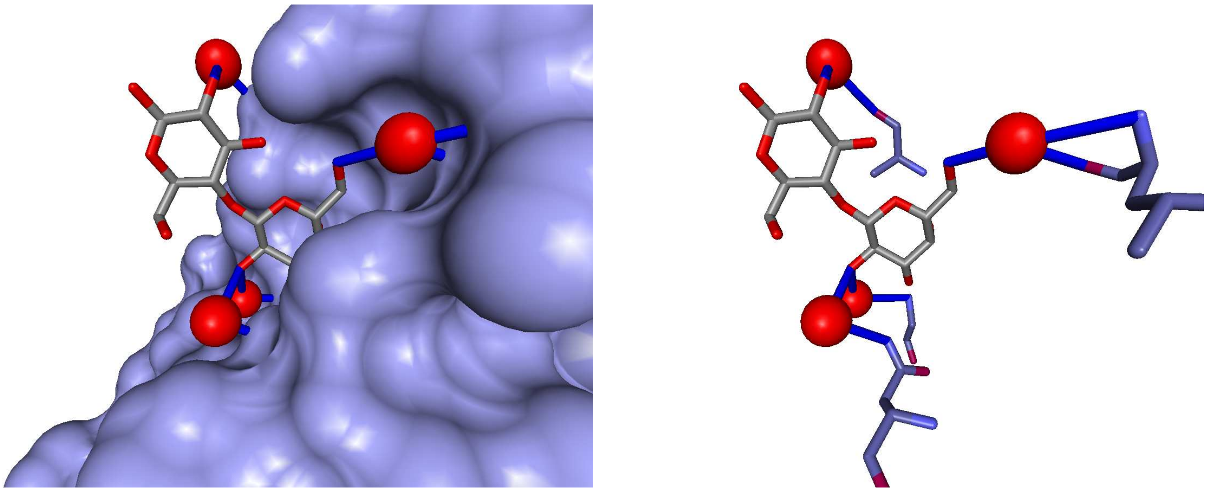


Figure 5.4.: Lac binding to ECorL: Gal binds deeply into the binding pocket while Glc is directed into the solvent building water bridges. Left hand side: the ligand in the binding site with the presumed water bridges. Right hand side: the participating side chains in the lectin.

site. Some of these waters even form hydrogen bonds to the protein, thus establishing a water mediated hydrogen bond between ligand and protein.

The X-ray data available for the ECorL/Lac complex (1AX0) demonstrate the importance of water molecules very well. In Fig. 5.4, the water molecules found in the X-ray data as well as possible hydrogen bonds between water, ligand and proteins are shown. Obviously these water bridges, which are neither modelled by SLICK/score nor by AutoDock, contribute significantly to the energy and thus influence the binding pose. Additionally, the temperature factors of these water molecules near the binding sites indicate a high probability to observe water molecules at these positions. Since AutoDock does not take these water molecules into account, positioning the second, water-surrounded sugar ring is rather difficult. Hence, scores of such complexes tend to be worse than expected.

Another interesting question is whether there are contributions which influence the quality of the scoring more than other terms. Therefore, the influence of every scoring contribution on every complex was analysed. The bottom line of that analysis is that every term is strictly necessary. For every energy term there is at least one example where this energy term performs well and another where it performs poorly. There are complexes that are dominated by hydrogen bonds and/or CH$\cdots\pi$ interactions but generally, all contributions are necessary. In most cases it is exactly the balanced combination of the terms that produces reasonable rescoring results. Fig. 5.5 shows some examples comparing the influence of the individual contributions on the overall score. The first row of plots in Fig. 5.5 shows the AIA/Me-Man complex (1J4U). This lectin does not have any aromatic rings in the binding site, which is why the CH$\cdots\pi$ score is 0 for every candidate. The plots reveal that for this complex hydrogen bonding is of great importance, which corresponds exactly to the data found in the crystal structure. Additionally, the van der Waals energies show the correct tendency for giving better candidates better scores. The electrostatics term seems to favour other positions, though. Because electrostatic contributions are scaled down and hydrogen bonding is weighted quite strongly, SLICK/score is able to identify

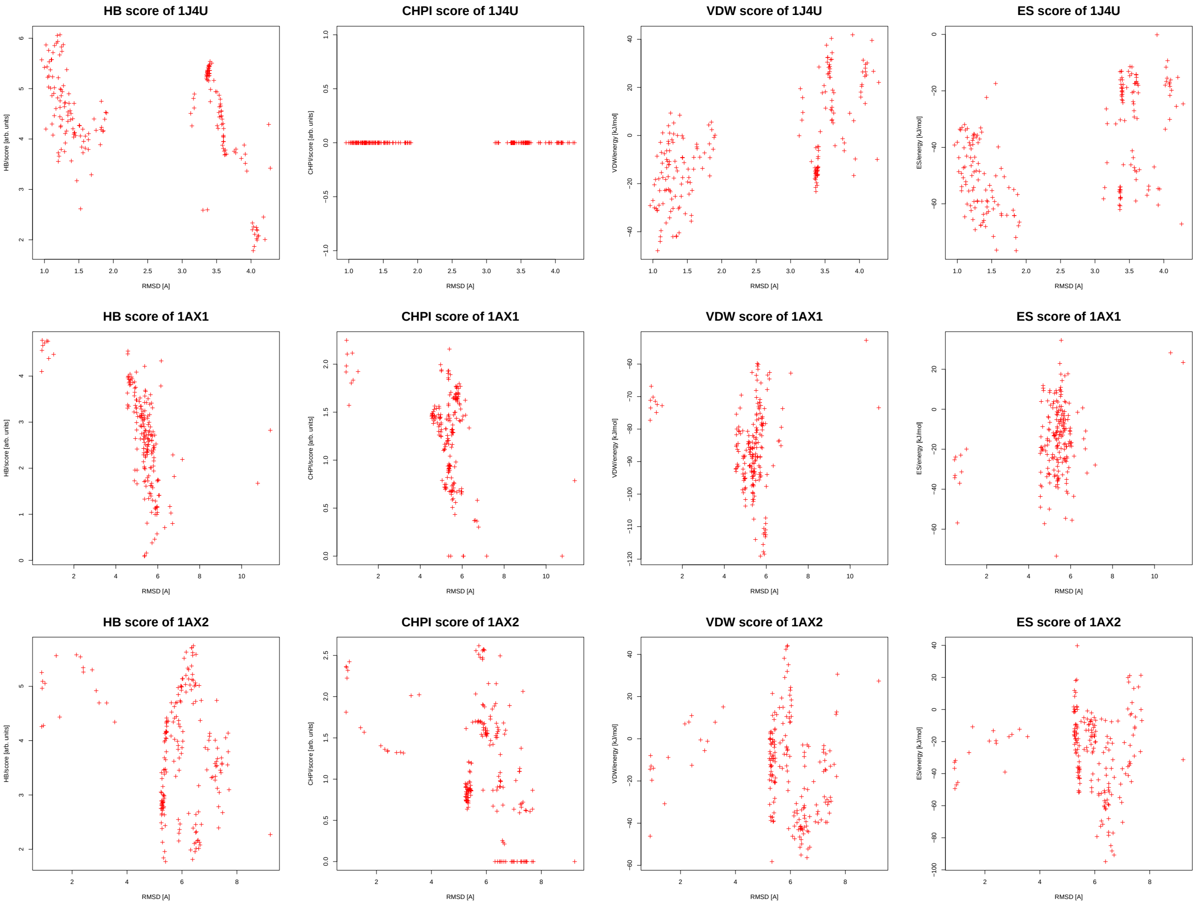


Figure 5.5.: Influence of individual scoring terms on the overall score. The contributions are ordered as follows. First column: hydrogen bonds, second column: CH$\cdots\pi$, third column: van der Waals, fourth column: electrostatics. Note that hydrogen bonds and CH$\cdots\pi$ contributions are scores (bigger is better) and van der Waals and electrostatics are energies (lower is better). Details are given in the text.

the correct binding conformations in this case.

The second row shows the individual contributions to the scoring of the complex of ECorL with Lac (1AX1). In this case, hydrogen bonds and CH$\cdots\pi$ clearly dominate the binding conformation, although both contributions also score a conformation that deviates by 6 Å. The van der Waals scores are bad, while electrostatics do at least not worsen the overall score. Both contributions seem to favour the other position, which surely is the reason why AutoDock generates many candidates in that cluster. In the third row, the picture gets worse. This complex is ECorL/LacNAc (1AX2), which is almost identical to ECorL/Lac (1AX1). The ligands differ only in the NAc group. Here the hydrogen bond and CH$\cdots\pi$ term also score the distorted conformation very high, thus being unable to compensate for the bad scoring of van der Waals and electrostatics.

The reason for the “bad” van der Waals and presumably the electrostatics scoring is the Glc ring of Lac and LacNAc, which in the crystal structure is almost completely surrounded by

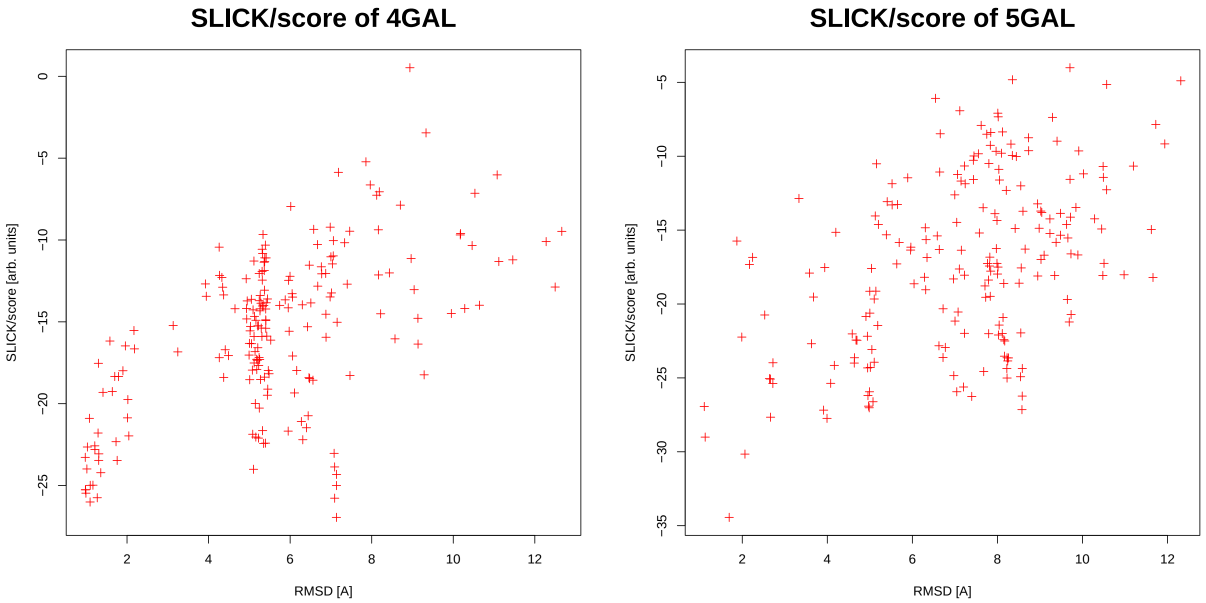


Figure 5.6.: Results of rescoring two galectin complexes with SLICK/score

the solvent. AutoDock tries to maximise van der Waals contacts and thus creates conformations where Glc lies directly on the surface of the protein. These conformations produce a large number of favourable van der Waals contacts dominating the overall score. In addition, the proximity of the ligand atoms increases the energy of the electrostatic interactions and the probability of finding CH$\cdots\pi$ and hydrogen bonds.

5.3.2. SLICK/score and Non-Plant Lectins

Although it could be shown that the scoring function works on the plant lectins of the calibration set, its performance on anything else is still elusive. So the test case for showing SLICK/score's ability was chosen to be human galectin-7 (hGal-7) in complex with either Lac or LacNAc. Both complexes are available through PDB (4GAL and 5GAL) with reasonably high resolutions. Additionally, binding free energies of these complexes were reported in [127].

Docking candidates for hGal-7 with its ligands were created in the same way as for the plant lectin rescoring. Again, the 200 candidates obtained from AutoDock were rescored with SLICK/score. The results are shown in Fig.5.6. As for hGal-7 binding with Lac, SLICK/score is able to identify the binding site, although some conformations with higher RMSDs are ranked better than the actual binding conformation. Structural analysis shows that in this case a similar behaviour as in ECorL/Lac complexes is observable. The Glc ring of Lac is extended into the solvent in the crystal structure. The docking programme does not include solvation effects nor water bridges. Consequently, conformations that lie on the protein surface are scored higher than those with the second ring sticking out into the water. The results of the hGal-7/LacNAc complex are better but still the tendency is the same.

5.3.3. Comparing SLICK/score with AutoDock Energies

The AutoDock programme does not distinguish between scoring and energy function. Therefore, the energies calculated by AutoDock should also be applicable as a filtering criterion. In this section, the fitness of the AutoDock energy function as a filter for protein-carbohydrate interactions will be addressed and compared to the results from SLICK/score rescoring. The complete set of RMSD plots of the complexes ranked with AutoDock is given in appendix B.1.2. Tab. 5.2 lists the respective numbers including highest ranks, best RMSDs and difference between best-scored RMSD and best RMSD.

Complex	R_{ftp}	d_{ftp} [Å]	$R_{\min}$	$d_{\min}$ [Å]	n-mer
1J4U	3	1.41	51	0.96	1
5CNA	1	0.89	90	0.46	1
1GIC	1	0.90	120	0.59	1
1QDO	(175)	(1.49)	175	1.49	2
1QDC	–	–	169	2.19	2
1ONA	3	1.14	16	1.12	3
1DGL	9	1.04	46	0.62	3
1AXZ	1	1.20	139	0.66	1
1AXO	1	0.88	111	0.70	1
1AX1	78	0.49	99	0.46	2
1AX2	93	1.01	106	0.89	2
2BQP	73	0.47	86	0.42	1
1BQP	40	0.81	79	0.71	1
1QF3	54	0.98	107	0.73	1
2PEL	(119)	(1.18)	166	1.01	2
1EHH	91	1.47	91	1.47	3
1EN2	15	1.43	34	0.88	4
1K7U	–	–	44	2.02	2
mean	33.07	0.91	94.75	0.82	–

Table 5.2.: AutoDock results: Please find the description of the columns in Tab.5.1. Complexes with an R_{ftp} above 100 are treated as unsuccessful docking runs and are not included in the average of R_{ftp} and d_{ftp}. These numbers are given in brackets.

Comparing Tab. 5.2 with Tab. 5.1, the first striking observation is the high mean rank of the first true positive using the AutoDock energy function as scoring method. While the RMSDs of the first true positives are comparable to those of SLICK/score, the binding conformations clearly cannot be identified with the AutoDock energy function. The plots of the candidate RMSDs against AutoDock energies (cf. Appendix B.1.2) confirm this observation.

Summarising the results gained so far, the performance of SLICK/score for rescoring structures created with a docking programme is satisfying. Binding conformations are well identified with the exception of few problem cases, which presumably are a consequence of the difficulties of the underlying docking programme in coping with protein-carbohydrate complexes. Analysis indicates that these cases on the one hand clearly result from the inadequate energy function used during structure generation. On the other hand, water bridges seem to play an important role. The former problem can be solved by integrating SLICK/score into a docking programme, which will be addressed later in this thesis. The latter is a problem of the structure generator and not of the scoring function.

5.4. SLICK/energy

Having a decent scoring function for filtering docking candidates is one step towards the creation of reasonable docking results. But we still need a function for predicting real binding energies which is the purpose of the energy function SLICK/energy. It basically has to cover the same interactions as SLICK/score does. Consequently, SLICK/energy is in some sense the "tougher" version of SLICK/score incorporating hydrogen bonds (S_{hb}), CH$\cdots\pi$ interactions ($S_{\mathrm{CH}\pi}$), van der Waals energies (ΔG_{vdw}) and electrostatic interactions ($\Delta G_{\mathrm{es}}^{\mathrm{int}}$). In contrast to SLICK/score, SLICK/energy also covers solvation effects ($\Delta G_{\mathrm{np}}^{\mathrm{solv}}$ and $\Delta G_{\mathrm{es}}^{\mathrm{solv}}$). The binding free energy ΔG is calculated by SLICK/energy as

$$\Delta G = c_0 + c_{\mathrm{CH}\pi}\Delta G_{\mathrm{CH}\pi} + c_{\mathrm{hb}}\Delta G_{\mathrm{hb}} + c_{\mathrm{vdw}}\Delta G_{\mathrm{vdw}} + c_{\mathrm{np}}\Delta G_{\mathrm{np}}^{\mathrm{solv}} + c_{\mathrm{es}}(\Delta G_{\mathrm{es}}^{\mathrm{solv}} + \Delta G_{\mathrm{es}}^{\mathrm{int}}) \tag{5.2}$$

with hydrogen bonding, CH$\cdots\pi$, and van der Waals terms being the same as in SLICK/score. The solvation effects and electrostatics are covered by a nonpolar solvation term $\Delta G_{\mathrm{np}}^{\mathrm{solv}}$ calculating interactions between molecules and solvent that are not caused by electrostatic effects. The polar solvation term represents the electrostatic interactions between molecules and surrounding solvent $\Delta G_{\mathrm{es}}^{\mathrm{solv}}$. The electrostatic interactions between protein and carbohydrate are covered by $\Delta G_{\mathrm{es}}^{\mathrm{int}}$.

All the effects caused by electrostatics, *i. e.* $\Delta G_{\mathrm{es}}^{\mathrm{solv}}$ and $\Delta G_{\mathrm{es}}^{\mathrm{int}}$, are calculated with the Jackson-Sternberg model [74] which was introduced and discussed in Section 3.4.9. The different polar interactions can be computed by using a Finite-Difference Poisson-Boltzmann solver (FDPB) or a generalised Born model (GB), which is a user-selectable option. In practise, FDPB results are superior to GB results in terms of prediction accuracy. Nevertheless, given its drastically shorter computation times, GB still performs reasonably well. The results presented in this thesis were completely calculated using the FDPB approach except where stated otherwise.

In the first SLICK/energy model, nonpolar solvation effects were calculated with the surface tension approach. The results were already encouraging, but improvement seemed possible. The next step was to include the more sophisticated SPT approach for the cavitational part of the solvation free energy and a Huron-Claverie term for the van der Waals interactions between solvent and solute. Unfortunately, practical application of this elaborate model showed that the Huron-Claverie term does not contribute significantly to the overall performance. In fact, coefficients obtained by multiple linear regression were too small to justify inclusion of this term into the calculation at all. It seems that the change in solvent-solute van der Waals interactions on binding is too small to contribute to the binding energy of small ligands like the ones under consideration here. Consequently, the nonpolar part of the change in solvation free energy on binding is only calculated with the SPT approach. Nevertheless, using SPT instead of surface tension models improved results noticeably.

5.4.1. Calibration and Statistical Validation of SLICK/energy

SLICK/energy was calibrated by fitting the predicted binding free energies against experimentally determined energies taken from literature. Please see Chapter 4 for experimental data and their sources.

The coefficients of SLICK/energy were fitted with multiple linear regression (MLR) models introduced in Section 3.5. The MLR was computed with the statistical software package R [145] assisted by the RPy software [146]. For the individual contributions of SLICK/energy using FDPB and GB electrostatics, MLR yielded the coefficients listed in Tab. 5.3.

Method	c_0	c_{hb}	$c_{\mathrm{CH}\pi}$	c_{vdw}	c_{np}	c_{es}
SLICK/energy (FDPB)	-2.72	-1.31	-0.74	0.022	0.50	-0.12
SLICK/energy (GB)	-2.51	-1.25	-0.65	0.017	0.41	-0.05

Table 5.3.: Coefficients of SLICK/energy obtained by MLR using FDPB and GB electrostatics

These coefficients emphasise the importance of solvation effects for calculating binding free energies of lectin sugar complexes. Nonpolar energies, which generally are energetically less important in terms of absolute numbers, are weighted quite strongly, while van der Waals energies, which tend to give larger absolute values, are scaled down. The electrostatics component contains interaction and solvation effects, which presumably is the reason for the negative coefficient. Favourable interactions seem to be compensated by solvation effects. Reducing the electrostatics part to interaction energies significantly worsens the prediction quality. It is possible that nonpolar effects are overestimated which is then compensated by the negative sign of the polar solvation contribution. The reasons for this behaviour are still elusive and should be addressed in further research.

The terms covering hydrogen bonding and CH$\cdots\pi$ interactions are scoring terms resulting in positive scores. As expected, These are contributing to the binding free energy as expected with a negative coefficient. The absolute coefficient of the hydrogen bonds term is larger than the CH$\cdots\pi$ coefficient, which is consistent with the fact that hydrogen bonds are known to be much stronger than CH$\cdots\pi$ interactions.

Plotting the predicted versus the experimentally determined energies, the graphs in Fig. 5.7 are obtained. Using FDPB electrostatics, the correlation coefficient of the calibration is 0.95, the maximum absolute error of the prediction is 3.13 kJ/mol while the average absolute error is 1.27 kJ/mol. With GB electrostatics, these numbers do not change much. The correlation coefficient of the GB version is 0.94, maximum absolute error is 3.33 kJ/mol and average absolute error stays at 1.27 kJ/mol. These encouraging numbers (listed in Tab. 5.4) indicate that SLICK/energy permits prediction of binding energies with high accuracy. Thereby FDPB method seems to perform slightly better. To ensure the robustness of the energy function, statistical methods will be employed.

The statistical validation consisted of an extensive cross-validation of the results. The robustness of the energy function against changes in the calibration set was assessed by a full leave-one-out (LOO) cross-validation and randomised 5-fold cross-validation. The randomised approach was averaged over 1000 runs. SLICK/energy achieved a mean absolute error of 2.1 kJ/-mol in LOO and 2.0 kJ/mol in average on the randomised 5-fold cross-validation. The histogram

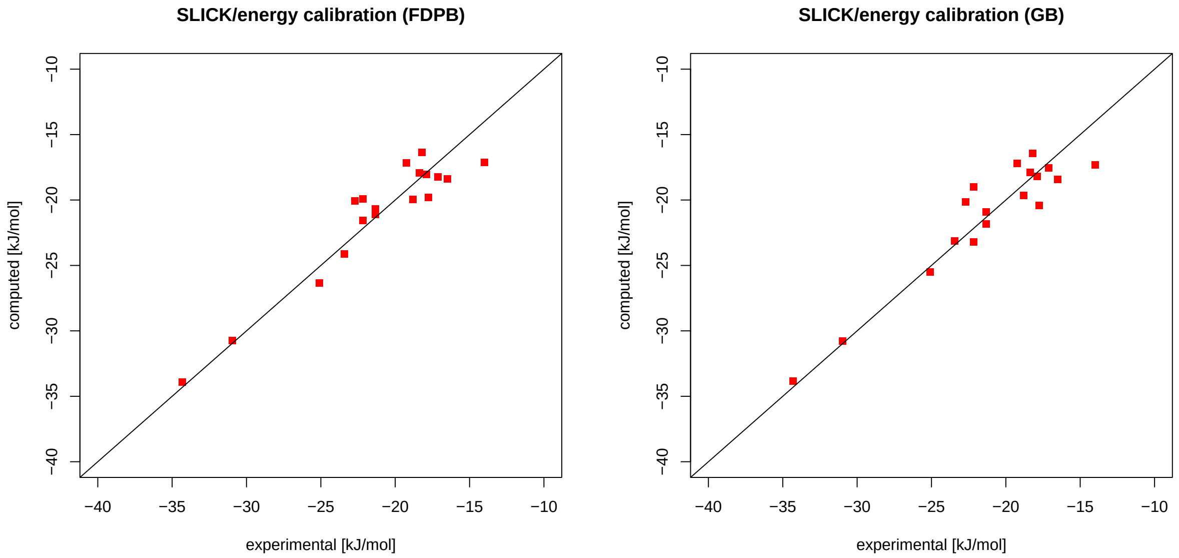


Figure 5.7.: Calibration graphs of SLICK/energy with FDPB (left, $R = 0.95$) and with GB electrostatics (right, $R = 0.94$).

Method	R	max(ΔE) [kJ/mol]	mean(ΔE) [kJ/mol]
SLICK/energy (FDPB)	0.95	3.13	1.27
SLICK/energy (GB)	0.94	3.33	1.27

Table 5.4.: Statistical data of the calibration with both electrostatics models. The correlation coefficient is denoted by R, max(ΔE) is the maximum absolute error and mean(ΔE) is the average absolute error of SLICK/energy in calibration.

in Fig. 5.8 shows the error distribution of the individual runs of the randomised cross-validation. Judging from these analyses, SLICK/energy should perform well even when calibrated on different sets of structures and energies.

After statistical assessment, SLICK/energy was applied for the prediction of the binding free energies of the two galectin complexes. The results are listed in Tab. 5.5. For the galectin complexes, the absolute error of the predicted energies is higher than the mean absolute error of the calibration but well within the limits of the maximum absolute error. Deducing a general tendency from two numbers is impossible. However, the energy differences suggest that calibrating SLICK/energy on plant lectins does not largely diminish its ability to predict lectins from other domains as well.

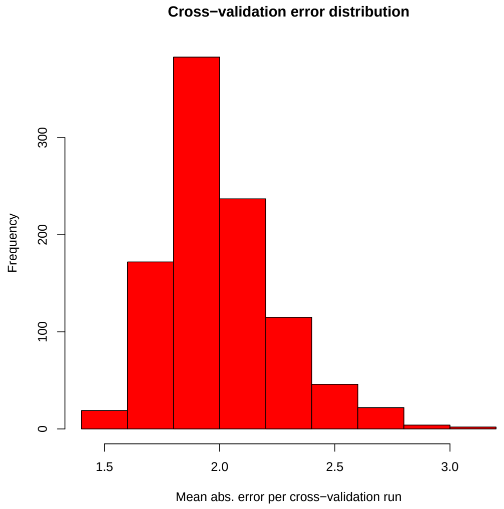


Figure 5.8.: The error distribution of the 1000 runs of randomised 5-fold cross-validation.

Complex	ΔG_{exp} [kJ/mol]	ΔG_{comp} [kJ/mol]	ΔE [kJ/mol]
4GAL	-19.25	-17.67	1.58
5GAL	-18.11	-15.64	2.47

Table 5.5.: Binding free energies of the galectin complexes predicted with SLICK/energy.

Running times

As expected, the running times of SLICK/energy are significantly higher as those of SLICK/score. The following numbers were computed on an AMD Opteron 250 with a CPU frequency of 2.4 GHz and 4 GB of main memory. On the calibration set, the computation times of SLICK/score range between 83 and 482 seconds. The average run time is 343.7 seconds. SLICK/energy needs between 1085 and 3151, with an average run time of 1587 seconds. Thus, SLICK/energy needs about 4.6 times more average computation time than SLICK/score. The computation time is dominated by the electrostatics component. Although SLICK/energy's electrostatics are calculated on an already reduced system, the simple Coulomb interaction calculated in SLICK/score is much faster than the full FDPB calculation.

5.4.2. Using SLICK/energy for Energetic Evaluation

Having both functions ready, the next step of validating their usefulness is to create a docking scenario. SLICK/score's ability to score docking candidates was assessed in Section 5.3. Now SLICK/energy is used for energetic evaluation of the docking candidates. First, binding free energies of the first true positives are computed and compared to the experimentally known values. Furthermore, the ability of SLICK/energy to identify the correct binding pose within the top ten ranked candidates is assessed.

Table 5.6 displays the results of energetic evaluation. The binding free energies of the first true positives deviate by about 7.6 kJ/mol from the experimental energies. The largest deviations

Complex	ΔG_{exp}	R_{ftp}	d_{ftp}	ΔG_{ftp}	ΔE	n-mer
	[kJ/mol]	[kJ/mol]	[Å]	[kJ/mol]		
1J4U	-18.24	1	1.11	-22.52	4.28	1
5CNA	-22.18	1	0.46	-23.20	1.02	1
1GIC	-19.25	1	0.59	-21.60	2.35	1
1QDO	-28.45	2	1.49	-28.25	0.20	2
1ONA	-30.96	1	1.14	-44.01	13.05	3
1DGL	-34.41	26	1.04	-46.70	12.29	3
1AXZ	-18.20	1	0.78	-22.37	4.17	1
1AX0	-17.90	1	0.89	-26.88	8.98	1
1AX1	-18.80	1	0.60	-19.19	0.39	2
1AX2	-22.70	25	1.01	-32.74	10.04	2
2BQP	-14.00	2	0.45	-21.43	7.43	1
1BQP	-16.60	1	0.78	-22.17	5.57	1
1QF3	-16.96	5	0.88	-21.71	4.75	1
2PEL	-17.76	34	1.06	-33.91	16.15	2
1EHH	-21.34	30	1.47	-34.90	13.56	3
1EN2	-23.43	1	1.25	-40.67	17.24	4
mean	–	8.31	0.94	–	7.59	–

Table 5.6.: SLICK/energy results: ΔG_{exp} denotes the experimental binding free energy, ΔG_{ftp} is the computed binding free energy of the first true positive, d_{ftp} is its RMSD and ΔE is the energy difference between computed and experimental binding free energy. Complexes without true positives were left out.

are found for large ligands, which is consistent with the results of SLICK/score. Oligomers are more difficult to compute for both functions, but in the case of SLICK/energy, the errors are much larger. For monomeric ligands, the mean deviation is about 1.7 kJ/mol lower than for the complete set. Ironically, the lowest deviations are found for two dimers, namely the Me-Man dimer binding to ConA and Lac binding to ECorL. Most surprisingly, LacNAc binding to the same lectin deviates more than 10 kJ/mol from the experimental value.

At this stage it is important to find out if the energy function can rule out false positives, *i. e.* structures that were scored well although they do not resemble the binding conformation. A summary of the results is given in Tab. 5.7.

In Tab. 5.7, the filtering results are classified as true positives if the lowest energy conformation is close to the binding conformation, *i. e.* if the RMSD difference is below 1.5 Å. If there is no conformation with low RMSD within the ten top scored structures, a distinction between true and false positive is not feasible. The results displayed in this table indicate that SLICK/energy does not perform very well finding the binding conformation among the top ten scored structures. In runs that yielded conformations with sufficient quality, SLICK/energy gave true positives in only seven of 12 cases.

Energies tend to be strongly underestimated. Additionally, non-binding conformations frequently get much lower energies than binding conformations. Comparing ΔE_{d_l} and $\Delta E_{d_{\text{min}}}$

Complex	ΔG_{exp}	ΔG_{d_l}	d_l	ΔE_{d_l}	$\Delta G_{d_{\text{min}}}$	d_{min}	$\Delta E_{d_{\text{min}}}$	TP/FP
	[kJ/mol]	[kJ/mol]	[Å]	[kJ/mol]	[kJ/mol]	[Å]	[kJ/mol]	
1J4U	-18.24	-22.41	1.06	4.17	-20.41	1.06	2.17	TP
5CNA	-22.18	-24.38	0.56	2.20	-21.97	0.46	0.21	TP
1GIC	-19.25	-31.10	3.85	11.85	-18.60	0.59	0.65	FP
1QDO	-28.45	-40.36	4.13	11.91	-25.21	1.49	3.24	FP
1QDC	-22.18	-37.35	4.15	15.17	-25.84	2.19	3.66	–
1ONA	-30.96	-49.41	9.09	18.45	-39.21	1.12	8.25	FP
1DGL	-34.41	-48.95	1.94	14.54	-31.99	1.81	2.42	–
1AXZ	-18.20	-23.34	0.77	5.14	-23.34	0.77	5.14	TP
1AX0	-17.90	-24.51	0.89	6.61	-22.85	0.84	4.95	TP
1AX1	-18.80	-24.69	0.71	5.89	-20.54	0.48	1.74	TP
1AX2	-22.70	-45.32	6.70	22.62	-26.38	6.08	3.68	–
2BQP	-14.00	-25.29	0.68	11.29	-20.36	0.42	6.36	TP
1BQP	-16.60	-24.53	11.07	7.93	-20.20	0.74	3.60	FP
1QF3	-16.96	-23.61	4.19	6.65	-20.09	0.85	3.13	FP
2PEL	-17.76	-39.16	4.79	21.40	-30.53	4.48	12.77	–
1EHH	-21.34	-39.17	2.51	17.83	-35.03	1.90	13.69	–
1EN2	-23.43	-41.36	1.05	17.93	-37.67	0.88	14.24	TP
1K7U	-21.34	-41.42	2.11	20.08	-33.41	2.02	12.07	–
mean	–	–	3.35	12.32	–	1.57	5.66	–

Table 5.7.: Filtering results: These numbers refer to the 10 best scored candidates of every complex. ΔG_{exp} denotes the experimental value for the binding free energy, ΔG_{d_l} is the lowest computed binding free energy, d_l is the RMSD of the candidate with the lowest computed binding free energy, ΔE_{d_l} is the absolute difference between experimental energy and lowest computed energy, $\Delta G_{d_{\text{min}}}$ denotes the calculated binding free energy of the candidate with the lowest RMSD, d_{min} is the lowest RMSD, $\Delta E_{d_{\text{min}}}$ is the difference between experimental energy and the energy of the structure with lowest RMSD. TP/FP displays whether the lowest energy conformation of a complex is a true positive (TP) or a false positive (FP).

demonstrates that the candidates with lowest RMSD were always much better in terms of energy prediction quality than those with lowest energy. It seems that SLICK/energy is very well able to predict accurate binding energies for structures that are very close to the actual binding conformation found in the crystal structure. But it fails when it comes to evaluating docking results.

The question is why SLICK/energy does so badly although SLICK/score performs quite well in identifying the binding site. The functions only differ in solvation and electrostatics. Consequently, there has to be a systematic error in those contributions. In order to find the culprit, the influence of every energy contribution was analysed (listed in Tab. 5.8). In this table, stronger influence means higher contribution of one energy term to the overall energy compared to the the other nine candidates of the top ten scored structures.

It seems that false positives are in most cases strongly influenced by the electrostatics term. A possible reason for the supposedly bad influence of the polar term is the parametrisation of

Complex	TP/FP	HB	CH···π	VDW	NP	ES
1J4U	TP	++	0			+
5CNA	TP				+	+
1GIC	FP		++	+		++
1QDO	FP	++				++
1ONA	TP	+			++	+
1AXZ	TP	+	+			
1AX0	TP	+	++			
1AX1	FP			+	+	+
2BQP	FP	++			+	
1BQP	FP	+			++	++
1QF3	TP	+				
1EN2	TP		+			

Table 5.8.: Dominating contributions to the lowest energy candidates. Abbreviations used: HB (hydrogen bonds), VDW (van der Waals term), ES (electrostatics term including polar solvation), NP (nonpolar solvation term). A plus sign "+" means noticeable influence, two plus signs "++" denote strong influence on the the energy. A zero means "no influence" at all. Only structures with positives within the top ten ranked structures are shown.

charges and especially radii. SLICK/score and SLICK/energy use PARSE charges and radii in the receptor. The charges of the ligand are computed from semi-empirical models while its radii are also taken from the PARSE parameter set. If the radii do not reproduce actual atomic radii for that case, the solvation term will produce bad results because radii have a strong influence on the calculation. This is true for both electrostatics models. The PARSE set was designed for reproducing accurate solvation free energies for small compounds similar to amino acids, which is the reason why it was chosen for these calculations. The results presented in the original PARSE publication [89] were of high accuracy as were our own computations on small molecules that were conducted before using this parameter set in the energy term. It is possible that this set of parameters does not perform very well on protein-carbohydrate complexes, especially when the radii are mixed with charges from semi-empirical calculations.

Finding a better source for electrostatics parameters is therefore imperative for the application of SLICK/energy to the evaluation of docking results if the scoring function allows bad conformations to be scored high. The same is probably true for the nonpolar solvation term. The solvation models themselves could also be the main error source, but judging from both published results and the assessment done prior to the incorporation of the solvation models into SLICK/energy, the parameter sets seem to be the more probable source of error. The binding free energies of candidates with low structural deviation from the native conformation are nevertheless reasonably well predicted.

5.5. Docking with SLICK

The analysis of SLICK/score showed that it is capable of identifying the binding conformations of lectins when rescoring docking candidates generated by AutoDock. Since AutoDock does not use SLICK/score in the structure generation process, the sets of conformations are not optimal for sugar docking, because important interactions are not embedded into the structure generator. Hence, a docking programme incorporating SLICK/score into the structure generation is the next aim in order to create candidates more suitable for the lectin-sugar docking problem.

When docking ligands to a receptor, the binding site of this receptor is usually known. Therefore, docking runs are limited to the region of the actual binding site in order to shorten run times. If the binding site is not known, one has to dock against the whole receptor surface which is very time consuming. In order to avoid global docking in the case of unknown binding sites, a method for finding putative binding sites on receptor surfaces can be employed.

This section will report on the integration of SLICK/score into an existing docking programme and discuss the results. Additionally, a binding pocket finder for lectins will be introduced. Together these two programmes should provide a full featured docking suite for protein-carbohydrate docking and even for lectins with unknown binding sites.

5.5.1. BALLDock/SLICK

SLICK was integrated into the docking programme BALLDock by Fuhrmann [7], which is a grid-based flexible ligand docking programme. It is based on a genetic algorithm for searching the conformational space of a ligand. For details on BALLDock beyond the introduction given in Section 3.3.2, please see [7]. It should be noted that, although BALLDock uses ideas from AutoDock, it does not include a local search for optimising ligand conformation before the creation of new generations.

The speed of a docking programme is dominated by the time necessary for calculating the energy of a ligand conformation. Grid-based methods use spatial discretisation for precomputing energy contributions of ligand atoms in order to speed up energy calculations during a docking run. The idea is to screen space with different probe groups that represent ligand atoms before the actual docking begins. First, space is discretised into a three-dimensional grid. Then for every grid point, a probe group is placed at the grid point location and the energy of that probe group is calculated. This energy is stored at that position in the grid. During the docking, the energy contribution of a ligand atom is then rapidly calculated by interpolating between the grid points occupied by this atom instead of computing the full energy term.

Because BALLDock employs a genetic algorithm, the ligand state has to be encoded into genes and chromosomes. Ligands are described by their translation, rotation and the torsion angles of every rotatable bond. A chromosome in BALLDock hence consists of $7 + n$ floating point numbers, three for the coordinates of the position of the ligand, four for a quaternion defining its orientation and one for each of the n rotatable bonds of the ligand. From these numbers, the coordinates of every ligand atom can be computed easily. The atomic coordinates represent the phenotype of a ligand and will be evaluated together with information about the properties of each atom by the fitness function.

BALLDock allows for docking being limited to a user-definable space. Ligand conformations are only permitted within this defined region. Obviously, a reasonable definition of this region is a sufficiently large area around the binding site providing enough freedom for generating many different conformations. BALLDock is able to define this region from the dimension of the ligand by calculating a bounding box of the ligand and extending it by a user-definable amount nto every direction. If the binding site is not known from the crystal structure, a binding site finding programme can be employed to define a suitable docking box.

The original energy function of BALLDock consists of three contributions. It comprises van der Waals energy, electrostatics and conformational energy of the ligand. Like SLICK/score, BALLDock uses a softened form of the AMBER implementation in BALL to calculate van der Waals energies. Electrostatics are calculated with the Coulomb formula. The conformational energy of the ligand is simply its AMBER energy.

BALLDock was adapted by incorporating the missing energy terms into the energy function. The first approach was to include the SLICK/score hydrogen bond and CH$\cdots\pi$ term into the grid building programme. Since hydrogen bonds and CH$\cdots\pi$ interactions are modelled in a purely geometric fashion, these models had to be adapted. In both cases, the position of the hydrogen is of great importance. During the grid building process, this position is not known. So the idea was to assume perfect hydrogen locations during the grid building process. This approach is based on the assumption that a ligand will orientate its hydrogens in such a way that favourable interactions can be established.

In the case of hydrogen bonds this meant to place a hydrogen donor atom on the grid point and locate possible acceptors in the receptor. Then for every possible acceptor, a hydrogen was added to the probe group at a position ideal for this particular hydrogen bond and the score was calculated. The maximum score was then stored in the grid. For the CH$\cdots\pi$ term, the approach was similar. Instead of searching hydrogen bond acceptors, aromatic rings in the receptor had to be located. A carbon atom is placed on the grid point and for every aromatic ring a hydrogen is placed at an ideal position. Again, the maximum score is stored in the grid.

In practise, this approach improved docking results for sugars but the impact was still small. The actual positions of hydrogen atoms of a putative ligand conformation proved too important to be approximated in the grid scores. Therefore, the real scoring functions were incorporated into the scoring computations instead of the grid based calculation. Thus, the actual conformation of a ligand is scored instead of interpolating between grid points. Following this approach, the docking results improved noticeably. The effect on running time is not very large because the scoring functions calculate simple geometry and are optimised for efficiency. The next section will present the results of this improved approach.

5.5.2. Sugar Docking Results

After integrating SLICK/score into BALLDock, the calibration data set was docked. Docking runs were performed with a large initial population of 5000 individuals while the population size during the docking was limited to 200 chromosomes for monomers and to 400 for larger sugars. Mutation rate was defined as 0.05. Docking was performed in a docking box which extended 8 Å from the bounding box of the ligand known from crystal structure. For every monomer complex, 600 runs were performed and analysed. For larger sugars, the number of runs was raised to 1000. The larger parameters for oligo-carbohydrates was necessary because of the high flexibility of the ligands along the glycosidic bonds. With smaller parameters, the conformational space of these ligands was not sufficiently scanned.

With these parameters, a sufficient coverage of the binding site was possible. Fig. 5.9 shows two

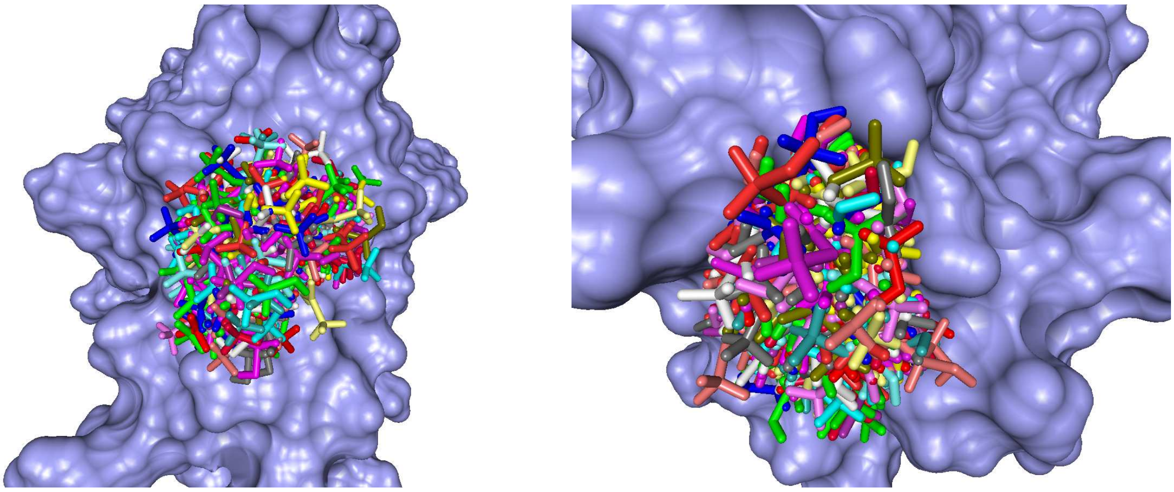


Figure 5.9.: BALLDock candidates covering the binding sites of ECorL (left, 1AX1) and PNA (right, 1QF3). Please see Fig. 5.1 for comparison with AutoDock candidates.

examples of binding site coverage which compare directly to the binding sites shown in Fig. 5.1 on page 67. Docking results are summarised in Tab. 5.9. Plots of RMSD vs. scores of all docking runs are given in appendix B.2.2 The RMSD distribution of the generated docking candidates are listed in appendix B.2.1.

Looking at Tab 5.9, two observations become apparent. First, BALLDock/SLICK is very well able to create and identify good approximations of binding conformations. In almost every case, the first true positive is the top ranked structure and the mean deviation is at only 0.85 Å. Second, the mean absolute error of the binding free energy for the first true positive is below 4 kJ/mol. This fact underlines the effectiveness of SLICK in docking protein-carbohydrate complexes.

Again, monomers seem to be easier to dock. This is not surprising because monomers are small and much less flexible than oligomers. It is also consistent with the results from rescoring the AutoDock candidates. Fig. 5.10 shows exemplary docking plots of BALLDock/SLICK results. All candidates with the exception of 2PEL show the correct tendency for scoring less deviating conformations better. But in almost every docking run, a cluster of numerous largely deviating structures is observable.

Analysis of the binding candidates produced by BALLDock/SLICK revealed the reasons for the deviating clusters. In the monomer case, one encounters many rotated rings. Compared to the binding conformation, these rings are rotated around the symmetry axis of the ring plane of the sugar. The scoring function is able to distinguish the correct pose from the incorrect one. The ConA/Me-Glc complex (1GIC) illustrates that very well. The scoring function identifies two highly scored clusters, one at about 1 Å RMSD and one at about 4 Å deviation. The latter cluster contains many rotated rings, which SLICK/score correctly identifies as a cluster of false positions. An exemplary illustration of rotated monomers is given in Fig 5.11.

In the case of dimers, clusters of large deviation are again caused by sugar residues that extend into the solvent. Fig. 5.12 shows an illustrating example for this behaviour. The second Man ring, which is surrounded by solvent, is not sterically only limited by the highly flexible glycosidic

PDB ID	ΔG_{exp}	R_{ftp}	d_{ftp}	ΔG_{ftp}	ΔE_{ftp}	R_{min}	d_{min}	ΔG_{min}	ΔE_{min}
	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]
1J4U	-18.24	1	0.54	-17.30	0.94	1	0.54	-17.30	0.94
5CNA	-22.18	1	0.91	-20.33	1.85	7	0.58	-19.00	3.18
1GIC	-19.25	1	0.48	-17.85	1.40	1	0.48	-17.85	1.40
1QDO	-28.45	8	0.83	-22.05	6.40	58	0.80	-19.11	9.34
1QDC	-22.18	2	0.80	-19.60	2.58	2	0.80	-19.60	2.58
1ONA	-30.96	1	1.32	-28.56	2.40	129	1.06	-21.84	9.12
1DGL	-34.41	2	1.43	-26.14	8.27	4	0.91	-28.81	5.60
1AXZ	-18.20	1	0.47	-15.79	2.41	15	0.26	-16.34	1.86
1AX0	-17.90	1	0.51	-21.06	3.16	2	0.29	-16.37	1.53
1AX1	-18.80	1	0.86	-19.39	0.59	10	0.60	-18.37	0.43
1AX2	-22.70	1	0.76	-19.12	3.58	2	0.53	-20.27	2.43
2BQP	-14.00	1	0.72	-16.30	2.30	3	0.42	-15.91	1.91
1BQP	-16.60	1	0.68	-15.65	0.95	2	0.64	-16.44	0.16
1QF3	-16.96	1	0.52	-18.26	1.30	19	0.24	-16.73	0.23
2PEL	-17.76	–	–	–	–	227	1.78	-19.79	2.03
1EHH	-21.34	1	1.47	-32.09	10.75	3	0.93	-25.49	4.15
1EN2	-23.43	1	1.17	-29.54	6.11	5	0.82	-27.06	3.63
1K7U	-21.34	1	1.04	-27.47	6.13	3	0.58	-28.35	7.01
mean	–	1.53	0.85	–	3.60	27.39	0.68	–	3.20

Table 5.9.: Results of docking the calibration set with BALLDock/SLICK. ΔG^{exp} denotes the experimental binding free energy, R_{ftp} is the rank of the first true positive, d_{ftp} its RMSD, ΔG_{ftp} its computed binding free energy. ΔE_{ftp} is the absolute difference between *DGexp* and ΔG_{ftp}. The respective numbers are also given for the candidate with minimal RMSD (variables with index min).

bond. BALLDock/SLICK creates many conformations with the tightly binding first ring in perfect position while the pose of the second ring is rotated away from the binding mode. The high flexibility of the glycosidic bond connecting these two rings makes thorough scanning of the ligand's conformational space imperative.

There remains only one problematic case, which is PNA binding Lac (2PEL). This is surprising because other complexes with Lac or LacNAc seem to work reasonably well. Additionally, the second PNA complex in the calibration set (1QF3) gives perfect results. The ligand in 1QF3 is Me-Gal. In most Lac or LacNAc complexes, Gal is the ring that binds directly to the receptor while the Glc ring extends into the solvent. Assuming this behaviour in the PNA/Lac complex as well, it is not obvious why docking Lac with its Gal ring into the binding site of PNA should fail.

A nice result is the good performance of the largest ligands in the calibration set. Docking a GlcNAc trimer and tetramer to UDA (complexes 1EHH and 1EN2) results in top ranked structures below 1.5 Å. The lowest deviations achieved are at 0.93 and 0.82 Å RMSD, respectively. These complexes are strongly influenced by CH$\cdots\pi$ interactions, which is reproduced by the model very well (cf. Fig. 5.13).

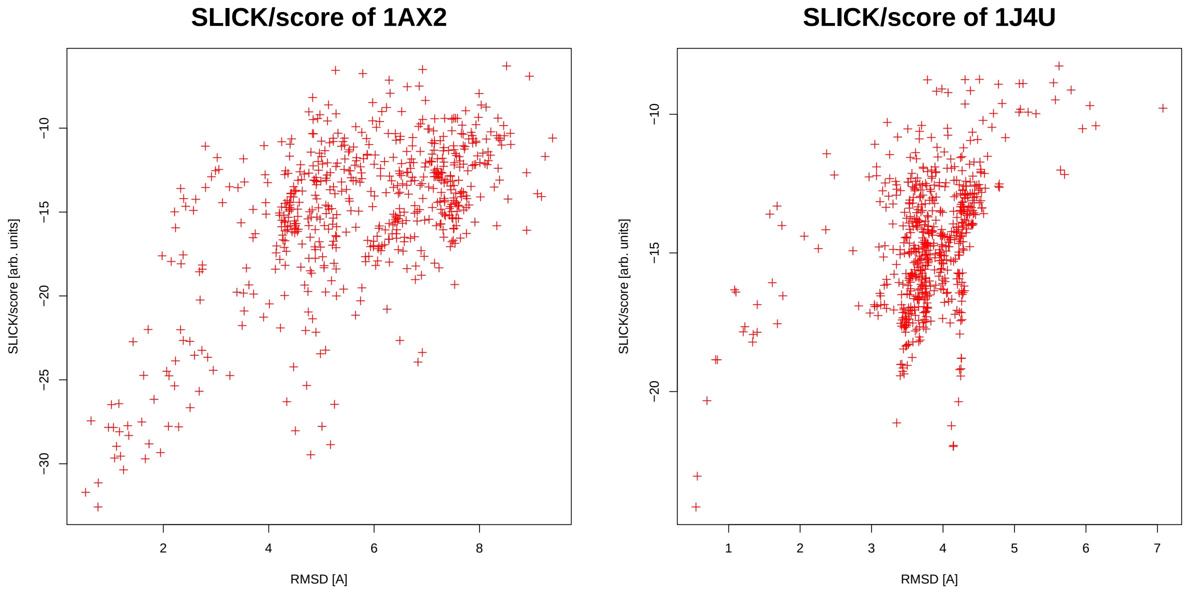


Figure 5.10.: Exemplary results of BALLDock/SLICK. Left: ECorL/LacNAc (1AX2). Right: AIA/Me-Man (1J4U). Note that 1J4U does not have CH···π interactions in the binding site.

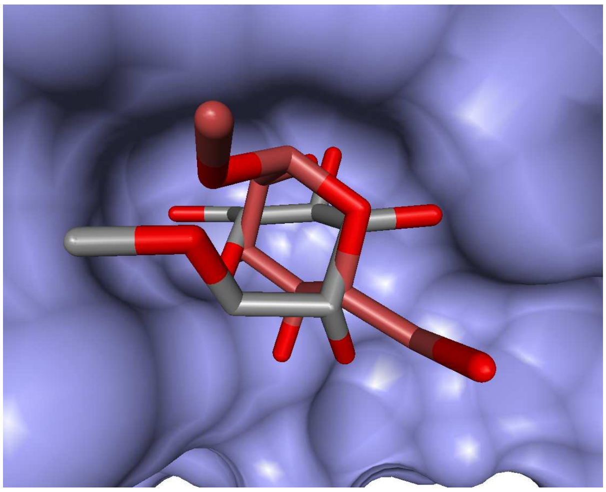


Figure 5.11.: Rotated monomer ring in the binding site of ConA. The best ranked structure (normal colours) has an RMSD of below 1 Å while the RMSD of the rotated monomer (reddish colours) is about 4 Å. Note the position of the ring oxygen.

The galectin test set was also docked and re-evaluated (see Tab. 5.10 and Fig. 5.14). As in the calibration set, the binding conformations of the two galectin complexes are found very accurately. The prediction of the binding free energy of the complexes, however, is worse than in the plant lectin case.

Docking an Extended Set of Sugar-Binding Proteins

For further assessment of the ability of BALLDock/SLICK to identify binding conformations of protein-carbohydrate complexes, an additional set of 20 lectins and sugar-binding proteins

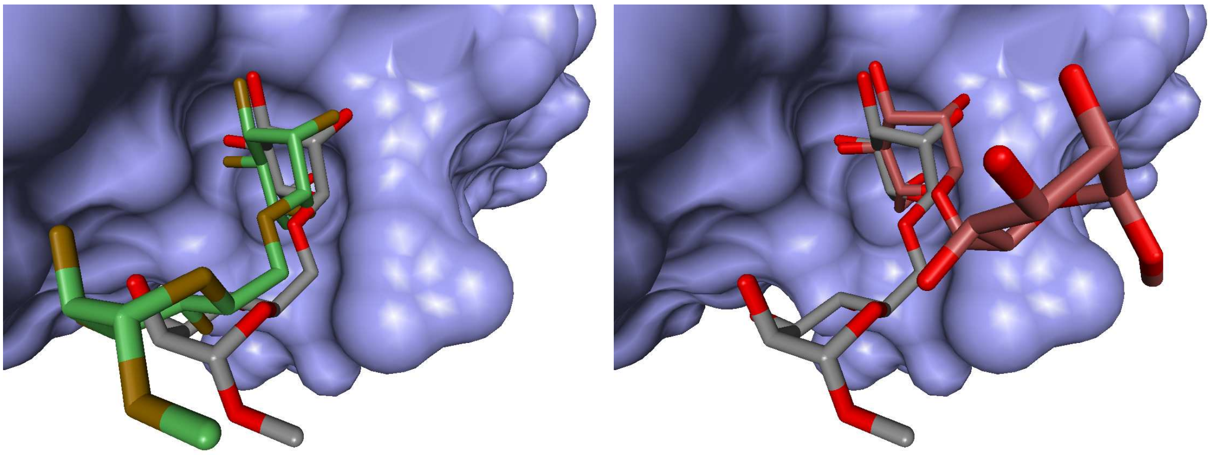


Figure 5.12.: Mannose-dimer binding to ConA. Left: comparison between crystal structure (normal colours) and best scored docking candidate (greenish colours, RMSD 1.70 Å, pivotal ring RMSD 0.96 Å). Right: crystal structure (normal colours) and one of many conformations with the second ring rotated out of the binding site (reddish colours, RMSD 4.50 Å, pivotal ring RMSD 0.67 Å).

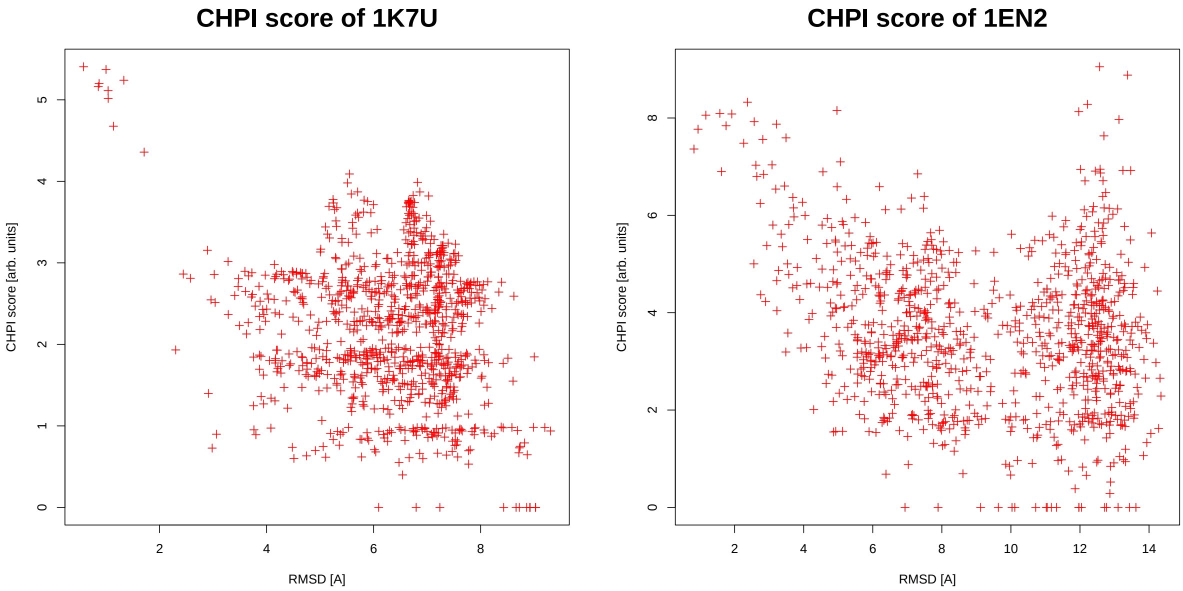


Figure 5.13.: CH···π scores of WGA and UDA complexes.

PDB ID	ΔG_{exp}	R_{ftp}	d_{ftp}	ΔG_{ftp}	ΔE_{ftp}
	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]
4GAL	-19.25	1	1.04	-29.58	10.33
5GAL	-18.41	2	1.33	-31.84	13.43
mean	–	1.50	1.19	–	11.88

Table 5.10.: Results of docking the galectin set with BALLDock/SLICK. Please find the description of the columns in Fig. 5.9.

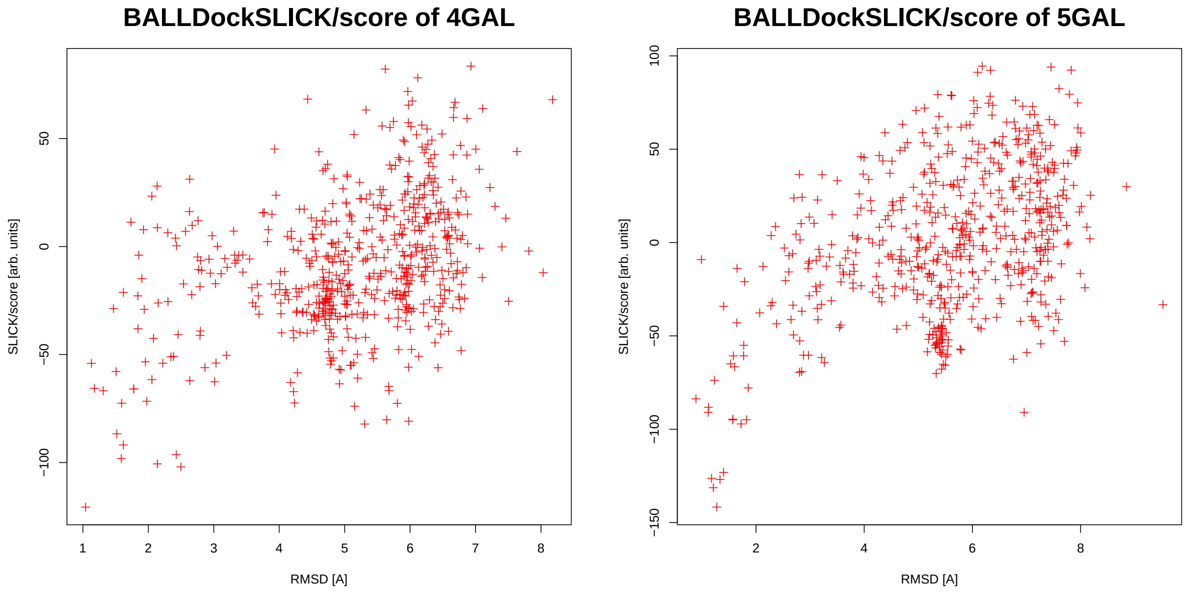


Figure 5.14.: Results of docking the galectin set with BALLDock/SLICK

was docked. The docking set is two-fold. One part contains only plant lectins, while the other consists of animal lectins and general sugar-binding proteins. For these complexes no binding free energies were found in literature. Consequently, the performance of SLICK/energy in evaluating the docked compounds energetically is not assessable. The results of these docking runs are summarised in Tab. 5.11. The docking plots are given in appendix B.2.3.

The docking set does not only cover non-plant lectins, it also contains sugar ligand that were not included in the calibration set. The latter only contains sugars built from Man, Glc and Gal monomers and their methylated and acetylated derivatives. In the docking set we find fructose, fucose and maltose. While Man, Glc and Gal are pyranoses in D-conformation, fructose is a furanose and fucose is a deoxy-L-galactose. Maltose is a Glc dimer. Since these ligands differ much from the calibration set, docking these sugars will permit to assess whether BALLDock/SLICK is able to cope with a broad range of sugar ligands or whether SLICK/score is biased too much by the choice of the calibration set. In addition, some of the proteins in the docking set have very deep binding pockets in contrast to the rather shallow binding sites of most plant lectins in the calibration set.

Complex	R_{ftp}	d_{ftp}	$R_{\min}$	$d_{\min}$	n-mer
		[Å]		[Å]	
	Plant Lectins				
`1KJ1`	1	0.92	11	0.46	1
`1KUJ`	1	0.54	1	0.54	1
`1MVQ`	1	0.51	4	0.29	1
`1WBL`	1	0.31	3	0.31	1
`1FNZ`	1	1.00	2	1.00	1
`1C3M`	1	0.27	1	0.27	2
`1GZC`	1	0.93	2	0.45	2
`1JOT`	1	1.43	12	0.93	2
`1PUM`	10	0.74	114	0.51	1
`1PUU`	17	0.43	29	0.40	2
`1HKD`	1	1.25	11	0.70	1
`1RIN`	2	0.86	3	0.44	1
`1OFS`	7	1.25	7	1.25	2
mean	3.46	0.80	15.38	0.58	–
	Non-Plant Lectins				
`1DIW`	17	1.13	202	1.01	1
`1GLG`	1	0.42	75	0.23	1
`1K12`	3	0.74	7	0.21	1
`1NL5`	1	1.03	1	1.03	2
`2GAL`	(117)	(0.54)	152	0.26	1
`1C1L`	1	0.68	1	0.68	2
`1SLT`	1	1.10	8	0.75	2
mean	4.00	0.85	63.71	0.59	–
total mean	3.63	0.82	32.3	0.59	–

Table 5.11.: Results of BALLDock/SLICK on the docking set. Please find the description of the columns in Tab.5.1. Complexes with an R_{ftp} above 100 are considered unsuccessful and discarded from the FTP mean. These numbers are given in brackets.

In this section, the structural analysis of the docking runs is of great importance, especially when it comes to non-plant lectins, because only on this basis insights for the further development of BALLDock/SLICK and SLICK in general can be gained. Therefore, problem cases that occur in these docking runs are analysed in depth in order to gain knowledge on the interactions and the errors made by BALLDock/SLICK.

For the 13 plant lectins, the docking runs confirm the results achieved with the calibration set. In nine cases, the first true positive is at the same time the top ranked structure. For another two complexes the first true positive is found among the top ten candidates. The mean RMSD of first true positives is only 0.8 Å. The two *viscum album* lectin complexes (1PUM, 1PUU) seem to be harder to dock. Their lowest RMSD structures rank at 114 and 29, respectively, although in both cases candidates with very small deviation exist.

In these cases, the docking plots reveal two highly scored regions. In one case, good approximations of the binding mode are scored well while in the other there is a cluster at about 3.5 Å deviation. The latter cluster is dominated by electrostatic energy. Structural analysis showed that in the close vicinity of the ligands five aspartic acid side chains (ASP 23, 26, 27, 28 and 45) and one asparagine (ASN 47) can be found. This could explain the strong electrostatic interactions dominating the scoring. Although there is a tryptophane in the binding site, which in the crystal structure clearly is participating in many CH$\cdots\pi$ bridges, the score for these interactions is too low to compensate for the strong electrostatics term.

Some docking runs produced top ranked structures with large deviations from the native binding mode. In the case of 1RIN this is an outlier in an otherwise perfect plot. In 1OFS, the bad conformations with high scores are dominated by van der Waals energies. In this case, the ligand is sucrose which consists of fructose and glucose. In contrast to all other monomers encountered so far, fructose is a furanose, which means that the ring consists of four carbons and an oxygen instead of five carbons and an oxygen. Thus, fructose is smaller than glucose and fits better into the pocket sterically. Consequently, conformations which place fructose into the binding site are scored better. The crystal structure, however, reveals that the bound conformation is exactly opposite. The glucose ring binds to the receptor while the fructose ring is extended.

In summary, the results for the plant lectin part of the docking set is very satisfactory. In comparison, the non-plant lectins results are only slightly inferior. In four cases the bound conformation of the ligand is unambiguously found. The mean values of d_{ftp} and d_{min} are 0.85 and 0.59 Å, respectively, which compares well to the plant lectin docking set.

In the case of 1K12 the tendency of the scoring function is correct, but the monomer is found in many twisted conformations. In addition, there are several aromatic groups in the vicinity of the binding site. Astonishingly, these groups seem to have little influence on the actual binding conformation, judging from the pose found in the crystal structure (Fig. 5.15). Twisted rings that build CH$\cdots\pi$ bridges are thus scored higher than those resembling the bound state, which is compensated only partially by the other interactions.

The same seems true for the 1DIW complex. Here one tyrosine is the only aromatic side chain in the binding site. But this side chain does rather contribute electrostatically than as CH$\cdots\pi$ partner in the crystal structure. Located below the sugar ring is an aspartic acid that can act as hydrogen bond partner. But in the crystal structure, the Gal seems to build hydrogen bonds to the backbone. Consequently, twisted conformations that build hydrogen bonds to the ASP are ranked very high, which makes it difficult for SLICK/score to identify the real binding pose. Additionally, the binding site is partially built from a very flexible coil, shown in Fig. 5.15. The question remains whether the crystal conformation is the only binding pose of Gal in this case.

The hGal-7/Gal complex (2GAL) is another example where twisted rings are favoured. In the crystal structure, the only axial hydroxyl group of Gal is coordinated within the binding site by two hydrogen bonds. The docking programme tends to twist the ring such that this axial group stays out of the binding site. In this case, van der Waals, electrostatics and hydrogen bond term agree that the twisted conformation is better than the crystal structure. Unfortunately, the binding modes of Lac and LacNAc in other complexes reveal exactly the same behaviour relying on the coordination of the axial hydroxyl group of Gal. Consequently, in this case SLICK/score is completely unable to score the real binding pose accurately.

Considering the two complexes with deep binding pockets, SLICK/score is able to identify the

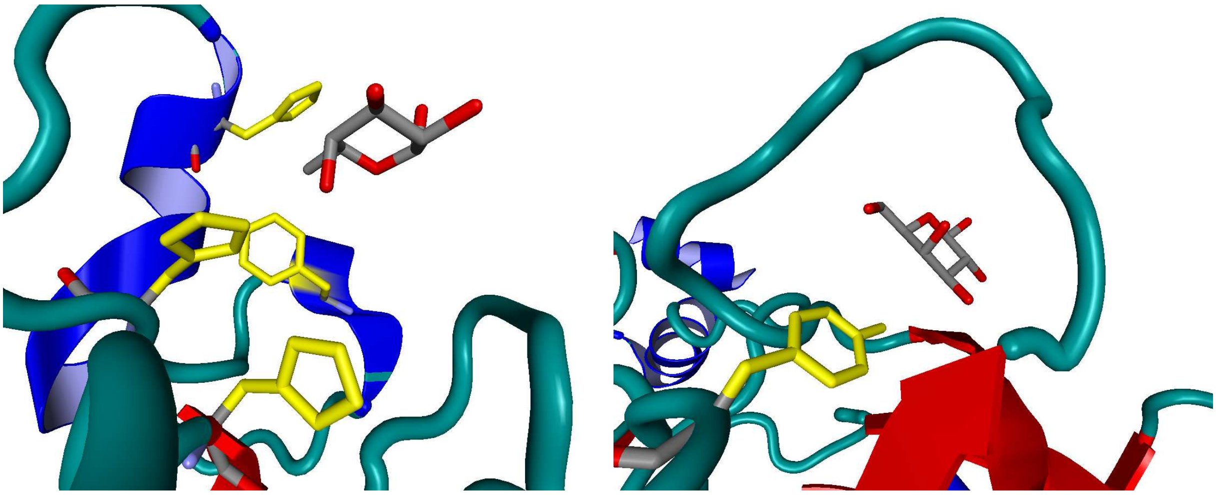


Figure 5.15.: Left: binding site of 1K12 with many aromatic side chains. No CH$\cdots\pi$ interactions can be observed. Right: binding site of 1DIW. TYR 1180 is the only aromatic side chain in the vicinity, but it seems to contribute electrostatically only. A CH$\cdots\pi$ interaction cannot be built. Note the flexible coil above the sugar.

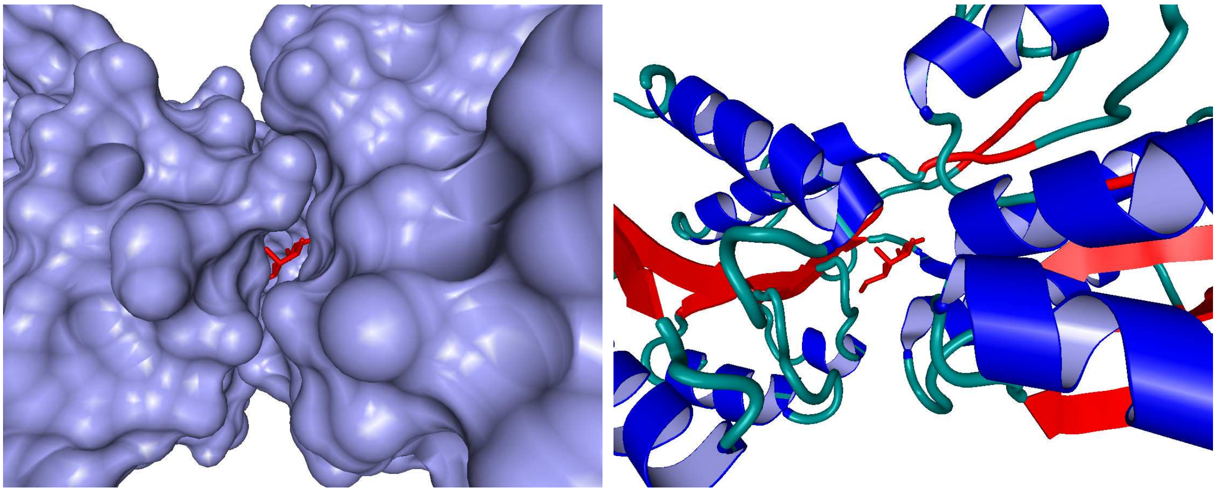


Figure 5.16.: The deeply buried ligand in the binding site of 1GLG.

binding positions correctly. The very limited space that the bound conformations of the proteins provide for the ligands clearly impacts the conformational space of both ligands, making the prediction rather easy. While in 1GLG the binding pocket is very narrow (see Fig. 5.16), in 1NL5 the ligand has a bit more freedom, which is easily observable when comparing the RMSD plots of both complexes.

In summary, the results prove that BALLDock/SLICK is well able to dock protein-carbohydrate complexes from different domains and with a large variety of sugars as ligands. Although there are some problem cases, the RMSD plots of the docked structures clearly indicate that SLICK/score does in most cases identify the binding conformation. Additionally, the deviation of the first true

PDB ID	ΔG_{exp} [kJ/mol]	R_{ftp}	d_{ftp} [Å]	ΔG_{ftp} [kJ/mol]	ΔE_{ftp} [kJ/mol]	R_{min}	d_{min} [Å]	ΔG_{min} [kJ/mol]	ΔE_{min} [kJ/mol]
1J4U	-18.24	3	1.41	-37.28	19.04	51	0.96	-36.40	18.16
5CNA	-22.18	1	0.89	-39.91	17.73	90	0.46	-38.99	16.81
1GIC	-19.25	1	0.90	-39.25	20.00	120	0.59	-37.53	18.28
1QDO	-28.45	(175)	(1.49)	-42.93	14.48	175	1.49	-42.93	14.48
1QDC	-22.18	–	–	–	–	169	2.19	-44.52	22.34
1ONA	-30.96	3	1.14	-64.06	33.10	16	1.12	-61.59	30.63
1DGL	-34.41	9	1.04	-63.85	29.44	46	0.62	-59.58	25.17
1AXZ	-18.20	1	1.20	-34.90	16.70	139	0.66	-30.92	12.72
1AX0	-17.90	1	0.88	-41.59	23.69	111	0.70	-38.83	20.93
1AX1	-18.80	78	0.49	-39.91	21.11	99	0.46	-38.45	19.65
1AX2	-22.70	93	1.01	-46.07	23.37	106	0.89	-45.56	22.86
2BQP	-14.00	73	0.47	-35.27	21.27	86	0.42	-35.10	21.10
1BQP	-16.60	40	0.81	-35.35	18.75	79	0.71	-34.81	18.21
1QF3	-16.96	54	0.98	-32.51	15.55	107	0.73	-31.76	14.80
2PEL	-17.76	(119)	(1.18)	-42.26	24.50	166	1.01	-38.87	21.11
1EHH	-21.34	91	1.47	-49.04	27.70	91	1.47	-49.04	27.70
1EN2	-23.43	15	1.43	-60.33	36.90	34	0.88	-58.20	34.77
1K7U	-21.34	–	–	–	–	44	2.02	-53.60	32.26
mean	–	28.94	0.88	–	22.71	94.75	0.82	–	21.09

Table 5.12.: AutoDock energies for the calibration set. Please see Tab. 5.9 for an explanation of the columns. Complexes with an R_{ftp} above 100 are considered unsuccessful and discarded from means of R_{ftp} and d_{ftp}. These numbers are given in brackets.

positives created with BALLDock/SLICK is very low with many candidates below 1.5 Å RMSD.

5.5.3. Comparison with existing Docking Programmes

The ability of docking carbohydrates into protein binding sites clearly distinguishes BALLDock-/SLICK from many other docking programmes. In this section, a short comparison of BALLDock/SLICK with two existing docking methods will be given.

AutoDock

The results of docking the calibration set with AutoDock were already shown in Section 5.3.3. Here, the results will be summarised briefly. A comparison of BALLDock/SLICK and AutoDock shows that the latter is clearly unable to identify binding modes of such complexes while the former finds good approximations of the bound conformation with high accuracy. The structure generator of AutoDock covers the conformational space of the ligand quite well and conformations of low deviation are generated during a docking run. Although some binding modes were identified correctly by the AutoDock energy function, most binding poses were ranked very poorly.

When comparing the predicted binding energies, the picture gets worse. AutoDock energies of the first true positives deviate by as much as 22.7 kJ/mol in the mean. The energy function of

Complex			**SLICK**				**AutoDock**			
PDB	n-	ΔG_{exp}	R_{ftp}	d_{ftp}	ΔG_{comp}	ΔE	R_{ftp}	d_{ftp}	ΔG_{comp}	ΔE
ID	mer	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]
Reduced Calibration Set										
1J4U	1	-18.24	1	1.11	-20.38	2.14	3	1.41	-37.28	19.04
5CNA	1	-22.18	1	0.46	-22.23	0.05	1	0.89	-39.91	17.73
1AXZ	1	-18.20	1	0.78	-21.57	3.37	1	1.20	-34.90	16.70
1BQP	1	-16.60	1	0.78	-21.10	4.50	40	0.81	-35.35	18.75
1AX2	2	-22.70	25	1.01	-30.44	7.74	93	1.01	-46.07	23.37
2PEL	2	-17.76	34	1.06	-33.90	16.14	(119)	(1.18)	-42.26	24.50
1ONA	3	-30.96	1	1.14	-36.87	5.91	3	1.14	-64.06	33.10
1EHH	3	-21.34	30	1.47	-36.48	15.14	91	1.47	-49.04	27.70
mean		–	11.75	0.98	–	6.87	33.14	1.13	–	22.61
Energy Validation Set										
1GIC	1	-19.25	1	0.59	-19.91	0.66	1	0.90	-39.25	20.00
1AX0	1	-17.90	1	0.89	-24.98	7.08	1	0.88	-41.59	23.69
2BQP	1	-14.00	2	0.45	-21.35	7.35	73	0.47	-35.27	21.27
1QF3	1	-16.96	5	0.88	-20.50	3.54	54	0.98	-32.51	15.55
1QDO	2	-28.45	2	1.49	-25.98	2.47	(175)	(1.49)	-42.93	14.48
1AX1	2	-18.80	1	0.60	-21.19	2.39	78	0.49	-39.91	21.11
4GAL	2	-19.25	2	1.27	-32.21	12.96	67	1.29	-37.61	18.36
5GAL	2	-18.41	3	1.13	-33.13	14.72	18	1.13	-41.17	22.76
1DGL	3	-34.41	26	1.04	-41.85	7.44	9	1.04	-63.85	29.44
1EN2	4	-23.43	1	1.25	-37.53	14.10	15	1.43	-60.33	36.90
mean		–	4.40	0.96	–	7.27	35.11	0.96	–	22.36

Table 5.13.: Results of rescoring and evaluating AutoDock candidates of calibration and validation set with SLICK. R_{ftp} denotes the rank of the first true positive candidate (RMSD $<$ 1.5Å), d_{ftp} is the RMS deviation of the first true positive from crystal structure, ΔG_{comp} denotes the computed binding free energy and ΔE is the deviation of the predicted energy from the experimental binding free energy ΔG_{exp}. Numbers are given for SLICK and AutoDock. The complexes 1QDC and 1K7U of the calibration set are not shown. For orientation, the number of monomers in the ligand is given in column n-mer. Complexes with an $R_{\text{ftp}} > 100$ were considered as unsuccessful docking runs and were discarded from the average. These numbers are given in brackets.

AutoDock underestimates energies systematically. Although SLICK/energy does have weaknesses when using it for filtering purposes, its performance on reasonable structures is satisfactory. AutoDock is, of course, a general flexible docking programme while BALLDock/SLICK was specially designed for protein-carbohydrate complexes.

Doing comparisons on SLICK's calibration set alone is surely not fair. Consequently, the calibration set was split into two sets, one for recalibrating SLICK and one for prediction. The reduced calibration set was chosen to include plant lectins from every family. The validation set contains the remaining plant lectins plus the two complexes of human galectin-7. Thus, the comparison of AutoDock with SLICK rescoring to original AutoDock scoring should become more robust. Results of this comparison are given in Tab. 5.13. Evidently, even on the drastically reduced set, SLICK still improves docking results of this validation set compared to standard AutoDock results.

FlexX

One of the most popular docking programmes is FlexX [47], which uses an incremental construction algorithm for creating putative binding conformation. FlexX is known for its speed and often used in screening large substance libraries for possible binding candidates. But it is also known for very accurate predictions of binding conformations of drug-like substances, which is the reason for choosing this programme for comparison purposes.

Like AutoDock, FlexX is a programme for flexible ligand docking with the receptor kept rigid throughout the computations. FlexX starts by separating the ligand into small fragments along rotatable bonds. One of these fragments is chosen for the role of *base fragment* and placed into the receptor binding site. A base fragment is chosen based on its size and interaction potential. Usually, relatively large fragments with a large number of possible interaction partners are chosen. The base fragment is then placed according to FlexX's energy function. After the base fragment has been put into the binding site, adjacent fragments are connected to the base fragment using rotamer libraries for defining the dihedral angles of a newly connected fragment. This procedure is repeated until all fragments have been consumed and the ligand has been re-built incrementally. FlexX uses several heuristics to speed up the construction process and to avoid combinatoric explosion, which will not be detailed here.

The first step in comparing FlexX with BALLDock/SLICK on protein-carbohydrate complexes was docking the calibration set of SLICK with FlexX release 2.02. In these docking experiments, FlexX standard parameters were used. Structures were taken from the PDB entries. The protonation state was determined automatically by FlexX. Atomic parameters for the energy function were automatically assigned. Ring conformations of the sugar rings of the ligand were taken from the crystal structure instead of using CORINA conformations, which is reasonable given the very rigid nature of sugar-rings. Receptor binding sites were defined by using spheres of 6.5 Å around ligand atoms, which is the default method. The base fragment was chosen automatically, as well as the placement of the base fragment in the binding site.

Table 5.14 shows the result of these docking attempts. As in all previous analyses, only solutions with a heavy-atom RMSD of below 1.5 Å were considered as true positives. FlexX was only able to create nine solutions of this quality. While most ligand conformations are at least in the vicinity of this limit, two oligomers (GlcNAc trimer and tetramer binding to UDA) are very far away from the native conformation with 5.7 Å and 9.4 Å deviation. While the ranking of true positives is acceptable, the binding free energy estimates calculated by FlexX for the first true positives are deviating by 11.4 kJ/mol in the mean, which is better than standard AutoDock but still does not reach the prediction quality of SLICK.

The results obtained from FlexX indicate that it performs well in ranking the lectin-sugar complexes as long as reasonable candidates are produced. In five of nine cases, the first true positive is also the best ranked conformation. But in two cases (1AX0 and 2PEL), the rank of the near-native conformation is very bad. Judging from these results, it seems that the FlexX energy function seems to cope with protein-carbohydrate interactions only partially. The question remains whether the overall performance is dominated by the structure generator or the energy

PDB	n-	ΔG_{exp}	R_{ftp}	d_{ftp}	ΔG_{ftp}	ΔE_{ftp}	R_{min}	d_{min}	ΔG_{min}	ΔE_{min}
ID	mer	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]		[Å]	[kJ/mol]	[kJ/mol]
1J4U	1	-18.24	1	0.76	-12.82	5.42	2	0.67	-12.70	5.54
5CNA	1	-22.18	1	0.89	-9.23	12.95	1	0.89	-9.23	12.95
1GIC	1	-19.25	1	0.84	-15.11	4.14	15	0.33	-9.26	9.99
1QDO	2	-28.45	–	–	–	–	194	2.03	4.93	33.38
1QDC	2	-22.18	1	1.21	-13.62	8.56	2	0.80	-13.41	8.77
1ONA	3	-30.96	3	1.47	-9.57	21.39	7	1.04	-7.37	23.59
1DGL	3	-34.41	1	0.75	-18.11	16.30	1	0.75	-18.11	16.30
1AXZ	1	-18.20	–	–	–	–	16	2.24	-3.48	14.72
1AX0	1	-17.90	32	1.22	-8.02	9.88	46	1.14	-6.57	11.33
1AX1	2	-18.80	–	–	–	–	115	3.43	-3.74	15.06
1AX2	2	-22.70	–	–	–	–	190	2.00	0.48	23.18
2BQP	1	-14.00	–	–	–	–	80	2.15	-4.16	9.84
1BQP	1	-16.60	–	–	–	–	139	1.91	4.16	20.76
1QF3	1	-16.96	–	–	–	–	108	1.51	-1.18	15.78
2PEL	2	-17.76	70	1.39	-5.55	12.21	109	0.69	-4.10	13.66
1EHH	3	-21.34	–	–	–	–	7	5.67	7.82	29.16
1EN2	4	-23.43	–	–	–	–	9	9.41	18.77	42.20
1K7U	2	-21.34	6	1.19	-9.28	12.06	6	1.19	-9.28	12.06
mean	–	–	12.89	1.08	–	11.43	58.17	2.10	–	17.68

Table 5.14.: FlexX energies for the calibration set. Please see Tab. 5.9 for an explanation of the columns.

function.

In order to assess the influence of FlexX's energy function, the structures generated by FlexX were re-evaluated with SLICK, shown in Tab. 5.15. From the structures generated during the FlexX docking, the 200 best structures were taken. The procedure for re-evaluation was the same as for rescoring AutoDock candidates. Again, from the nine candidates below 1.5 Å, five first true positives are at the same time the highest ranked structures. Interestingly, the complexes with that high prediction quality differ between the FlexX scoring and the SLICK scoring. Considering 5CNA, the FlexX score ranked a true positive at the top of the list, while SLICK/score only achieves rank 26. On the other hand, 1AX0 and 2PEL are still under the top ten according to SLICK/score while FlexX ranks a true positive at 32 and 70, respectively. The average first true positive rank of the SLICK/score is at 4.7, which is considerably better than the mean rank of 12.9 achieved by FlexX. The biggest difference is found in the energy estimates. The energy difference between calculated and experimental values is at 3.1 kJ/mol when using SLICK/energy for energy calculations, which is well within the estimate obtained from the calibration of SLICK/energy.

Finally, the performance of FlexX was compared to BALLDock/SLICK on the docking set that was already employed earlier. Tab. 5.16 shows a summary of the results obtained by this comparison. Apparently, FlexX achieves significantly better results on the docking set than on the calibration set. Of the 20 structures of the docking set, 16 could be docked. The mean rank of the first true positive is 4.0 and the mean RMSD of the first true positive is as low as 1.1 Å. BALLDock/SLICK produces three more successful docking runs and provides a better ranking

PDB ID	n-mer	ΔG_{exp} [kJ/mol]	R_{ftp}	d_{ftp} [Å]	ΔG_{ftp} [kJ/mol]	ΔE_{ftp} [kJ/mol]	R_{min}	d_{min} [Å]	ΔG_{min} [kJ/mol]	ΔE_{min} [kJ/mol]
1J4U	1	-18.24	1	0.67	-17.71	0.53	1	0.67	-17.71	0.53
5CNA	1	-22.18	26	1.27	-21.17	1.01	35	0.89	-27.36	5.18
1GIC	1	-19.25	1	0.63	-19.36	0.11	3	0.33	-21.08	1.83
1QDO	2	-28.45	–	–	–	–	139	2.03	-20.19	8.26
1QDC	2	-22.18	2	0.80	-23.46	1.28	2	0.80	-23.46	1.28
1ONA	3	-30.96	1	1.47	-29.91	1.05	38	1.04	-40.24	9.28
1DGL	3	-34.41	1	1.03	-27.89	6.52	3	0.75	-30.00	4.41
1AXZ	1	-18.20	–	–	–	–	83	2.24	-24.76	6.56
1AX0	1	-17.90	4	1.22	-18.12	0.22	129	1.14	-19.95	2.05
1AX1	2	-18.80	–	–	–	–	9	3.43	-18.47	0.33
1AX2	2	-22.70	–	–	–	–	168	2.00	-28.45	5.75
2BQP	1	-14.00	–	–	–	–	19	2.15	-15.80	1.80
1BQP	1	-16.60	–	–	–	–	45	1.91	-15.16	1.44
1QF3	1	-16.96	–	–	–	–	15	1.51	-14.57	2.39
2PEL	2	-17.76	5	0.88	-24.16	6.40	12	0.69	-24.26	6.50
1EHH	3	-21.34	–	–	–	–	1	5.67	-26.24	4.90
1EN2	4	-23.43	–	–	–	–	26	9.41	-38.04	14.61
1K7U	2	-21.34	1	1.40	-32.38	11.04	5	1.19	-32.09	10.75
mean	–	–	4.67	1.04	–	3.13	25.33	0.83	–	4.65

Table 5.15.: SLICK re-evaluation of calibration set structures generated with FlexX. Please see Tab. 5.9 for an explanation of the columns.

and deviation, but the differences are rather small.

In summary, the comparison of several different docking methods showed that SLICK is able to enhance results on the energy level as well as in scoring putative binding poses. Tab. 5.17 and 5.18 summarise this comparison. On the calibration set, using SLICK enhances the results of AutoDock and FlexX in terms of ranking, RMSD, and energy calculation. For FlexX, the correlation coefficient of the energy calculations also improve when using SLICK, although the correlation of BALLDock/SLICK, FlexX and FlexX/SLICK results are not very good. Astonishingly, the correlation of AutoDock is much better and is decreased when using SLICK on AutoDock results. This contradicts the fact that SLICK was calibrated on exactly this data and produced very good results, even in cross-validation. AutoDock's energies are off by over 20 kJ/mol, but if this constant is valid for all sugar binding predictions with their energy function, the reasons for this good correlation in comparison to the significantly worse ones produced with SLICK should be investigated.

The performance differences are less pronounced on the docking set. These calculations were only done with BALLDock/SLICK and FlexX. BALLDock/SLICK is able to successfully dock three more complexes than FlexX, but the mean R_{ftp} differs by only 0.37 in favour of BALLDock/SLICK. Looking at the deviation, the difference between the mean d_{ftp} only 0.28 Å, again in favour of BALLDock/SLICK.

PDB ID	n-mer	FlexX				BALLDock/SLICK			
		R_{ftp}	d_{ftp}	R_{min}	d_{min}	R_{ftp}	d_{ftp}	R_{min}	d_{min}
		Plant Lectins							
1KJ1	1	10	1.41	17	1.37	1	0.92	11	0.46
1KUJ	1	1	0.94	2	0.53	1	0.54	1	0.54
1MVQ	1	1	0.67	81	0.59	1	0.51	4	0.29
1WBL	1	–	–	5	1.84	1	0.31	3	0.31
1FNZ	1	1	1.21	1	1.21	1	1.00	2	1.00
1C3M	2	1	1.25	5	0.43	1	0.27	1	0.27
1GZC	2	–	–	44	3.48	1	0.93	2	0.45
1JOT	2	25	1.25	197	0.91	1	1.43	12	0.93
1PUM	1	6	1.06	23	1.02	10	0.74	114	0.51
1PUU	2	1	1.07	14	0.98	17	0.43	29	0.40
1HKD	1	1	0.66	10	0.50	1	1.25	11	0.70
1RIN	3	9	1.39	9	1.39	2	0.86	3	0.44
1OFS	2	1	1.07	13	0.53	7	1.25	7	1.25
mean	–	5.18	1.09	32.38	1.14	3.46	0.80	15.38	0.58
		Non-Plant Lectins							
1DIW	1	–	–	179	2.58	17	1.13	202	1.01
1GLG	1	1	1.14	23	0.50	1	0.42	75	0.23
1K12	1	3	0.67	3	0.67	3	0.74	7	0.21
1NL5	1	1	1.47	10	0.57	1	1.03	1	1.03
2GAL	1	–	–	17	1.97	(117)	(0.54)	152	0.26
1C1L	2	1	0.84	1	0.84	1	0.68	1	0.68
1SLT	2	1	1.44	1	1.44	1	1.10	8	0.75
mean	–	1.40	1.11	33.42	1.22	4.00	0.85	63.71	0.60
total mean	–	4.00	1.10	32.75	1.17	3.63	0.82	32.30	0.59

Table 5.16.: Comparison of results on the docking set. For FlexX and BALLDock/SLICK, the rank of the first true positive (R_{ftp}) and deviation of the first true positive (d_{ftp}) are given. Additionally, rank and deviation of the structure with minimal RMSD are shown (R_{min} and d_{min}).

Method	Successfully docked	R_{ftp} (mean)	d_{ftp} (mean)	ΔE_{ftp} (mean)	R
AutoDock	14	33.07	1.01	23.17	0.834
AutoDock/SLICK	16	8.31	0.94	7.59	0.786
BALLDock/SLICK	17	1.53	0.85	3.60	0.602
FlexX	9	12.89	1.08	11.43	0.490
FlexX/SLICK	9	4.67	1.04	3.13	0.616

Table 5.17.: Statistical data on the performance of different methods applied to the 18 structures of the SLICK calibration set.

Method	Successfully docked	R_{ftp} (mean)	d_{ftp} (mean)
BALLDock/SLICK	19	3.63	0.82
FlexX	16	4.00	1.10

Table 5.18.: Statistical data on the performance of different methods applied to the 20 structures of the SLICK extended docking set.

5.5.4. A Programme for Finding Sugar Binding Sites

If the binding site of a protein is not known, a *global docking* has to be performed, *i. e.* the ligand is docked to the whole receptor surface. Global docking can be very time-consuming. One way to reduce the necessary computing power is the deployment of an automated pocket finder in order to conduct only local searches after identifying regions of high interest.

The idea behind the programme LecXplorer [147] for finding sugar binding sites is based on the Hammerhead programme by Jain, Ruppert and Welch [148, 149] and the GRID programme by Goodford [150] for examination of binding pockets. While the methods for screening and probing follow the ideas of Hammerhead, the potentials and the scoring function are partially based on GRID. Clustering methods and the adaptation to protein-carbohydrate complexes were implemented independently from these previous efforts. In contrast to these other programmes, LecXplorer does not consider hydrogen atoms. The idea was to create a pocket finder that is independent of hydrogen positions which are in most cases optimised with force fields, thus introducing an additional level of parameterisation.

The programme roughly works as follows: The surface is scanned with probe groups and scores are computed for every probe group position. Positions with high scores are then clustered in order to find regions with high affinity. The clustering is done in two steps in order to find clusters of the correct size and form. The overall scheme is as follows:

1. Find screening points on the surface
2. Compute scores for probe groups placed on these points
3. Cluster points with high scores
4. Rank clusters according to their overall score

Scores are composed of weighted individual terms in the same fashion as ordinary scoring functions are. The main difference between scoring functions like SLICK/score and the scoring function we use in the pocket finding programme is that the probe groups do not contain information about the geometry of a putative ligand. Therefore, the terms included in the scoring function have to be purely radially symmetric, which poses a certain limitation to terms like the CH$\cdots\pi$ interaction which is completely based on geometric considerations.

Scoring function

The scoring function consists of four contributions which are weighted according to the results of a manually conducted simple optimisation scheme for the contribution weights. The potentials were chosen from the original GRID scoring function with the SLICK scoring function in mind while at the same time focusing on speed. Nevertheless, the most important goal was the identification of suitable binding sites on lectin surfaces.

The scoring function of the binding site finder consists of a van der Waals contribution, a term scoring possible hydrogen bonds, simple electrostatics and a CH$\cdots\pi$-like potential which basically resembles the SLICK/score function. The Van der Waals potential is the standard AMBER Lennard-Jones term. The electrostatics term scores polar interactions only and does certainly not include solvation effects. Electrostatic interactions are computed as Coulomb interactions with the potential used in the GRID implementation. This formulation (eqn. (5.3)) assumes a planar interface between a homogeneous protein phase of dielectric constant ε_p and homogeneous solvent of ε_s.

$$E_{\text{es}} = \frac{q_i q_j}{K\varepsilon_p}\left(\frac{1}{d} + \frac{\frac{\varepsilon_p - \varepsilon_s}{\varepsilon_p + \varepsilon_s}}{\sqrt{d^2 + 4s_i s_i}}\right) \tag{5.3}$$

The electrostatic interaction term estimates the spatial arrangement of protein atoms according to approximations for macromolecules by Hopfinger [151] using the so-called nominal depths s_i and s_j of protein atoms i and j. For details on the spatial approximation please see the original publication [150] of the GRID method.

Hydrogen bonds are scored with a pair potential similar to the one used in van der Waals calculations:

$$E_{\text{hb}} = \left(\frac{A_{ij}^{\text{hb}}}{r_{ij}^6} - \frac{B_{ij}^{\text{hb}}}{r_{ij}^4}\right) \tag{5.4}$$

In the original GRID implementation this potential had a directed term including hydrogen bond angles. Since there are no hydrogens in the structures used by this pocket finder, optimal angles are assumed.

The CH$\cdots\pi$ part of the scoring function is a flattened version of the CH$\cdots\pi$ terms used in SLICK/score and SLICK/energy. Because the hydrogen position is not known, an "ideal" hydrogen is attached to the C probe group which is directed towards the centre of the aromatic ring. This is in some sense the same model used in hydrogen bonds assuming ideal angles.

Finding probe group positions

Finding suitable positions for probe groups is quite important in order to obtain reasonable interaction points that provide the basis for finding high-affinity areas on the protein surface. We chose to use points on the solvent accessible surface (SAS) of the receptor in question. The SAS defines the surface which is still accessible for solvent molecules thus representing the nearest distance for any atom approaching the receptor. The implementation in BALL provided us with a convenient way of creating a raster of points over the protein surface which can be used to screen the receptor.

Clustering

Surface points are clustered in two steps. First, small and compact clusters are searched employing single linkage clustering. These small clusters represent high affinity regions on the protein surface. Second, proximate small clusters are combined to larger clusters using centroid method clustering. These larger clusters then represent the actual binding site for the ligand.

PDB ID	Rank with CH···π	Rank w/o CH···π	PDB ID	Rank with CH···π	Rank w/o CH···π
Calibration set					
1J4U	13	2	1AX1	2	3
5CNA	2	3	1AX2	1	3
1GIC	1	1	2BQP	2	5
1QDO	2	7	1BQP	1	3
1QDC	2	3	1QF3	1	34
1ONA	3	8	2PEL	2	20
1DGL	2	1	1EHH	2	12
1AXZ	1	1	1EN2	4	34
1AX0	1	5	1K7U	6	18
Test set					
4GAL	1	13			
5GAL	2	15			
1I3H	2	7			

Table 5.19.: Ranks of the clusters resembling the lectin binding site. Ranks are given with and without CH···π integration in the scoring function of the pocket finder.

Single linkage clustering (SLC) and centroid method clustering (CMC) are both agglomerative clustering algorithms but differ in the distance function used for combining smaller clusters. SLC considers the minimum distance between any two elements from two different clusters. The method is very fast and well suited for finding small clusters. CMC uses the distance of the centres of gravity of two clusters for choosing the clusters which have to be merged. This method needs more computing time because the centres have to be recalculated in every clustering step. As the number of points is greatly reduced after the first clustering step, these clusters of clusters can be found efficiently.

Size and shape of the resulting clusters are controlled by defining upper limits on the distances allowed in one cluster. The first clustering step is supposed to find small high affinity regions. Consequently, the upper limit on point distances has to be rather small compared to ligand size. The second step combines these small high affinity regions to “sets” of attractive binding points which then represent the whole binding site. Thus the upper limit of the second step has to be chosen in the range of actual ligand dimensions.

Results

The cluster parameters were optimised manually on the calibration set and then tested on unrelated structures. Additionally, a comparison between the scoring function and the same term without CH···π contributions was conducted. Clusters were ranked according to their binding affinity and compared to the actual binding site found in the crystal structures.

Using 0.73Å as upper limit for the SLC step and 4.0Å for CMC and the SLICK/score scoring function, the known binding sites were found among the four top ranked clusters for all lectins of the calibration set with only two exceptions. In six cases, the top ranked cluster resembled the

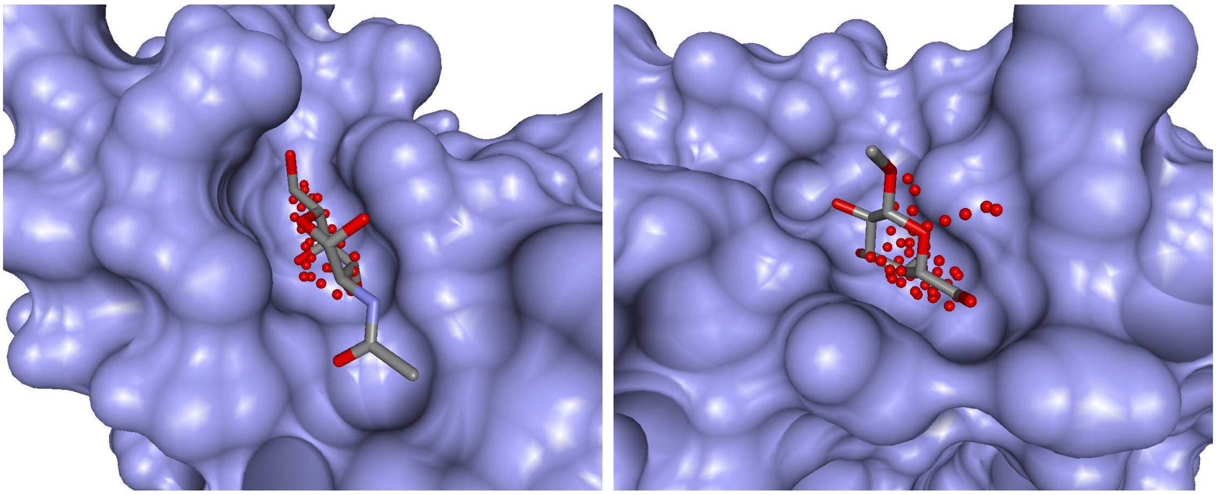


Figure 5.17.: Exemplary results of the LecXplorer programme: The best cluster (red balls) of complexes 1AX0 (left) and 1QF3 (right) with the ligand in the protein's binding site.

actual binding site. The binding site of AIA (PDB ID 1J4U) was only found at rank 13, which is probably due to the fact that there are no aromatic side chains in the binding site and thus no CH$\cdots\pi$ interactions are involved in binding. Therefore, the highly weighted CH$\cdots\pi$ contribution chose regions with aromatic side chains near the lectin surface. The binding site of WGA (PDB ID 1K7U) was found at rank six.

For a first validation, LecXplorer was used on three complexes that are not included in the calibration process Applying the pocket finder to the two hGal-7 complexes (PDB IDs 4GAL and 5GAL) showed that it also works for lectins not in the training set and even not in the domain of plant lectins. For a ConA/Man$_2$ complex (1I3H), LecXplorer identified the pocket at rank 2.

An interesting question is whether the inclusion of a CH$\cdots\pi$ term improves results. As shown in Tab. 5.19, using the scoring term without CH$\cdots\pi$ interactions resulted in a significantly worse identification of binding sites. As expected, complexes with a large number of CH$\cdots\pi$ interactions, *e. g.* UDA and WGA complexes, were ranked very poorly. This behaviour is consistent in the calibration set as well as the small validation set.

Knowing that the scoring function creates decent results, a second validation was conducted by applying LecXplorer to the complexes of the docking set. The results of this broader validation are given in Tab. 5.20. Although most binding sites could be identified with very good accuracy, five binding sites were more difficult to determine. Nevertheless, the mean rank of 3.7 suggests that the pocket finder provides sufficient accuracy in docking environments.

Plant Lectins		**Non-Plant Lectins**	
1KJ1	14	1DIW	14
1KUJ	1	1K12	2
1MVQ	1	1NL5	1
1WBL	8	1C1L	1
1FNZ	8	1SLT	1
1C3M	9	2GAL	1
1GZC	2	4GAL	1
1JOT	1	5GAL	2
1PUU	2	**mean**	2.88
1HKD	1		
1RIN	1	**total mean**	3.70
1OFS	3		
mean	4.25		

Table 5.20.: Results of applying LecXplorer to the docking set.

6. Discussion

The main goal of this thesis was the development of a computational model of protein-sugar interactions and the subsequent realisation of this model in a scoring function SLICK/score and an energy function SLICK/energy suitable for application in molecular docking. Furthermore, the actual effectiveness of both functions was shown by incorporation of the implemented functions into an existing docking programme and comparison of its results with the results of other docking programmes.

Both functions are empirical functions, which have to be calibrated using experimental data that were gained from databases and literature. They consist of several additive contributions, the calculation of which was implemented in C++ using the BALL framework for molecular modelling. The quality of the predictions of both functions was assessed with suitable measures yielding high quality results in the statistical assessment of both the energy function and the scoring function. Inclusion of these functions into BALLDock, a programme for docking flexible ligands into receptor binding sites employing genetic algorithms, yielded the first docking method especially designed to predict protein-carbohydrate interactions.

Thorough analysis of the two components of SLICK showed that this package is indeed a valuable tool for investigating protein-sugar interactions. The scoring function SLICK/score is able to correctly score docking candidates created by independent structure generators from different docking programmes. The energy function SLICK/energy predicts binding free energies of complex conformations near the native conformation with very high accuracy. Merging SLICK into BALLDock yielded very good docking results on an extensive set of structurally known protein-sugar complexes. In summary, the main result of this thesis is the creation of a docking method for protein-carbohydrate complexes that predicts bound conformations with high accuracy.

However, there still remain unsolved problems. The scoring function SLICK/score seems to rank almost all docking candidates it was applied to correctly. These results remain valid for candidates created with different structure generators. This means that the topography of the energy surface is at least approximatively reproduced by this very simple scoring function. Astonishingly, SLICK/energy seems to fail in ranking docking candidates correctly, although the only difference between SLICK/score and SLICK/energy lies in the treatment of solvation effects. Consequently, there must be a systematic error in the solvation handling component. The question is, which part of the solvation component does fail? Calculating correct energy contributions depends on

- accurate computational models for each energy term,
- correct implementation of the computational models, and
- adequate parameters for use in the calculations.

Each of these points has to be verified, but verification is only feasible if there is enough reliable data that can be used for comparison. Accurate experimentally determined solvation free energies are hard to obtain, even for very small molecules, let alone proteins. However, in the case of solvation models, the biggest problem is that polar and nonpolar contributions cannot be measured independently. Consequently, the verification of these models based on experimental data is not an easy task.

In literature, nonpolar solvation models are verified on sets of small molecules that do not contain polar groups, like N-alkanes. The verification is then based on the assumption that polar contributions are negligible and thus the solvation free energy from the experiment does only consist of nonpolar contributions. Models for the polar contribution to the solvation free energy are also verified on very small molecules, but this time these molecules have to be highly polar. Again, the assumption is made that the contribution in question is so strong that it dominates the whole solvation free energy and that other contributions can be neglected. In some cases, nonpolar contributions are calculated by established methods and then integrated into the verification of polar models, which adds uncertainty to the verification process. For proteins, polar models even are verified by comparing calculated energies to other calculated energies, assuming that polar contributions computed with FDPB models are accurate.

But not only are the two contributions to solvation free energy not directly accessible. The computational models in question employ different parameters in their calculation, namely atomic radii and atomic partial charges. These quantities themselves are approximations of the real world, which means they also are results of computational models. There is a set of atomic radii by Bondi [105] that contains experimental van der Waals radii for nonbonded atoms, but these are hardly applicable for bonded atoms in different chemical environments. In the simplified view of molecular mechanics, an aliphatic carbon will most probably have a different radius than an aromatic one. Therefore, a wide variety of different parameter sets for atomic features exist, which in most cases are the result of fitting computational models against experimental data. In an exaggerated manner, one could say that the verification of these computational models depends on computational models, which in turn depend on computational models and some experimental data.

To make things worse, the models available for solvation effects are based on the description of macroscopic effects but are applied to microscopic systems. It is not evident that these macroscopic models are transferable to molecular or even atomic length scales. In some formulations, additive constants are included to account for effects on small length scales, but this can be approximative at best. Nevertheless, the models for polar and nonpolar solvation models used and analysed for this thesis seem to represent the state-of-the-art regarding the calculation of solvation effects in molecular systems.

In this thesis, different combinations of models for solvation free energies were tested on sets of small molecules because these were the only reliable sources of data at hand. The tests were performed for several different models using several parameter sets. The calculations suggested that the combination of models implemented in SLICK/energy should provide the best prediction accuracy. Obviously this is only the case for conformations that are close to the native one. Using SLICK/energy for rescoring docking candidates is not possible at the moment. At the same time, the calibration of an energy function that neglects solvation produces drastically worse correlations and predictions. Consequently, the conclusion is that solvation terms are very important to the energy function and that the existing terms have to be analysed and revised. Judging from the analysis of the different contributions to ΔG calculated by SLICK/energy, it seems that the polar contribution has the strongest influence on erroneous energy predictions.

This suggests that the Jackson-Sternberg model using PARSE parameters is the culprit. On the other hand, the observed effects might be the result of overfitting. But further investigation is necessary to confirm one of these hypotheses.

One possibility to rectify the behaviour of the solvation term could be the readjustment of the parameters used in the calculations. Although the parameters used so far seem to produce accurate numbers on small molecules, they might not be efficient in docking environments. A new parameterisation specifically designed for SLICK could provide better scoring efficiency. Nevertheless, this additional analysis goes beyond the scope of this thesis and should be addressed by further research.

In addition to the influence of parameter sets of SLICK itself, the optimisation results gained with AMBER, Glycam and GAMESS cannot be neglected. Hydrogen positions are crucial to CH$\cdots\pi$ and hydrogen bond calculations. The force fields used in optimisation do not include these effects in their computations. Thus, the results of the calibration of SLICK might not be optimal regarding these two contributions. A recalibration of SLICK with structures optimised with a force field capable of reproducing these interactions might yield better results with respect to ligand placement. The first force field including CH$\cdots\pi$ is the CHARMM force field, which was just recently adapted. Literature does not yet provide information on how this new formulation of CHARMM integrates with the carbohydrate solution force field (CSFF) which is based in the previous CHARMM energy function. Further investigation is needed to assess the effectiveness of this approach for a possible reparameterisation of SLICK.

Another important issue for docking sugars in lectin binding sites and small ligand docking in general is the treatment of water-mediated hydrogen bonds. Several studies emphasise the importance of these water bridges for lectin-sugar complexes. The analyses presented in the results of this thesis support the observation that ligands are strongly depending on these types of interactions if parts of the ligand are extending into the solvent. For sugar-oligomers this is seemingly often the case. The handling of water-mediated hydrogen bonds is still an unresolved problem in molecular docking, but it is predominantly an issue of the structure generator, not the energy function. SLICK offers a term for hydrogen bonds that is in principle applicable to water bridges, but it depends on water molecules placed in or around the binding site in order to function properly.

Although the current results of SLICK are very encouraging, there is still much room for improvement. First of all, the calibration and the validation of SLICK needs more reliable experimental data. One way of enlarging the data basis include non-plant lectins or general sugar binding proteins into the calibration set. But this approach would also demand an additional model for the coordination of ligands by metal ions in the binding site, because in animal lectins and enzymes, the binding is heavily influenced by such ions, which directly interact with the ligand. In plant lectins, metal ions usually influence the structure of the binding site and are several Ångström away from the ligand, which justifies the assumption that considering the electrostatic effects only is sufficient. Additionally, new experimental data should be generated to close the gap between available structural and thermodynamic data.

The van der Waals model used in this study employs a very simple form of softening energy contributions. Although the softening approach improved results, there might be better ways of doing so. Until now, the standard Lennard-Jones form of the van der Waals interaction model was used, which is based on a repulsion term with exponent 12. Changing the exponent to 10 or 9 could produce better results, but this also requires readjusting the van der Waals parameters.

Including CH$\cdots\pi$ interactions into the energy function in order to cover ring stacking effects distinctly enhanced the prediction quality for lectin-sugar complexes. However, aliphatic-aromatic ring stacking is only one type of ring stacking observed in molecular complexes. Ring stacking of two aromatic systems might be just as important in ligand docking. There are many drugs which rely on this so-called $\pi\cdots\pi$ stacking when binding to their targets. The FlexX energy function already includes aromatic interactions based on interaction surfaces. Developing a geometric model similar to the CH$\cdots\pi$ component of SLICK could improve on the FlexX approach and reproduce $\pi\cdots\pi$ stacking more accurately. An additional interaction term for $\pi\cdots\pi$ stacking would broaden the applicability of SLICK far beyond the lectin-sugar case.

A more ambitious goal could be the transformation of SLICK into a force field. With such a force field, the optimisation of docked complexes in the binding site as well as thorough energetic analysis of the receptor surface and dynamics simulations of such systems would be possible. Unfortunately, this transformation is not as straight-forward as it seems. An energy function that is supposed to act in a force field has to be differentiable. Unfortunately, some of the models used in SLICK do not have this property. The geometric approaches to CH$\cdots\pi$ and hydrogen bonds can be turned into differentiable functions with little effort as long as the sigmoid form is used. The solvation component poses a bigger problem. There are derivatives for electrostatics terms, but nonpolar models at the SPT level have not been differentiated for use in force field calculations so far. Creating a consistent differentiable solvation term thus needs further work.

Considering the docking efforts, trying a different structure generation algorithm than the genetic approach could improve results or at least reduce computational demand of the method. At the moment, docking runs using the genetic algorithm for structure generation need a lot of computing time, especially when compared to very fast algorithms like FlexX. On the other hand, the extreme flexibility of oligo-sugars combined with the shallow binding grooves of lectins supposedly are a serious problem for construction algorithms. FlexX uses a heuristic based on energy estimates in order to avoid combinatorial explosion. If energy differences between two distinct candidate conformations are very small, the algorithm is likely to fail because the decision for the "correct" candidate cannot be made. Judging from the comparison between structures generated by FlexX and BALLDock/SLICK, it seems that the construction algorithm is more dependent on steric features of the binding site of the receptor. However, this could change if the energy function of FlexX would include additional interactions.

In conclusion, the results obtained in this thesis show that sugar docking is possible with high accuracy if the interactions that have proven important to lectin-sugar binding are included in the calculations. With BALLDock/SLICK, a first docking method for protein-carbohydrate complexes has been developed based on the thorough investigations that built the basis for SLICK.

With this method, various pharmaceutical applications could be devised. One example stems from a research project called GELENA ("Nichtvirales Gentransfersystem auf Basis Lektin-funktionalisierter Nanopartikel") conducted at Saarland University from 2000 to 2002. The goal of the project was the creation of a non-viral gene transfer system based on silica nanoparticles functionalised by specifically designed lectins. It is based on the observation that cells can be identified very specifically by sugars coating the cell surface. Because lectins bind to sugars very specifically, the idea was to load charged nanoparticles with DNA, functionalise the particles with specially designed lectins that were built for identifying a certain cell type and thus make the loaded particles bind to exactly the type of cell which has to be targeted (see Fig. 6.1). These nanoparticles would then be internalised triggered by the binding process and release the loaded

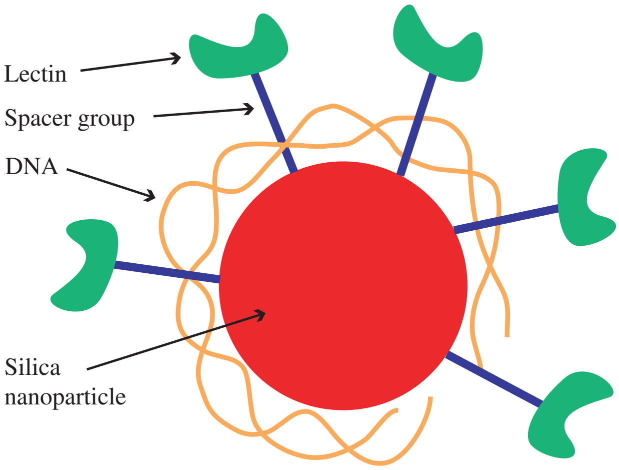


Figure 6.1.: Lectins as functionalising groups in a gene transfer system. A lectin with designed specificity is connected to a nanoparticle by a spacer group. The charged nanoparticle is then loaded with DNA fragments that have to be transported into a cell.

DNA fragments into the cytosol, where it would eventually be transported into the cell nucleus and replicated through mitosis.

For such a system to work properly without having to synthesise and experimentally test every thinkable lectin, it is obviously necessary to have means of predicting binding specificities of artificially designed lectins derived from natural ones. Therefore a tool for molecular docking designed for the computation of binding modes and energies of protein carbohydrate complexes is essential.

Another example presents drugs based on sugars. *Helicobacter pylori* is a wide spread pathogen which resides in the stomach and is held responsible for gastric ulcer as well as some forms of

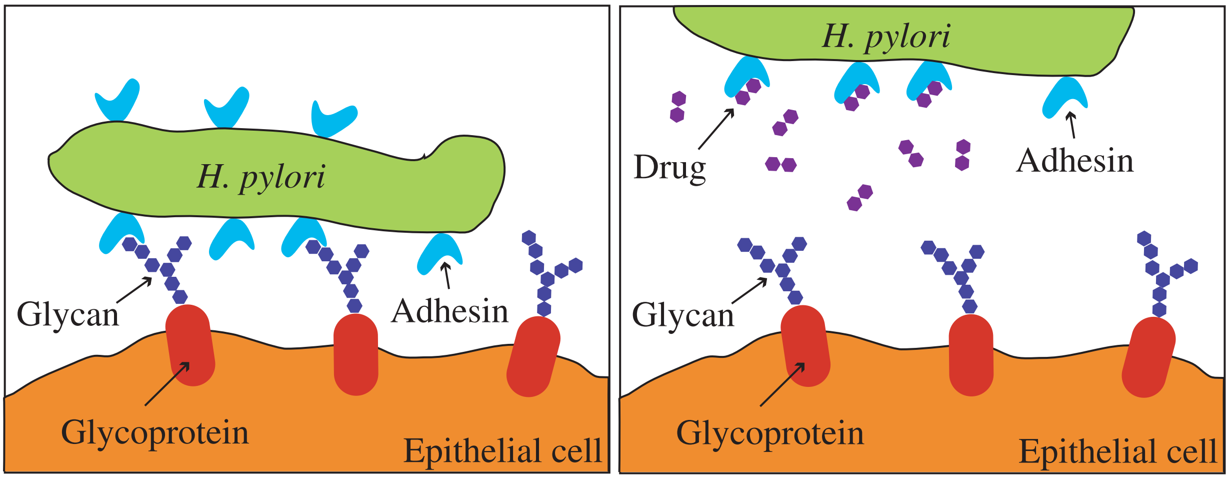


Figure 6.2.: Sugar mimetics as anti-microbial drugs. Please see text for details.

gastric cancer. This bacterium binds to gastric epithelial cells by lectinlike adhesins binding specifically to certain carbohydrate epitopes on the cells. Usual approaches to cure *H. pylori* infections are based on antibiotics combined with drugs lowering the gastric acidity, but this kind of therapy has some drawbacks. There are the general negative effects of antibiotics on patients along with increasing allergy and drug-resistance of pathogens.

Instead of using traditional antibiotics alone for eradicating *H. pylori* from patients' gastrointestinal tract, a combination with drugs inhibiting the adhesion of the bacterium to epithelial cells would be most promising. One way of doing this is the design of carbohydrates that bind to the adhesins of the pathogen surface and inhibit their binding to the epithelial cells. Thus, the bacterium would be prevented from settling down in the gastric environment. There are experiments of using porcine milk which contains similar carbohydrate epitopes as presented on the surface of epithelial cells in order to block adhesion of *H. pylori* [152]. The next step could be the design of sugars or sugar-like drugs with predefined features from the known specificity of the *Helicobacter* adhesins. The approach presented in this thesis could accelerate and simplify the development of such drugs.

A. Implementation

The methods described in this thesis were realised in C++ using BALL [103]. BALL is an ANSI/ISO a C++ framework for rapid software prototyping in molecular modelling, which was designed using state-of-the-art software engineering paradigms. It is released under an open-source license (LGPL [153]). The main goals behind the design of BALL are robustness, extensibility and ease of use. The BALL architecture is highly modular and allows for transparent integration of functionality based on well-defined and well-documented interfaces. The components comprising SLICK were integrated into BALL based on the same design goals. This chapter gives a brief overview of the development of SLICK.

A.1. Scoring Framework

BALL provides a framework for energy functions in the field of molecular mechanics (MM) force fields. Unfortunately, this framework cannot be used for SLICK because of two main reasons. First, the framework for MM energy functions relies on parameters being the same for every contributing energy term. For example, when calculating electrostatic interactions or van der Waals contributions, the radii stay the same. This is not the case in SLICK, where *e. g.* the nonpolar solvation contribution needs another set of radius parameters as the polar solvation component or the van der Waals term. Consequently, SLICK needs a new method for storing atomic parameters for every energy contribution.

Second, the MM energy functions are not designed to integrate purely intermolecular interaction terms. Calculating interaction energies with MM energy functions is only possible via the difference of bound and unbound states. Let A be the receptor and B the ligand, which form the complex AB, then the MM interaction energy $\Delta G_{\mathrm{AB}}^{\mathrm{int}}$ is given by

$$\Delta G_{\mathrm{AB}}^{\mathrm{int}} = \Delta G_{\mathrm{AB}}^{\mathrm{MM}} - (\Delta G_{\mathrm{A}}^{\mathrm{MM}} + \Delta G_{\mathrm{B}}^{\mathrm{MM}}) \tag{A.1}$$

But some terms included in SLICK, namely hydrogen bonding and CH$\cdots\pi$ contribution, calculate interactions directly. Therefore, a generalised formulation based on bound-unbound differences is not easily possible.

Consequently, a new, broader framework for energy and scoring functions in BALL is necessary. This new framework is strongly based on the MM force field implementation, but extends over the old framework by allowing the inclusion of very different energy contributions in one scoring or energy function. The new framework exclusively calculates interaction scores or energies and cannot be used for molecular mechanics at all, because there are virtually no restrictions for terms that can be included into such a function.

The class responsible for these kinds of functions is called `ScoringFunction`[1]. Every `ScoringFunction` contains a list of individual contributions, which are objects of the type `ScoringComponent`, and

[1] Please note that not only scoring functions can be implemented with this class. If the composition of contributions is chosen properly, a `ScoringFunction` can actually be an energy function

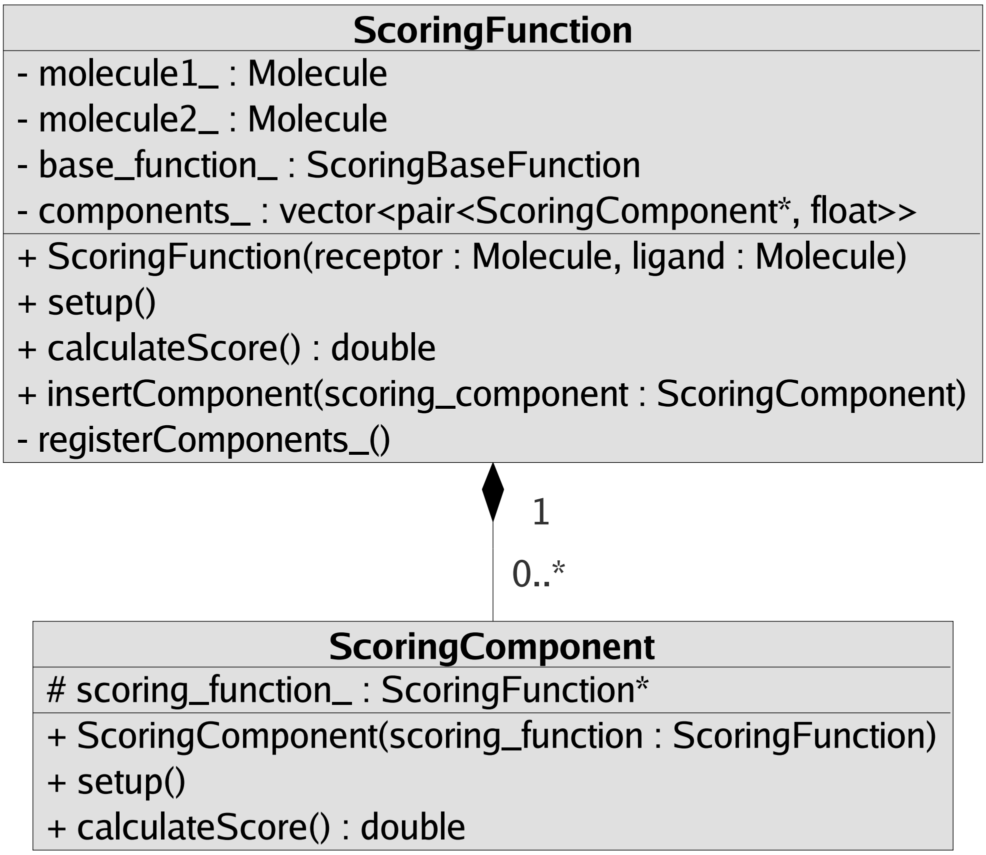

ScoringFunction
- molecule1_ : Molecule
- molecule2_ : Molecule
- base_function_ : ScoringBaseFunction
- components_ : vector<pair<ScoringComponent*, float>>
+ ScoringFunction(receptor : Molecule, ligand : Molecule)
+ setup()
+ calculateScore() : double
+ insertComponent(scoring_component : ScoringComponent)
- registerComponents_()
1
0..*
ScoringComponent
# scoring_function_ : ScoringFunction*
+ ScoringComponent(scoring_function : ScoringFunction)
+ setup()
+ calculateScore() : double

Figure A.1.: A simplified UML diagram of `ScoringFunction`.

two objects of type `Molecule` representing the receptor and the ligand, respectively. Fig. A.1 shows a simplified UML diagram of the class `ScoringFunction`.

With this framework, a more flexible approach to calculating scoring or energy functions is possible, including specialised parameterisations for every component and even every molecule. A `ScoringFunction` can integrate models that rely on calculating energy differences of whole systems as well as methods that calculate interaction scores or energies directly. However, the construction has two drawbacks. First, a `ScoringFunction` is not an energy function of a `ForceField`. Thus, it cannot be integrated in optimisation or molecular dynamics, even if the object calculates a differentiable energy function. Second, a `ScoringFunction` cannot benefit from the fast and optimised calculation methods used in `ForceField`, because the individual energy terms might not fit into the somewhat narrow definition of an energy function for molecular mechanics.

Using `ScoringFunction` for building a computational model is very easy. Every object of type `ScoringComponent` can be integrated into a scoring function with the help of the method `registerComponents_()`, which has to be called in the constructor of the `ScoringFunction` constructor. The `setup()` will then initialise every component with its own setup function. With `calculateScore()`, the actual scoring calculation can then take place.

A small example illustrating the usage of this framework is presented below. The first listing shows the header definition of a scoring function that only consists of one component. In this example, the scoring function just computes CH···π scores.

```
#include <BALL/SCORING/COMMON/scoringFunction.h>

namespace BALL
{
  class CHPIScoring
    : public ScoringFunction
  {
    public:

      CHPIScoring() throw();

      CHPIScoring(Molecule& protein, Molecule& ligand) throw();

      virtual ~CHPIScoring() throw();

    private:

      void registerComponents_() throw();

  };
}
```

In this header, the actual scoring components are not yet chosen. It merely provides the interface for something called `CHPIScoring`. It is important to declare the method `registerComponents_()` in this header, because it will be responsible for the composition of the scoring function. The actual definition of the scoring function is relatively easy. With the method `insertComponent()`, which is derived from the base class, `registerComponents_()` can integrate objects of type `ScoringComponent` into the scoring function. In all constructors, `registerComponents_()` has to be called. The listing below shows example code for our class `CHPIScoring`.

```
#include "CHPIScoring.h"
#include <BALL/SCORING/COMPONENTS/CHPI.h>

namespace BALL
{

  CHPIScoring::CHPIScoring() throw()
    : ScoringFunction()
  {
    registerComponents_();
  }

  CHPIScoring::CHPIScoring(Molecule& protein, Molecule& ligand) throw()
    : ScoringFunction()
  {
    setReceptor(protein);
    setLigand(ligand);
    registerComponents_();
    setup();
  }

  CHPIScoring::~CHPIScoring() throw()
  {
  }

  void CHPIScoring::registerComponents_() throw()
  {
    // This code defines the composition of the scoring function.
    // In this litte example, only a CHPI term is included.
    insertComponent(new CHPI(*this));
  }
}
```

The inclusion of such a scoring function into a programme is very simple. Assume you want to write a programme that just reads two molecules from files and calculate the CH···π score of the putative complex. The following listing shows such a programme.

```
#include <BALL/FORMAT/HINFile.h>
#include <BALL/KERNEL/molecule.h>
#include "CHPIScoring.h"
#include <iostream>

using namespace BALL;
using namespace std;

int main()
{
  // Load the receptor from a file.
  Molecule receptor;
  HINFile infile_receptor("receptor.hin");
  infile_receptor >> receptor;
  infile_receptor.close();

  // Load the ligand from a file.
  Molecule ligand;
  HINFile infile_ligand("ligand.hin");
  infile_ligand >> ligand;
  infile_ligand.close();

  CHPIScoring chpi_scoring(receptor, ligand);
  float score = chpi_scoring.calculateScore();

  cout << "The CH/pi score is " << score << endl;
}
```

By inserting other components of type `ScoringComponent`, the scoring function can be easily tailored for any application. Of course, this simple example did not include any code necessary to assign parameters of even check the contents of the files.

A.2. Implementation Details

This section presents some details on components that were implemented during the development of SLICK. Many techniques employed in the implementation make heavy use of the functionality provided by BALL.

The CH$\cdots\pi$ Component

The CH$\cdots\pi$ component is implemented in the class `CHPI`, which is derived from the class `ScoringComponent`. It contains the two nested classes `CHPI::AromaticRing` and `CHPI::CHGroup`, which are used for storing possible interaction partners of CH$\cdots\pi$ bridges. During the setup of this component, aromatic rings and CH groups are searched. All possible pairs of interaction partners are then stored in an STL vector. During the calculation of the scores, this vector is processed and the CH$\cdots\pi$ scores are calculated.

The nested classes provide functionality that simplifies the calculation of the necessary vector geometry. For example, `CHPI::AromaticRing` provides a helpful method `getNormalVector()`, which returns the normal vector of the plane defined by the planar aromatic ring. This vector is necessary to calculate the projection of the H atom into the ring plane.

A user may choose the base function (cf. section 3.4.1) of the scoring function through the method `setBaseFunction()` of the `ScoringFunction` interface. At the moment, two base functions are defined, the linear and the sigmoid one. The limits defining the transition interval of the base function are an option of the CH$\cdots\pi$ component and accessible via the datatype `Option` of the BALL framework. This datatype allows for easy modification of options.

Nonpolar Solvation Models

During the development of SLICK, several nonpolar solvation models were implemented and tested. The implementations are based on the processor concept that is realised in BALL. Processors are classes that apply a certain function to atoms of a system. BALL provides the necessary framework to use such constructs in a very simple way. Once a processor is defined, the method `apply()` takes care of the computation.

```
ExampleProcessor proc;
system.apply(proc);
```

Processors are derived from the generic processor base classes `UnaryProcessor` and `BinaryProcessor`. In the case of nonpolar solvation models, a unary processor was sufficient to compute the solvation effects. For example, computing the Uhlig solvation energy of a molecule is accomplished by the following code fragment:

```
PDBFile infile("infile.pdb");
System system;
infile >> system;
infile.close();

UhligCavFreeEnergyProcessor uhlig;
system.apply(uhlig);
float uhlig_energy = uhlig.getEnergy();
```

With the generic processor concept it was possible to implement one scoring component handling several different nonpolar solvation models without having to write the same code several times. Using dynamic binding, the method `calculateScore()` does not need to know, which model is chosen by the user. It just uses the interface of `UnaryProcessor` in order to compute the energy. The following code fragment illustrates the possibilities:

```
float calculateScore(UnaryProcessor* proc, System& system)
{
        system.apply(*proc);
        return(proc->getEnergy());
}

...

System system;
infile >> system;

UnaryProcessor* proc;

UhligCavFreeEnergyProcessor uhlig;
proc = &uhlig;
float uhlig_energy = calculateScore(proc, system);

PCMCavFreeEnergyProcessor reiss;
proc = &reiss;
float reiss_energy = calculateScore(proc, system);
```

Parameter Assignment

Assigning parameters to atoms is usually quite complicated. First, atom types have to be assigned to individual atoms. These atom types are determined by the atom's element and its chemical surrounding. For example, in most cases, aromatic carbons are treated differently than aliphatic ones. Second, the parameters like radius and charge have to be assigned based on that atom type.

Since element and chemical environment determine the atom type, it is often possible to define rules that unambigously identify atoms of a certain type. For this purpose, processors have been developed that assign values to atoms based on an extensive set of user-definable rules. These rules can be defined directly as an argument to the constructor of a rule processor. For convenience and easy modification, the user can create a file containing a large set of rules, based on the `INIFile` datatype and file format defined by BALL. Objects of type `ChargeRuleProcessor` and `RadiusRuleProcessor` can read such files and process them. The following listing shows some exemplary rules from the PARSE parameter set:

```
...

[ChargeRules:H]
; carboxylic acid groups
0.435 = connectedTo((-O(-C(=O))))
; hydroxyl groups
0.49  = connectedTo((-O))
; thiol hydrogens
0.29  = connectedTo((-S(-*)))

...

[RadiusRules:H]
; methyl groups at aromatic rings (e.g. methylbenzene)
; are considered as united atoms: charge 0 and radius 0
0.0  = connectedTo((-C(-*)(-*)(-*)))
; amine groups are also considered as united atoms
0.0  = connectedTo((-N(-*)(-*)))
1.0  = true()

...
```

The square brackets define a section of such a rule file. As the name suggests, the section `ChargeRules` defines rules for assigning charges to an atom, `RadiusRules` defines radius assignment. After the colon, the element of an atom is defined. In the example above, the rules apply to hydrogen atoms and are processed in the order in which they appear in the file. The first matching rule is applied to the hydrogen atom. Using a `RuleProcessor` is as simple as using the processors for nonpolar solvation. The user just needs to apply a processor to a system. Every atom of the system will be traversed and the rules will be applied to the currently processed atom.

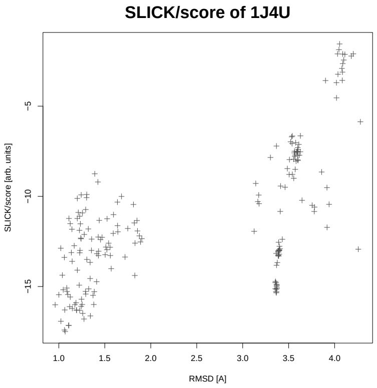
The construction of the rules is fairly straight-forward. The user defines a pattern that represents atoms that are bound to the current atom. The patterns are simple strings and can be arbitrarily long. Thus, the possible degree of detail of such a rule is virtually unlimited.

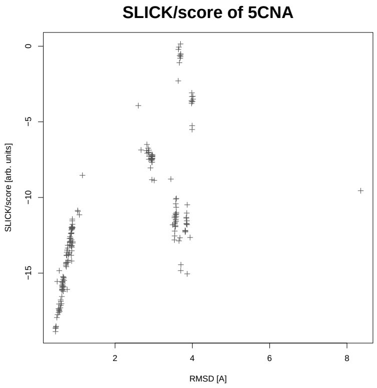
With this powerful framework, SLICK accomplishes almost all parameter assignments. PARSE parameters, Bondi radii and Glycam atom types are assigned by using rules and processors. The remaining parameters can be assigned from tables that are based on atom types.

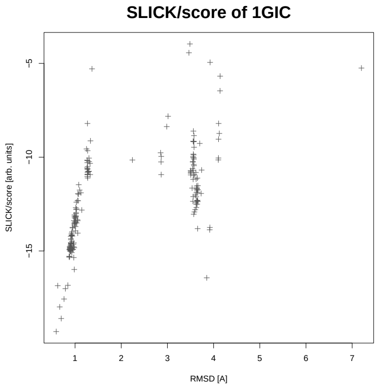
B. Detailed Results

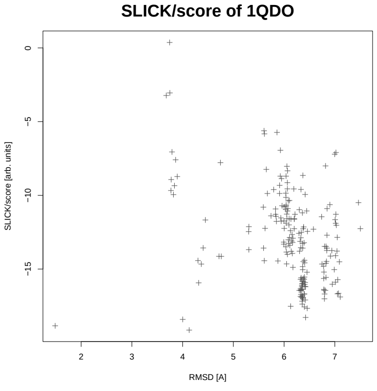
B.1. AutoDock Calibration Set Candidates

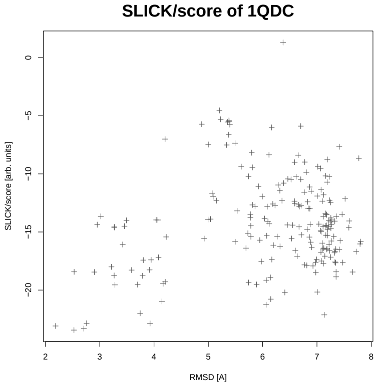
B.1.1. SLICK/score Rescoring of AutoDock Calibration Set Candidates

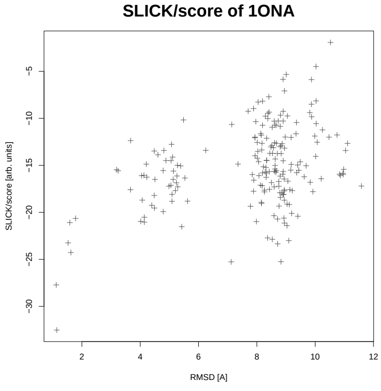




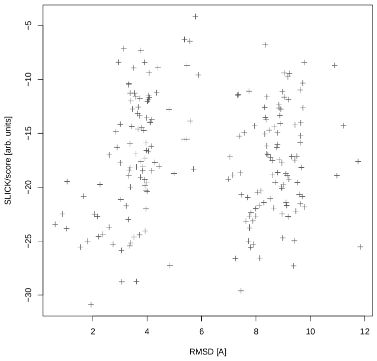




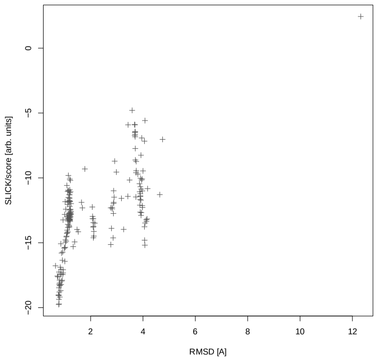




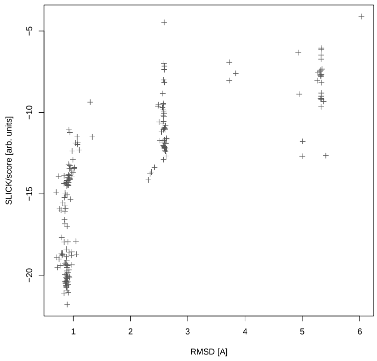
SLICK/score of 1DGL



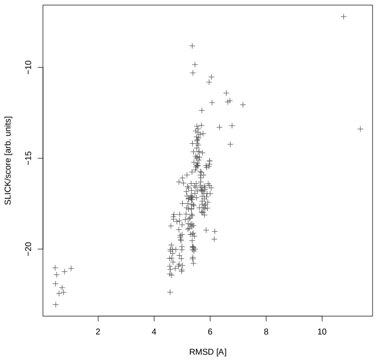
SLICK/score of 1AXZ



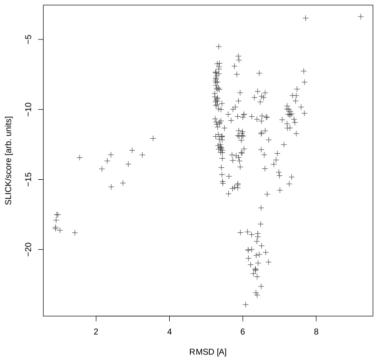
SLICK/score of 1AX0



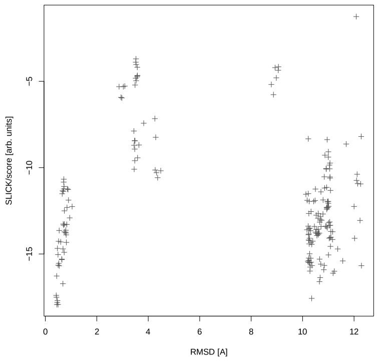
SLICK/score of 1AX1



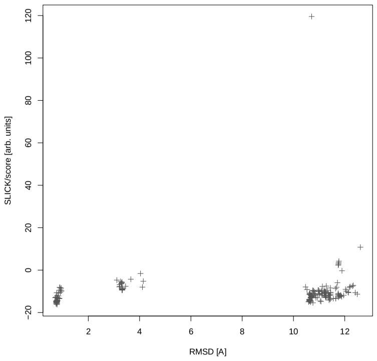
SLICK/score of 1AX2



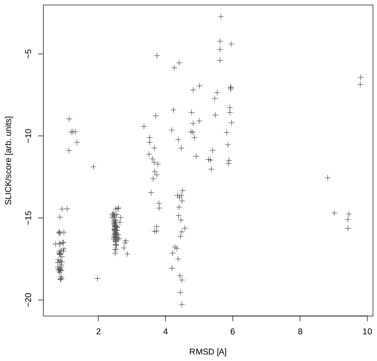
SLICK/score of 2BQP



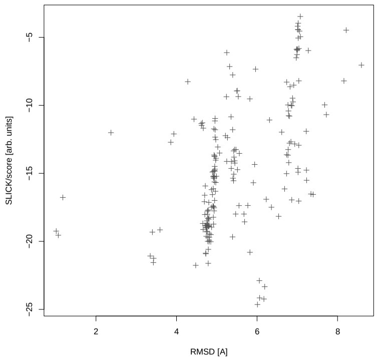
SLICK/score of 1BQP



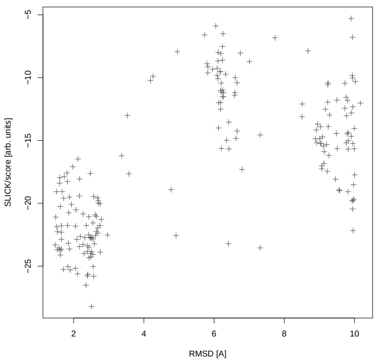
SLICK/score of 1QF3



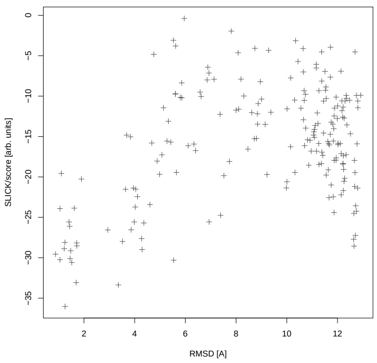
SLICK/score of 2PEL



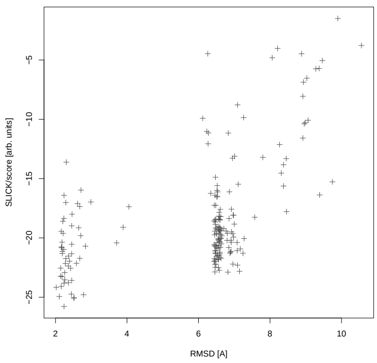
SLICK/score of 1EHH



SLICK/score of 1EN2

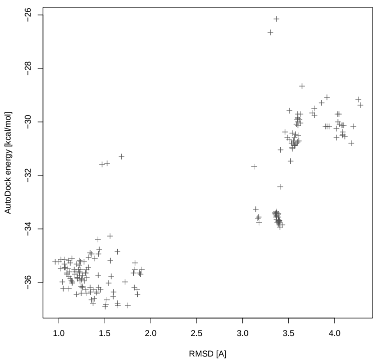


SLICK/score of 1K7U

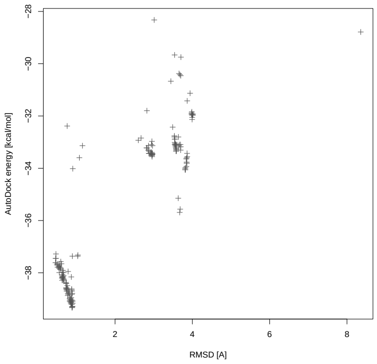


B.1.2. AutoDock Energies of the Calibration Set

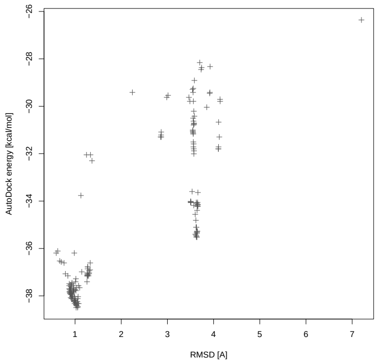
AutoDock energy of 1J4U



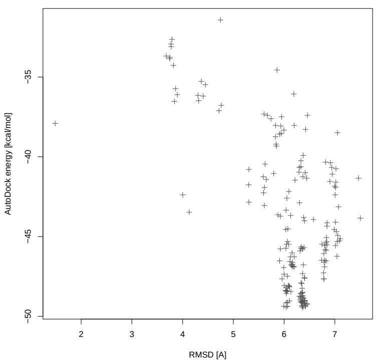
AutoDock energy of 5CNA



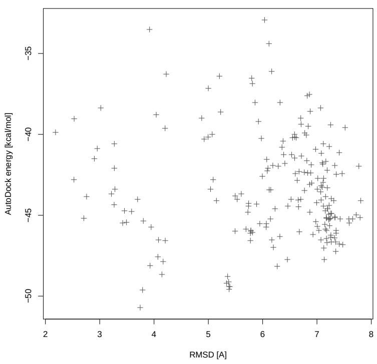
AutoDock energy of 1GIC



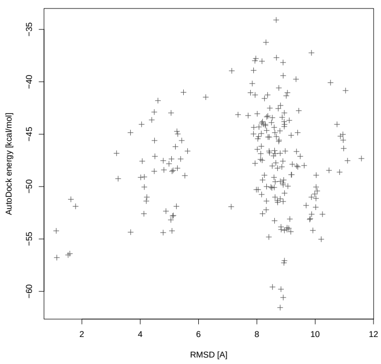
AutoDock energy of 1QDO



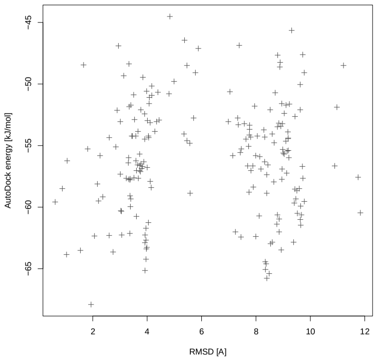
AutoDock energy of 1QDC



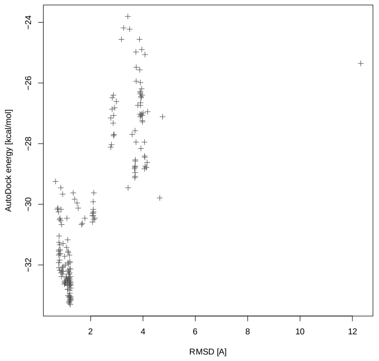
AutoDock energy of 1ONA



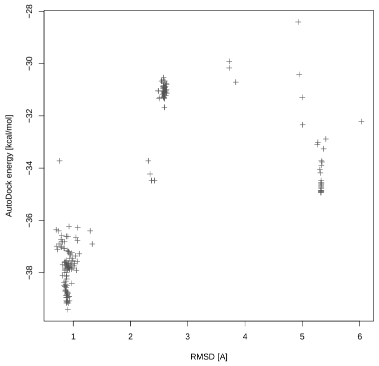
AutoDock energy of 1DGL



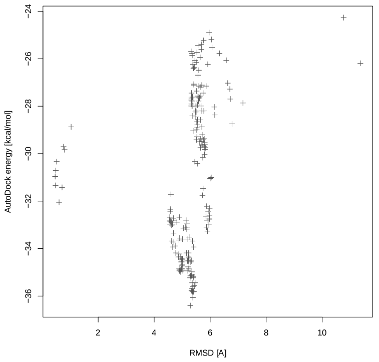
AutoDock energy of 1AXZ



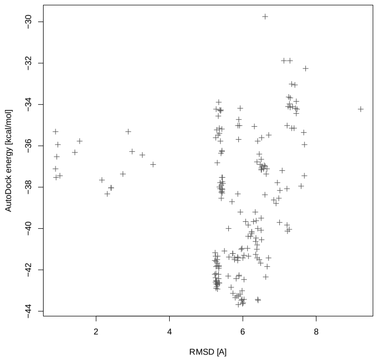
AutoDock energy of 1AX0



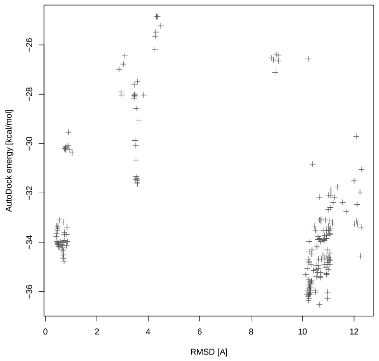
AutoDock energy of 1AX1



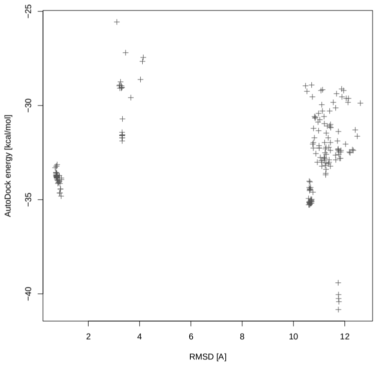
AutoDock energy of 1AX2



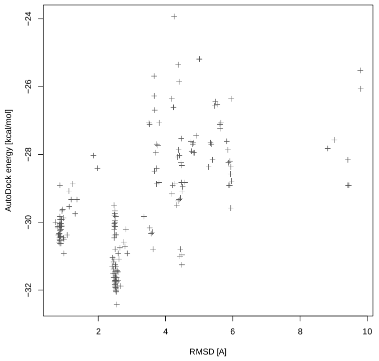
AutoDock energy of 2BQP



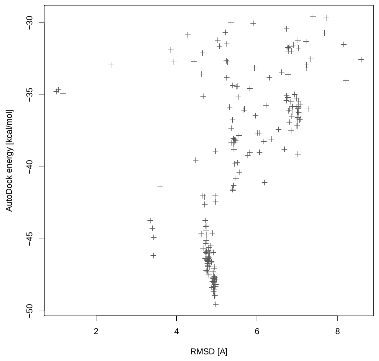
AutoDock energy of 1BQP

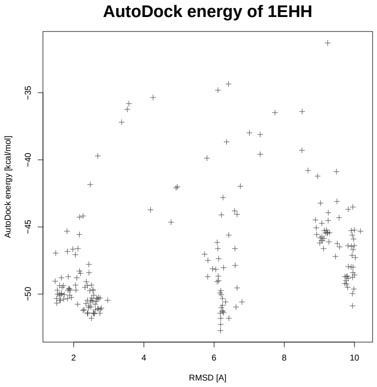


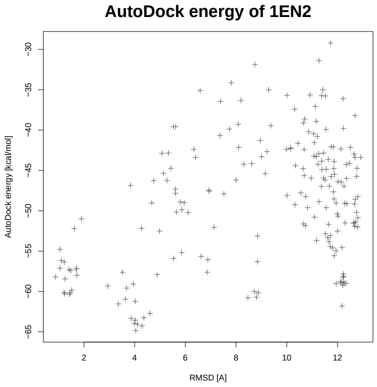
AutoDock energy of 1QF3

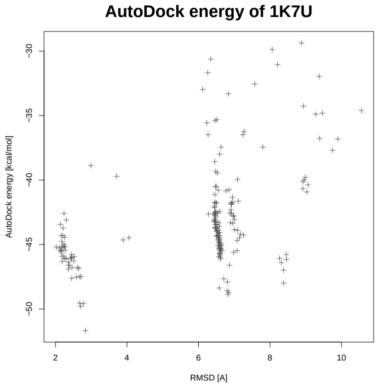


AutoDock energy of 2PEL

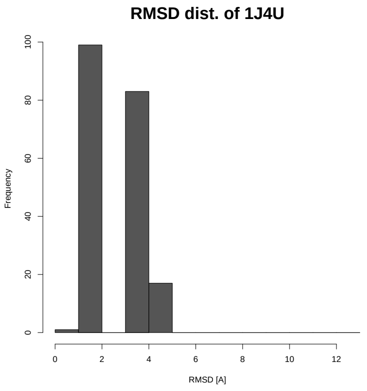


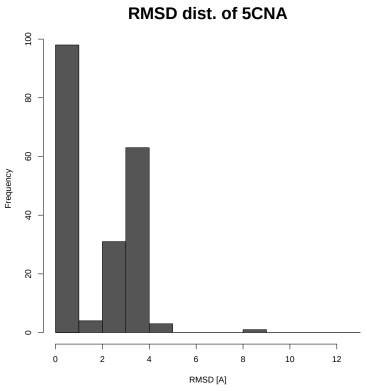

AutoDock energy of 1EHH

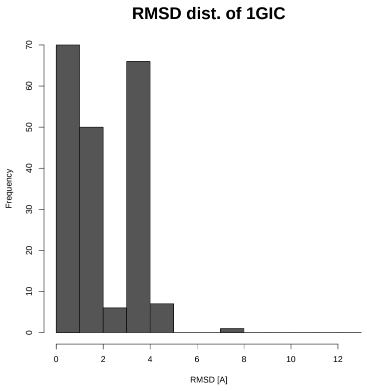

AutoDock energy of 1EN2

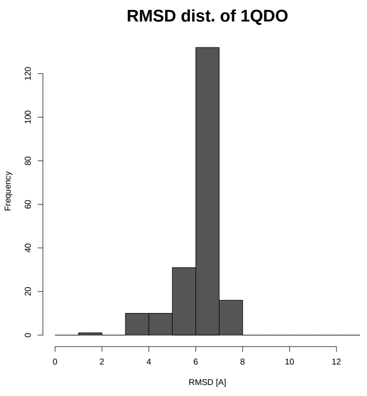

AutoDock energy of 1K7U

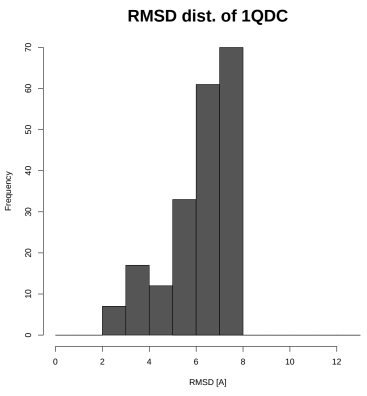
B.1.3. RMSD Distributions of the AutoDock-generated Candidates

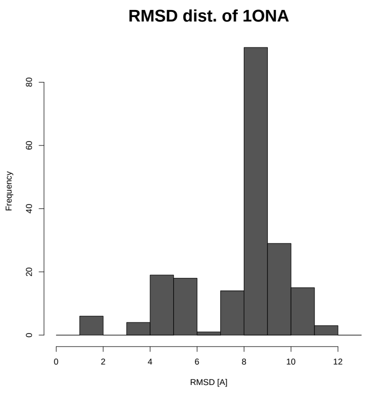

RMSD dist. of 1J4U

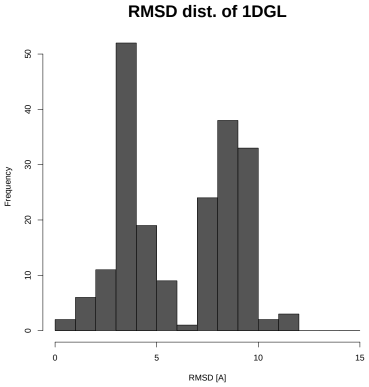

RMSD dist. of 5CNA

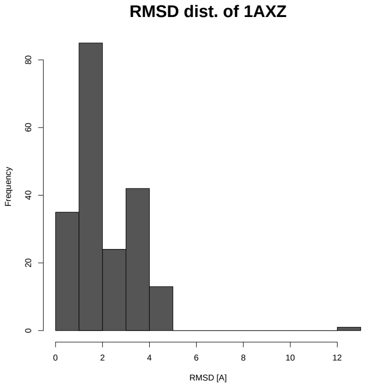

RMSD dist. of 1GIC

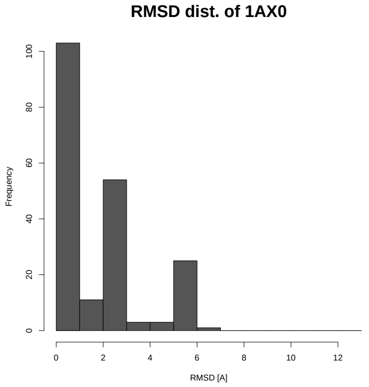

RMSD dist. of 1QDO

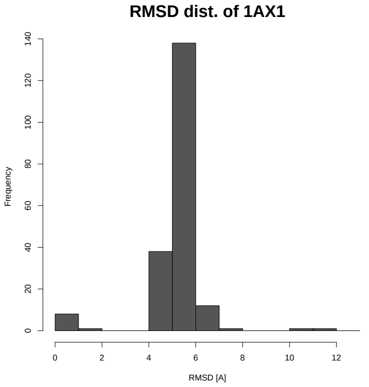

RMSD dist. of 1QDC

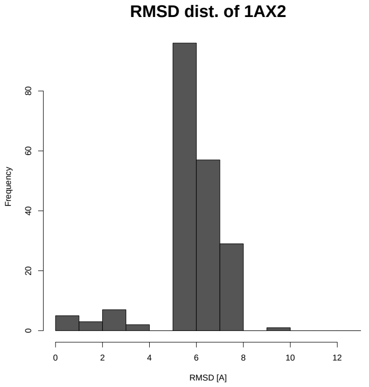

RMSD dist. of 1ONA

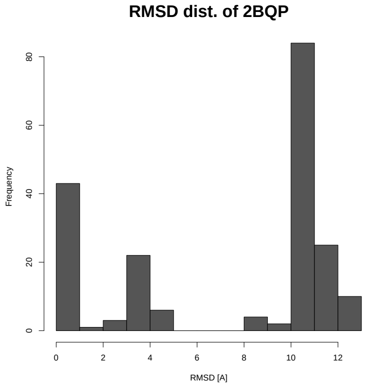

RMSD dist. of 1DGL
Frequency
0
10
20
30
40
50
0
5
10
15
RMSD [A]

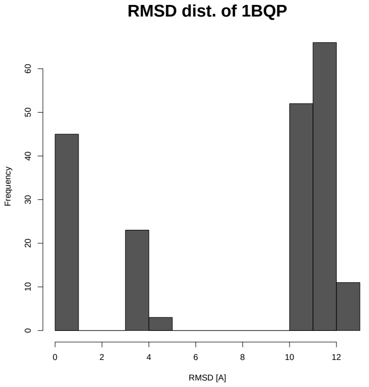

RMSD dist. of 1AXZ
Frequency
0
20
40
60
80
0
2
4
6
8
10
12
RMSD [A]

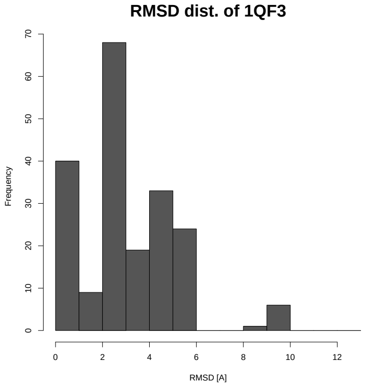

RMSD dist. of 1AX0
Frequency
0
20
40
60
80
100
0
2
4
6
8
10
12
RMSD [A]

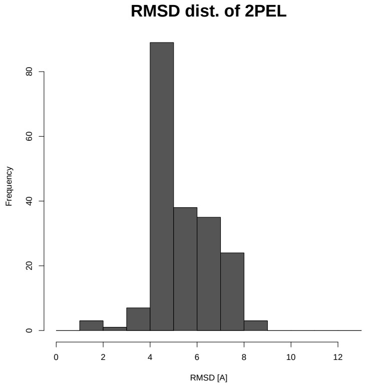

RMSD dist. of 1AX1
Frequency
0
20
40
60
80
100
120
140
0
2
4
6
8
10
12
RMSD [A]

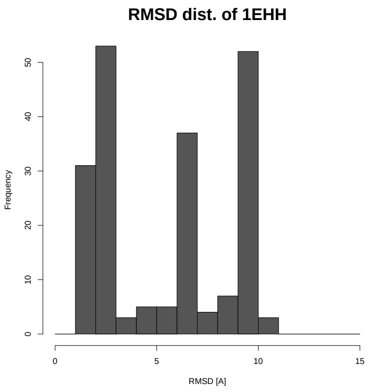

RMSD dist. of 1AX2
Frequency
0
20
40
60
80
0
2
4
6
8
10
12
RMSD [A]

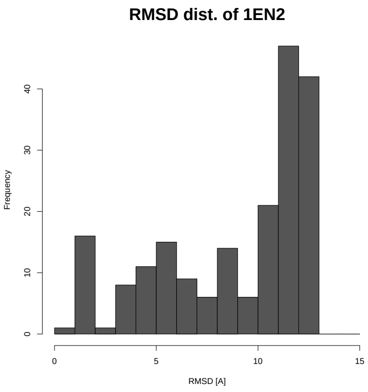

RMSD dist. of 2BQP
Frequency
0
20
40
60
80
0
2
4
6
8
10
12
RMSD [A]

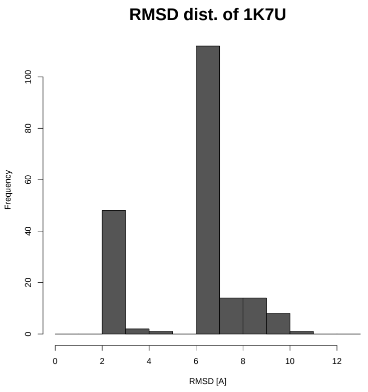

RMSD dist. of 1BQP
Frequency
0
10
20
30
40
50
60
0
2
4
6
8
10
12
RMSD [A]


RMSD dist. of 1QF3
Frequency
0
10
20
30
40
50
60
70
0
2
4
6
8
10
12
RMSD [A]


RMSD dist. of 2PEL
Frequency
0
20
40
60
80
0
2
4
6
8
10
12
RMSD [A]

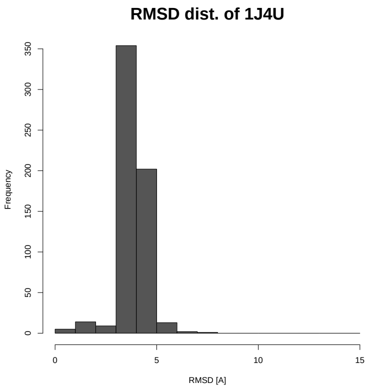

RMSD dist. of 1EHH

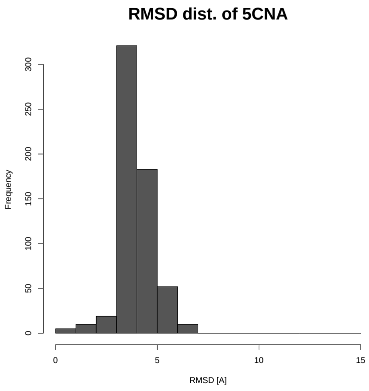

RMSD dist. of 1EN2

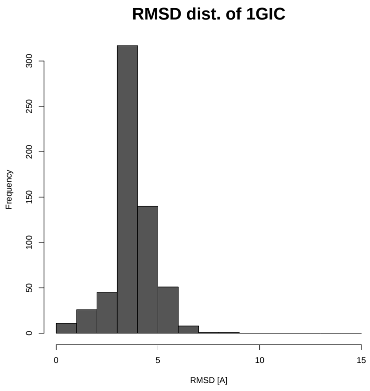

RMSD dist. of 1K7U

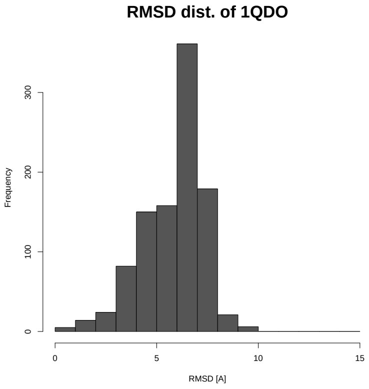
B.2. BALLDock/SLICK Results

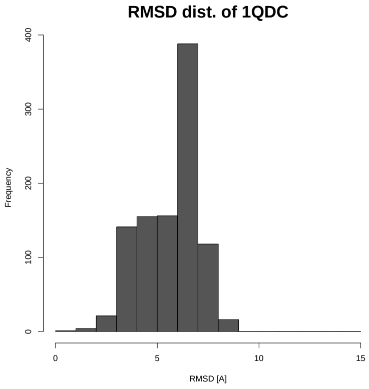
B.2.1. RMSD Distributions of the Calibration Set

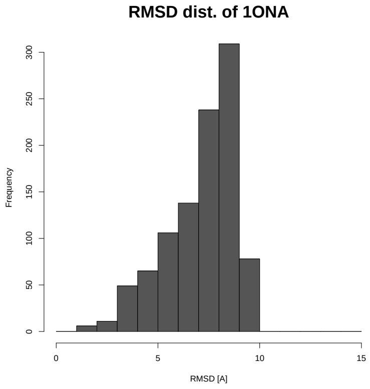


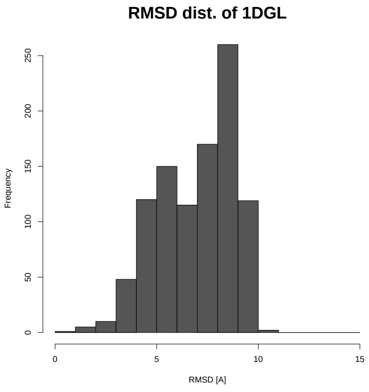


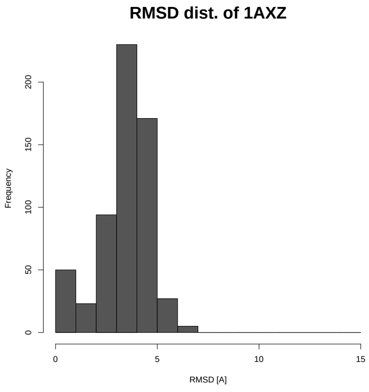


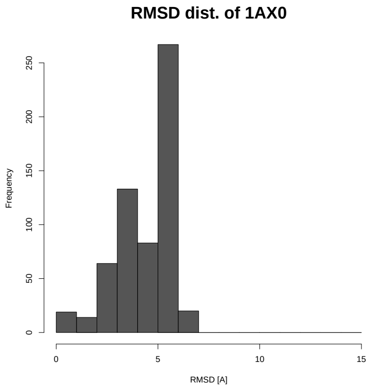




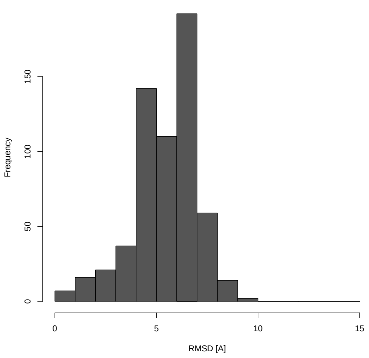




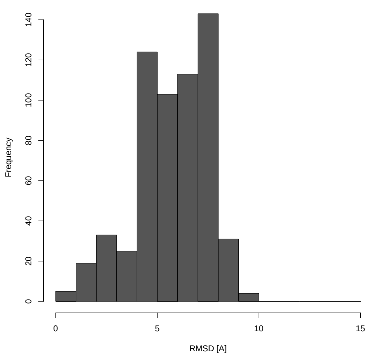




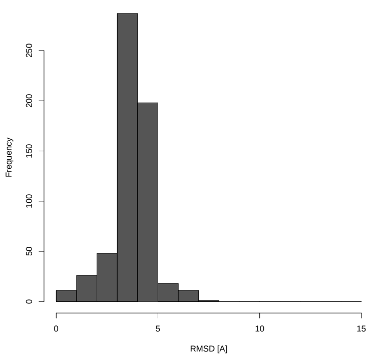
RMSD dist. of 1AX1



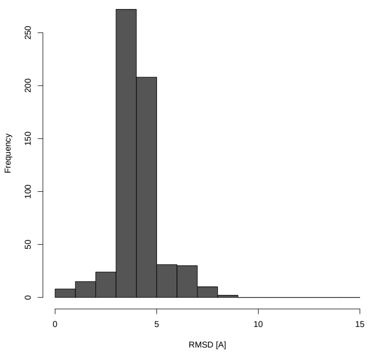
RMSD dist. of 1AX2



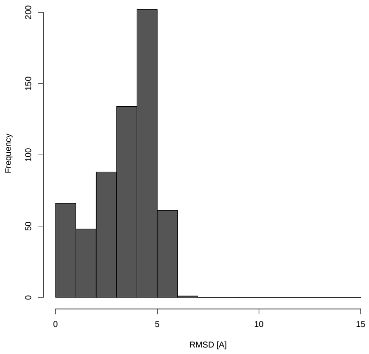
RMSD dist. of 2BQP



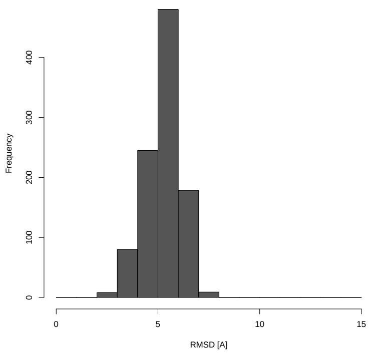
RMSD dist. of 1BQP



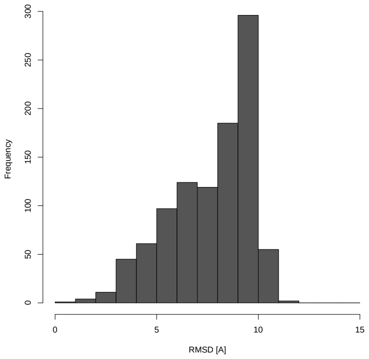
RMSD dist. of 1QF3



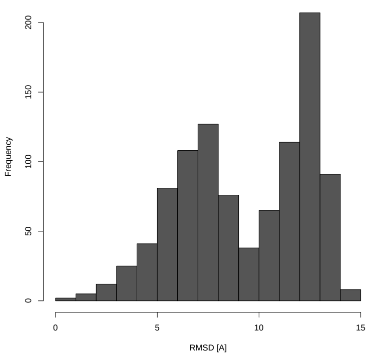
RMSD dist. of 2PEL



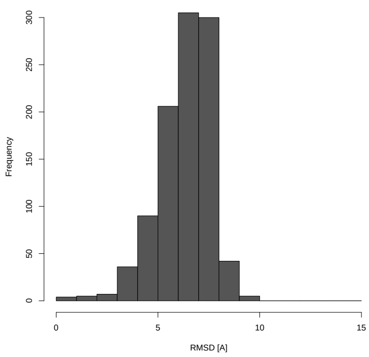
RMSD dist. of 1EHH



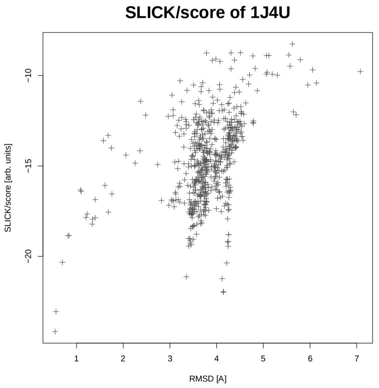
RMSD dist. of 1EN2

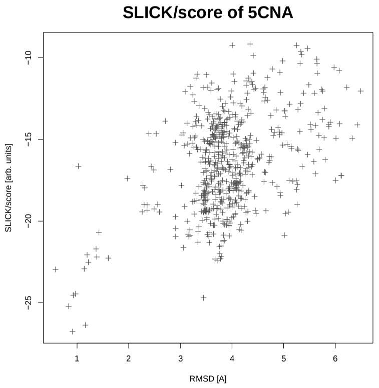


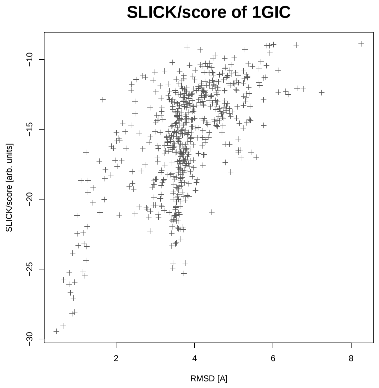
RMSD dist. of 1K7U

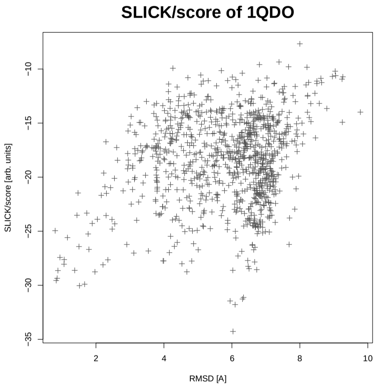


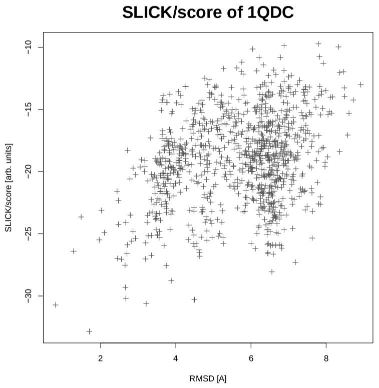
B.2.2. Calibration Set

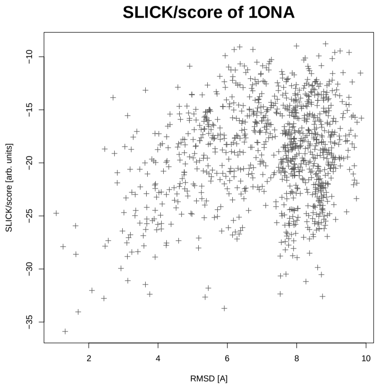


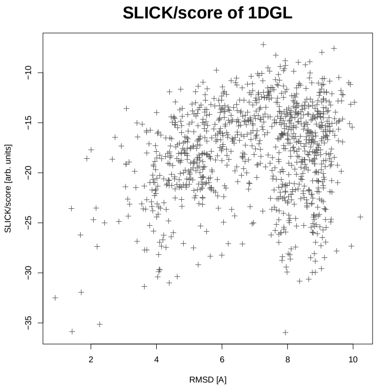


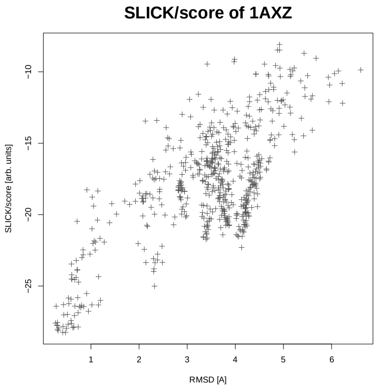


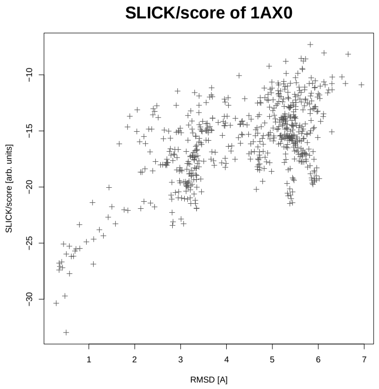


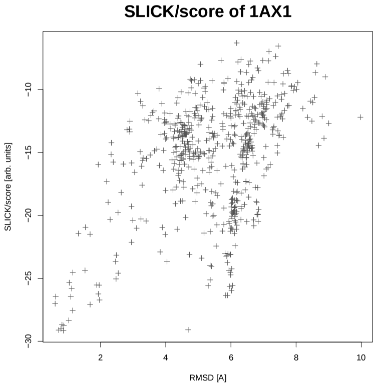


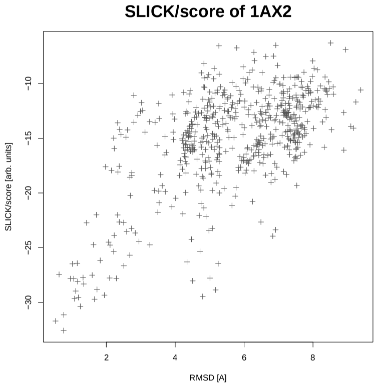


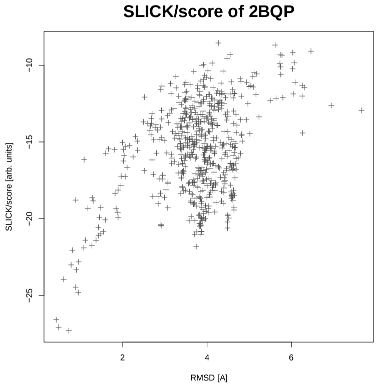


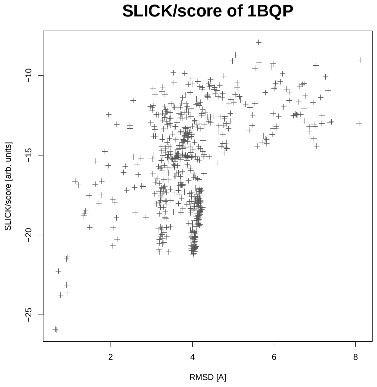


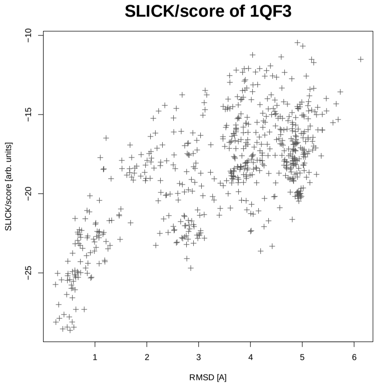


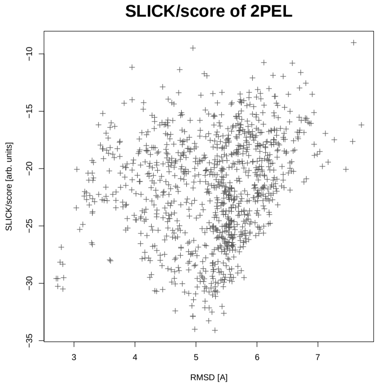

SLICK/score of 1AX1
SLICK/score [arb. units]
−10
−15
−20
−25
−30
RMSD [A]
2
4
6
8
10

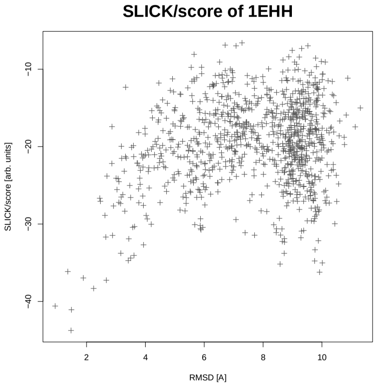

SLICK/score of 1AX2
SLICK/score [arb. units]
−10
−15
−20
−25
−30
RMSD [A]
2
4
6
8

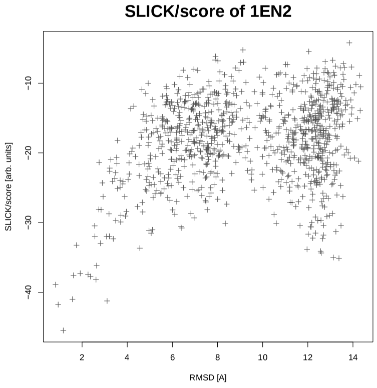

SLICK/score of 2BQP
SLICK/score [arb. units]
−10
−15
−20
−25
RMSD [A]
2
4
6

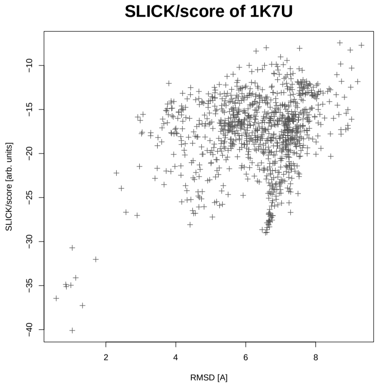

SLICK/score of 1BQP
SLICK/score [arb. units]
−10
−15
−20
−25
RMSD [A]
2
4
6
8


SLICK/score of 1QF3
SLICK/score [arb. units]
−10
−15
−20
−25
RMSD [A]
1
2
3
4
5
6

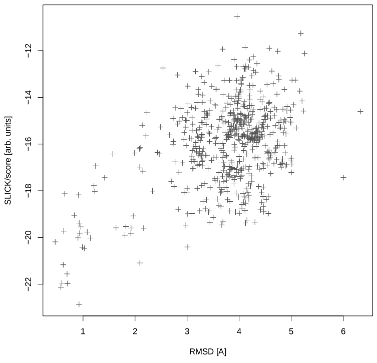

SLICK/score of 2PEL
SLICK/score [arb. units]
−10
−15
−20
−25
−30
−35
RMSD [A]
3
4
5
6
7

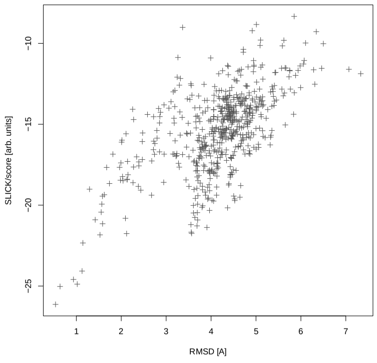

SLICK/score of 1EHH
SLICK/score [arb. units]
−10
−20
−30
−40
RMSD [A]
2
4
6
8
10

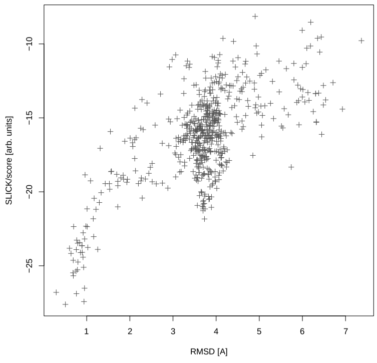

SLICK/score of 1EN2
SLICK/score [arb. units]
−10
−20
−30
−40
RMSD [A]
2
4
6
8
10
12
14

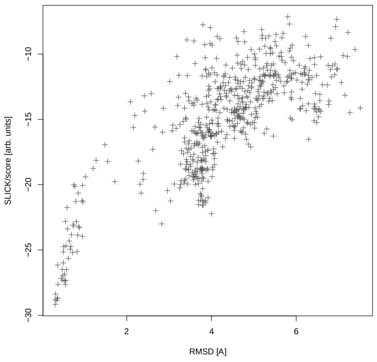

SLICK/score of 1K7U
SLICK/score [arb. units]
−10
−15
−20
−25
−30
−35
−40
RMSD [A]
2
4
6
8

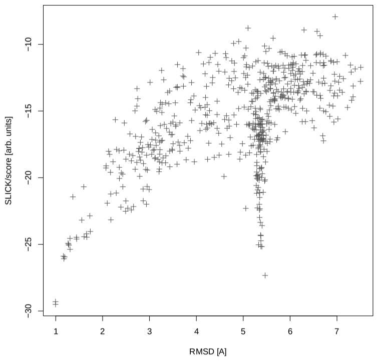
B.2.3. Docking Set – Plant Lectins

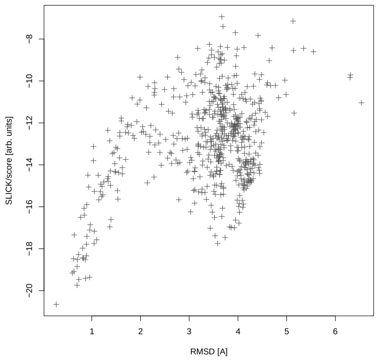

SLICK/score of 1KJ1

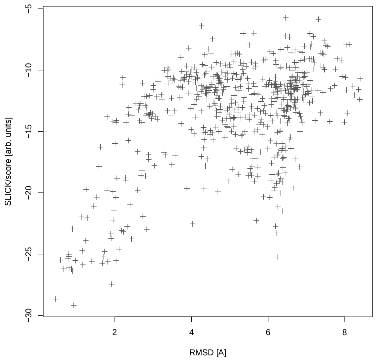

SLICK/score of 1KUJ

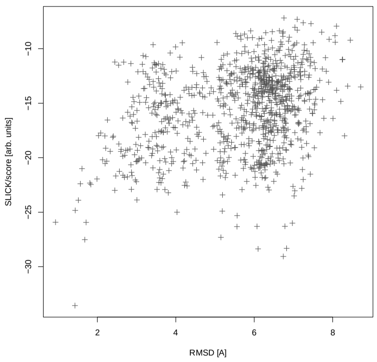

SLICK/score of 1MVQ

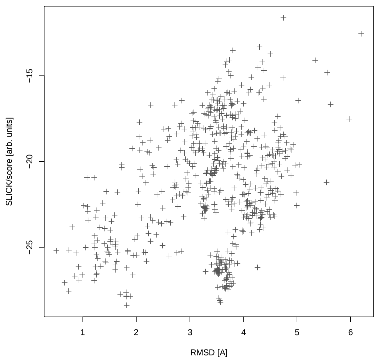

SLICK/score of 1WBL

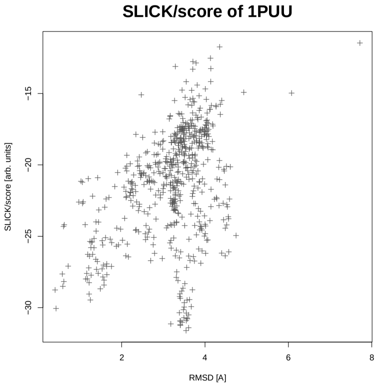

SLICK/score of 1FNZ

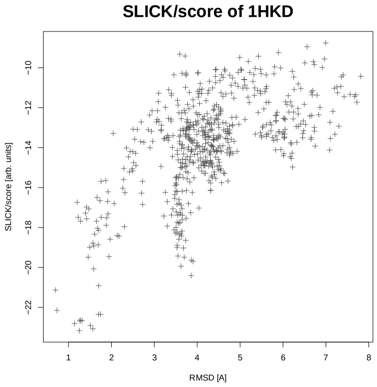

SLICK/score of 1C3M

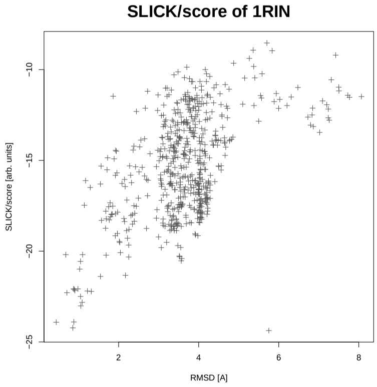

SLICK/score of 1GZC

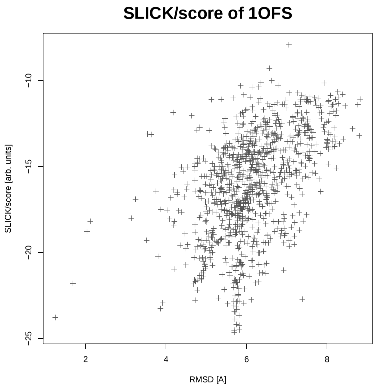

SLICK/score of 1JOT


SLICK/score of 1PUM

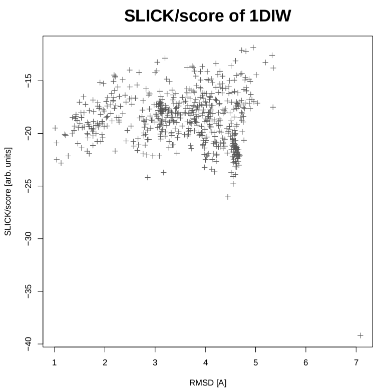

SLICK/score of 1PUU

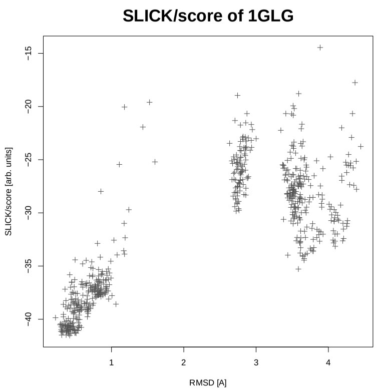

SLICK/score of 1HKD

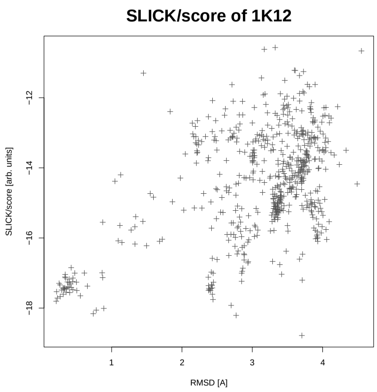

SLICK/score of 1RIN

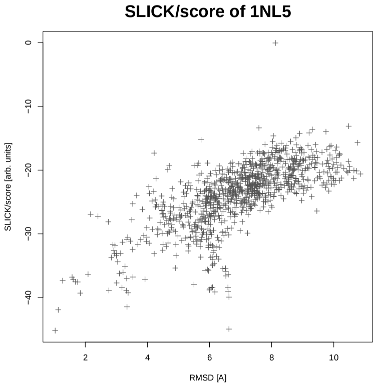

SLICK/score of 1OFS

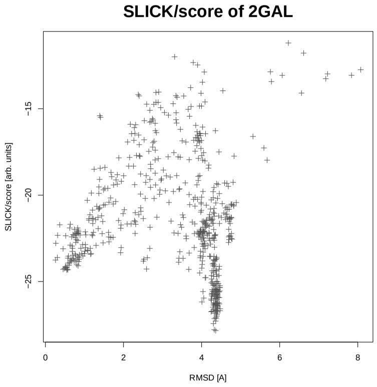
B.2.4. Docking Set – Non-plant Lectins

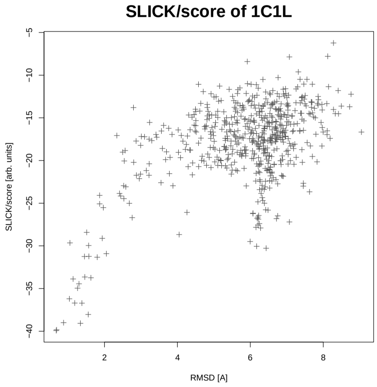

SLICK/score of 1DIW

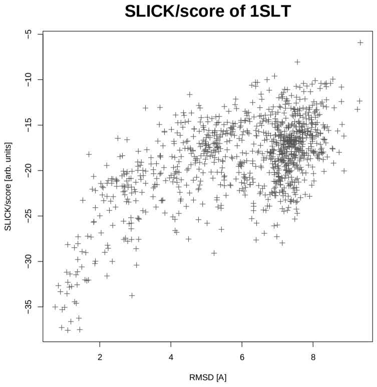

SLICK/score of 1GLG


SLICK/score of 1K12


SLICK/score of 1NL5
SLICK/score [arb. units]
0
−10
−20
−30
−40
RMSD [A]
2
4
6
8
10

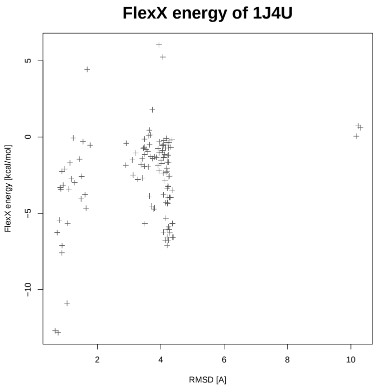

SLICK/score of 2GAL
SLICK/score [arb. units]
−15
−20
−25
RMSD [A]
0
2
4
6
8

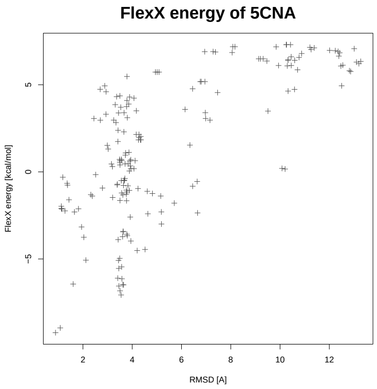

SLICK/score of 1C1L
SLICK/score [arb. units]
−5
−10
−15
−20
−25
−30
−35
−40
RMSD [A]
2
4
6
8

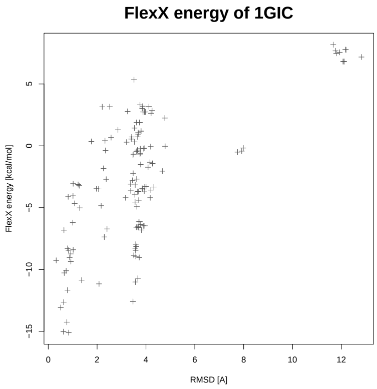

SLICK/score of 1SLT
SLICK/score [arb. units]
−5
−10
−15
−20
−25
−30
−35
RMSD [A]
2
4
6
8

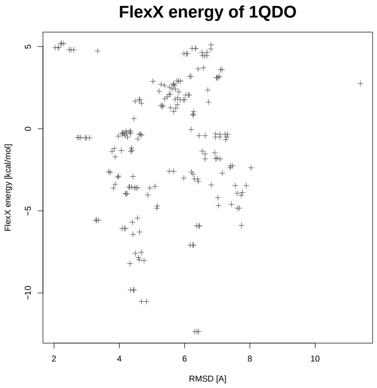
B.3. Results of FlexX Docking

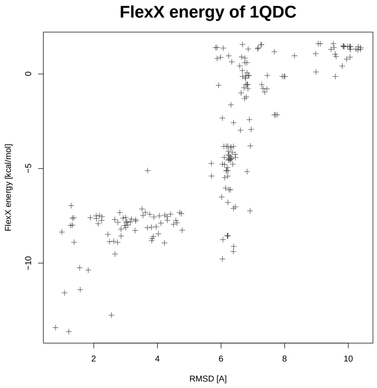
B.3.1. Calibration Set

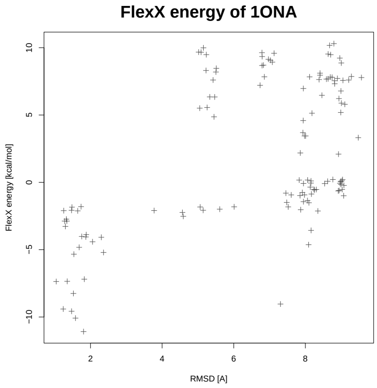


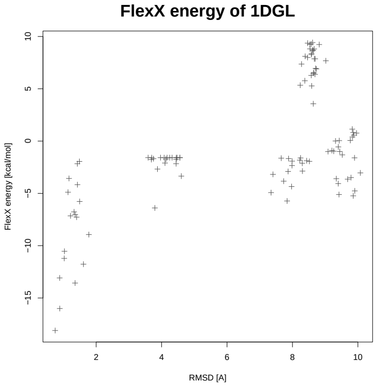


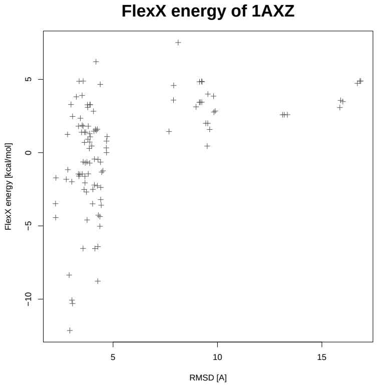


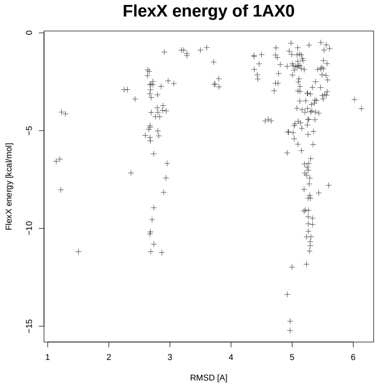


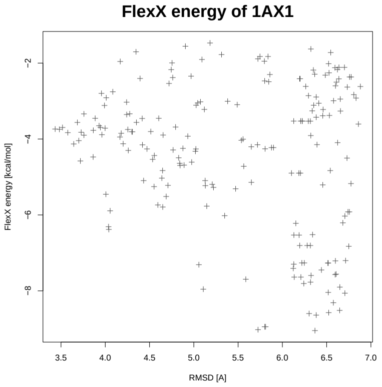


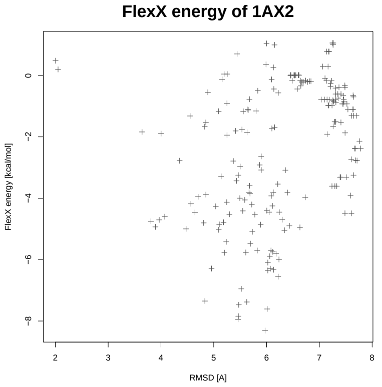


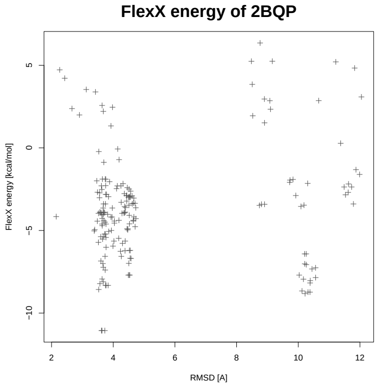


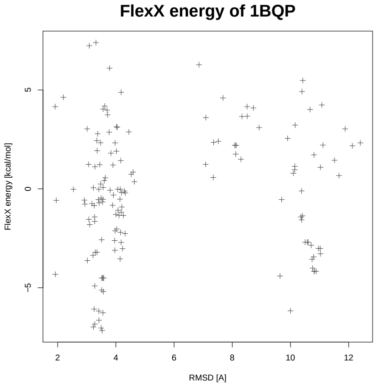


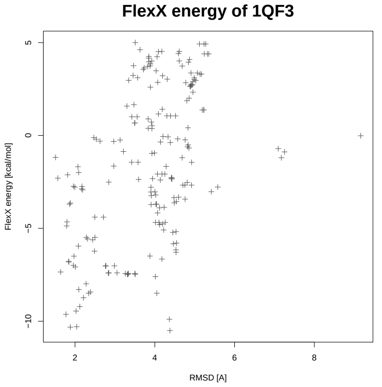


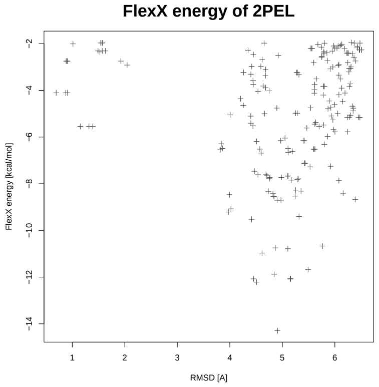

FlexX energy of 1AX1
FlexX energy [kcal/mol]
−2
−4
−6
−8
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
RMSD [A]

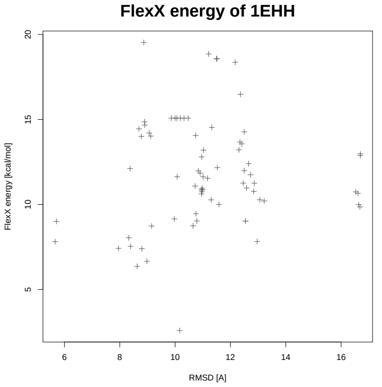

FlexX energy of 1AX2
FlexX energy [kcal/mol]
0
−2
−4
−6
−8
2
3
4
5
6
7
8
RMSD [A]

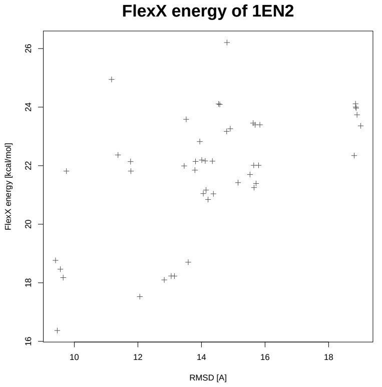

FlexX energy of 2BQP
FlexX energy [kcal/mol]
5
0
−5
−10
2
4
6
8
10
12
RMSD [A]

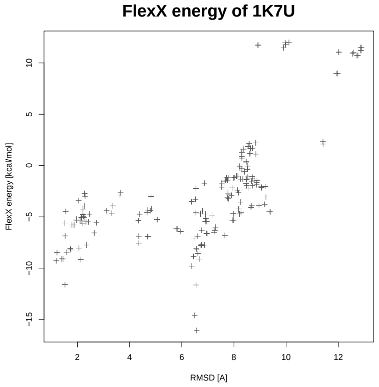

FlexX energy of 1BQP
FlexX energy [kcal/mol]
5
0
−5
2
4
6
8
10
12
RMSD [A]


FlexX energy of 1QF3
FlexX energy [kcal/mol]
5
0
−5
−10
2
4
6
8
RMSD [A]

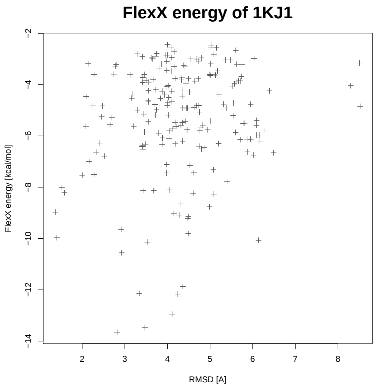

FlexX energy of 2PEL
FlexX energy [kcal/mol]
−2
−4
−6
−8
−10
−12
−14
1
2
3
4
5
6
RMSD [A]

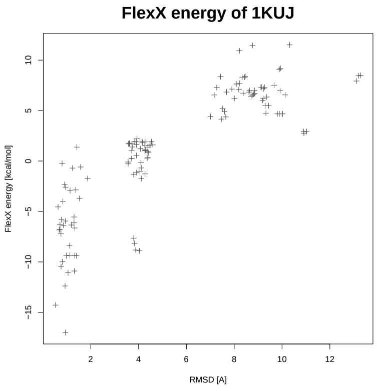

FlexX energy of 1EHH
FlexX energy [kcal/mol]
20
15
10
5
6
8
10
12
14
16
RMSD [A]

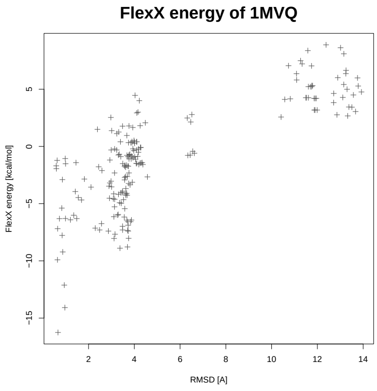

FlexX energy of 1EN2
FlexX energy [kcal/mol]
26
24
22
20
18
16
10
12
14
16
18
RMSD [A]

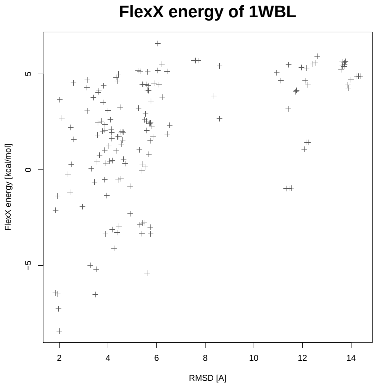

FlexX energy of 1K7U
FlexX energy [kcal/mol]
10
5
0
−5
−10
−15
2
4
6
8
10
12
RMSD [A]

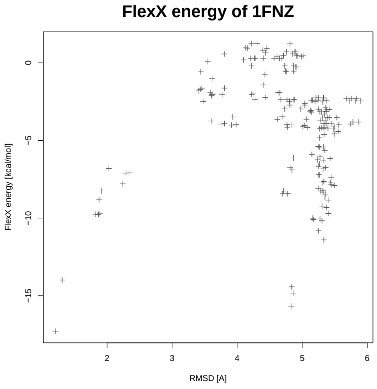
B.3.2. Docking Set – Plant Lectins

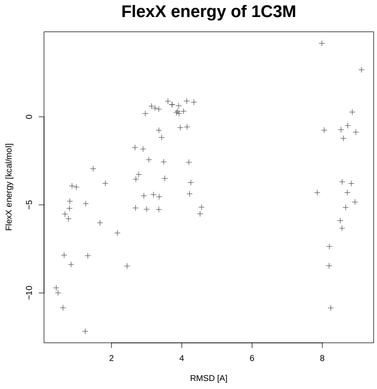


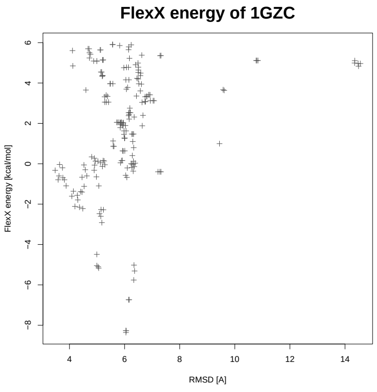


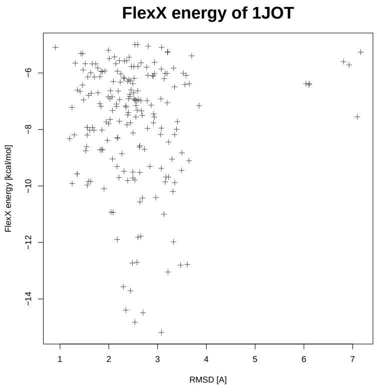


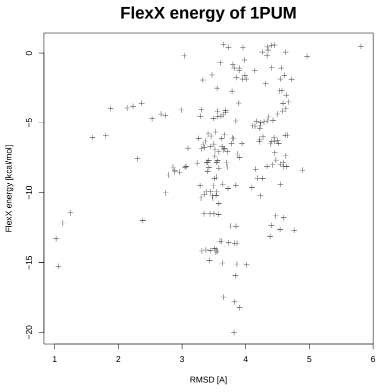


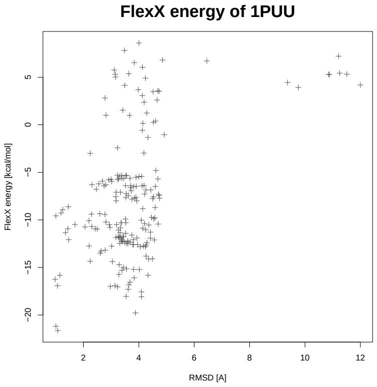


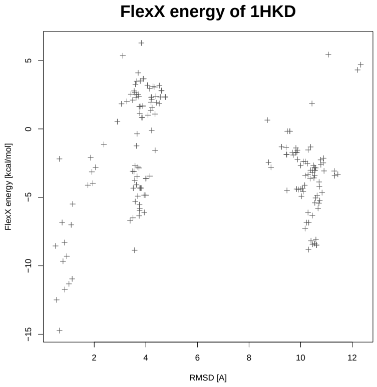


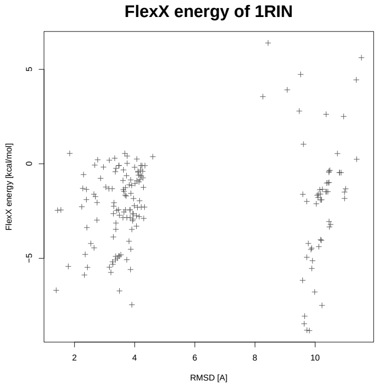


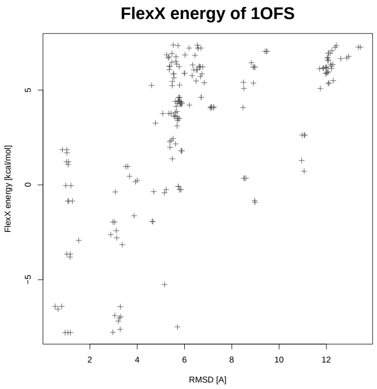




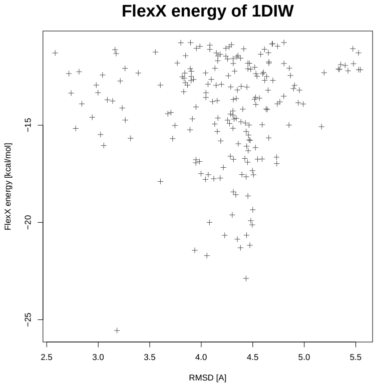


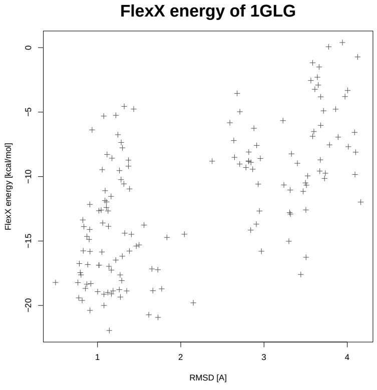


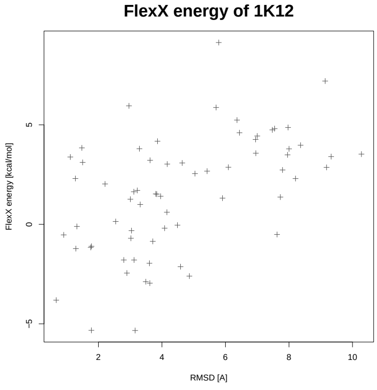


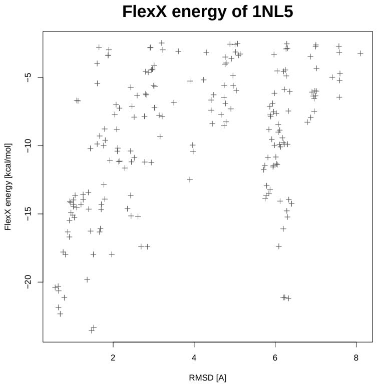


B.3.3. Docking Set – Non-plant Lectins

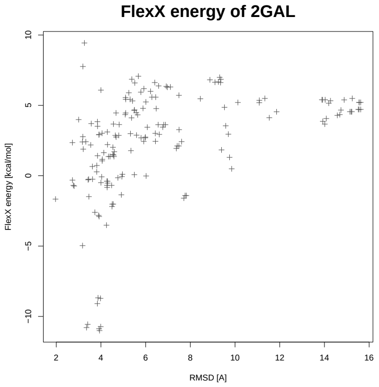




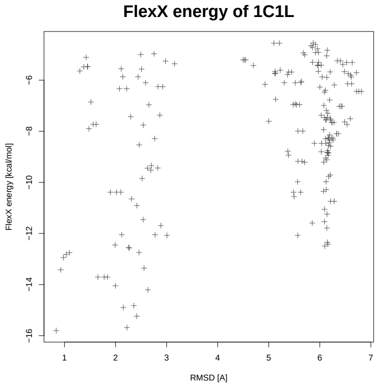




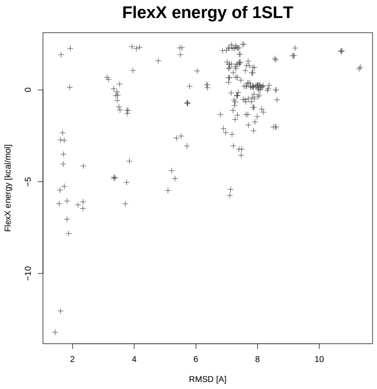
FlexX energy of 1NL5



FlexX energy of 2GAL



FlexX energy of 1C1L



FlexX energy of 1SLT

C. Mathematical Details

C.1. Deriving Fermi Parameters from Switching Function Limits

The aim is to derive parameters a', b' that ensure that the slope of linear base function and sigmoid function are equal at the centre x of the interval defined by a and b.

$$x = a + \frac{1}{2}(b - a) \tag{C.1}$$

Exploiting the fact that both functions have the value $\frac{1}{2}$ at point x gives the first constraint:

$$\frac{1}{1 + \exp(-a'x + b')} = \frac{1}{2} \quad \Rightarrow \quad b' = a'x = a'(a + \frac{1}{2}(b - a)) \tag{C.2}$$

Now the derivatives of both functions have to be computed.

$$\frac{d}{dx}\left(\frac{1}{1 + \exp(-a'x + b')}\right) = \frac{a' \exp(-a'x + b')}{(1 + \exp(-a'x + b'))^2} \tag{C.3}$$

$$\frac{d}{dx}\left(1 - \frac{x - a}{b - a}\right) = \frac{1}{a - b} \tag{C.4}$$

Inserting (C.1) and (C.2) into (C.3) yields

$$\frac{a' \exp(-a'x + a'x)}{(1 + \exp(-a'x + a'x))^2} = \frac{a'}{(1 + 1)^2} = \frac{a'}{4} \tag{C.5}$$

Equating (C.4) and (C.5) yields

$$\frac{a'}{4} = \frac{1}{a - b} \quad \Rightarrow \quad a' = \frac{4}{a - b} \tag{C.6}$$

C.2. Recalculation of Solvent-Solute Van der Waals Interactions

The solvent volume V_S will be defined as that portion of space that is not enclosed by the solvent accessible surface (SAS) of the solute molecule (see Fig. C.1). Since the SAS of the solute molecule (henceforth denoted S_M) is easily calculated, it is convenient to transform the volume integral into a surface integral. Huron and Claverie [71] employed Ostrogradsky's formula [154] for this

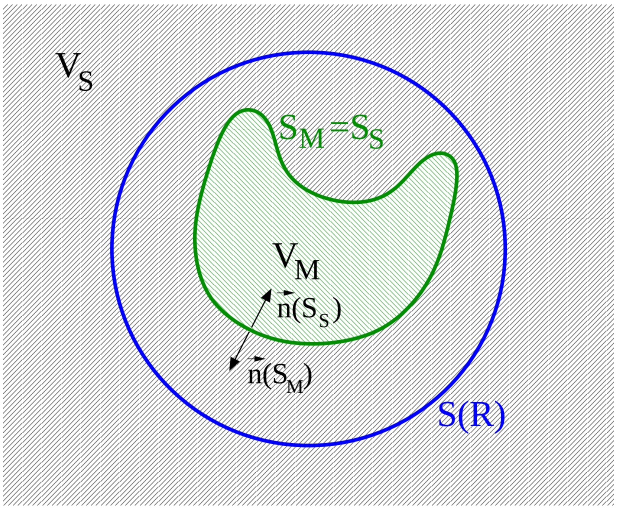


Figure C.1.: The definition of the volumes and surfaces used. The molecular surface S_M separates the solute volume V_M from the solvent volume V_S. It is identical to the solvent surface S_S, except for the surface normals which have opposing signs.

purpose. They introduce a sphere $S(R)$ of radius R that completely encloses S_M. The solvent volume $V_S(R)$ included between $S(R)$ and S_S is then

$$\iiint\limits_{V_S(R)} \operatorname{div} \mathbf{W}(\mathbf{r})\, dv = \iint\limits_{S_S+S(R)} \mathbf{W}(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r})\, ds \tag{C.7}$$

where $\mathbf{n}(\mathbf{r})$ are the unit normals to $S_S \cup S(R)$ directed towards the outside of the integration volume and $\mathbf{W}$ is a vector field. Although the solvent surface S_S and the molecular surface S_M are in principle identical their normals have opposing signs. Since the volume integral is over the solvent volume, we have to use S_S instead of S_M.

If the volume and the surface integrals converge for $R \to \infty$, as is the case for any reasonable choice of the Van der Waals interactions, the integral over the total solvent volume V_S can be written as

$$\iiint\limits_{V_S} \operatorname{div} \mathbf{W}\, dv = \iint\limits_{S_S} \mathbf{W} \cdot \mathbf{n}\, ds + \lim_{R\to\infty} \iint\limits_{S(R)} \mathbf{W} \cdot \mathbf{n}\, ds \tag{C.8}$$

A radial symmetric vector field $\mathbf{W}(r)$ has to be determined such that its divergence gives the function $F(r)$ that has to be integrated.

$$\begin{aligned} F(r) &= \operatorname{div} \mathbf{W} && \text{(C.9)}\\ \Rightarrow \mathbf{W}(\mathbf{r}) &= f(r)\mathbf{r} && \text{(C.10)}\\ \text{with} \quad F(r) &= \nabla \cdot (f(r) \cdot \mathbf{r}) && \text{(C.11)}\\ F(r) &= r\frac{d}{dr} f(r) + 3f(r) && \text{(C.12)} \end{aligned}$$

Solutions to to the differential equation (C.12) are of the form

$$f(r) \;=\; \frac{1}{r^3}\int x^2 F(x)\,dx + C \tag{C.13}$$

Together with (C.9) and (C.11), a solution for the vector field $\mathbf{W}$ is

$$\mathbf{W}(\mathbf{r}) = \left(\frac{1}{r^3}\int\limits_{r_0}^{r} x^2 F(x) dx\right)\cdot \mathbf{r} \tag{C.14}$$

The choice of the lower integration limit r_0 (which corresponds to the integration constant C in (C.13)) in this equation changes the values of the two integrals in (C.8). It seems convenient to choose r_0 such that one of these integrals vanishes. In fact, this is possible for the second integral

$$\lim_{r\to\infty}\iint\limits_{S(R)} \mathbf{W}\cdot\mathbf{n}\,ds \;=\; \lim_{R\to\infty}\iint\limits_{S(R)}\left(\frac{1}{R^3}\int\limits_{r_0}^{R} x^2 F(x)\,dx\right)\cdot\mathbf{R}\cdot\mathbf{n}\,ds \tag{C.15}$$

Since the surface integral is over the sphere $S(R)$, the surface normals $\mathbf{n}$ have the same direction as $\mathbf{R}$:

$$\mathbf{R}\cdot\mathbf{n} \;=\; \mathbf{R}\cdot\frac{\mathbf{R}}{R} = R \tag{C.16}$$

$$\Rightarrow \quad \lim_{R\to\infty} \quad \iint\limits_{S(R)}\left(\frac{1}{R^3}\int\limits_{r_0}^{R} x^2 F(x)\,dx\right)\cdot\mathbf{R}\cdot\frac{\mathbf{R}}{R}\,ds \tag{C.17}$$

$$= \lim_{R\to\infty}\frac{1}{R^3}\int\limits_{r_0}^{R} x^2 F(x) R \iint\limits_{S(R)} ds \tag{C.18}$$

$$= \lim_{R\to\infty}\frac{1}{R^3}\int\limits_{r_0}^{R} x^2 F(x)\,dx\cdot 4\pi R^3 \tag{C.19}$$

$$= 4\pi\int\limits_{r_0}^{\infty} x^2 F(x)\,dx \tag{C.20}$$

To make this integral vanish, $r_0 = \infty$ has to be chosen. Hence, the initial volume integral (C.8) becomes

$$\iiint\limits_{V_S} F(r)\,dv \;=\; \iiint\limits_{V_S} \operatorname{div}\mathbf{W}\,dv = \iint\limits_{S_M}\mathbf{W}\cdot\mathbf{n}\,ds + 0 \tag{C.21}$$

$$= \iint\limits_{S_M}\left(\frac{1}{r^3}\int\limits_{\infty}^{r} x^2 F(x)\,dx\right)\mathbf{r}\cdot\mathbf{n}\,ds \tag{C.22}$$

This result can now be applied to the energy functions of interest, *e. g.* the 6-12 potential used to describe the Van der Waals interaction energy $E(r)$ between two particle with distance r.

$$E = \frac{A}{r^{12}} - \frac{B}{r^{6}} \tag{C.23}$$

To determine the interaction energy of a single atom of the solute molecule m with a single atom s of all solvent molecules surrounding the solute, the whole solute volume, *i. e.* the universe outside the molecular surface of the solute, has to be integrated. Assuming an isotropic solvent with number density ρ_S, the Van der Waals interaction energy E^{vdw} of the two atoms m and s is

$$\begin{aligned}
\iiint\limits_{V_S} \rho_S E^{\mathrm{vdw}}(m,s)\,dv &= \rho_S \iiint\limits_{V_S} \frac{A}{r^{12}} - \frac{B}{r^{6}}\,dv && \text{(C.24)}\\
&\overset{\text{(C.22)}}{=} \rho_S \iint\limits_{S_S} \left(\frac{1}{r^3}\int\limits_{\infty}^{r} x^2\left(\frac{A}{x^{12}} - \frac{B}{x^{6}}\right)dx\right)\mathbf{r}\cdot\mathbf{n}(S_S), ds && \text{(C.25)}\\
&= \rho_S \iint\limits_{S_S} \left(\frac{1}{r^3}\int\limits_{\infty}^{r} \left(\frac{A}{x^{10}} - \frac{B}{x^{4}}\right)dx\right)\mathbf{r}\cdot\mathbf{n}(S_S)\,ds && \text{(C.26)}\\
&= \rho_S \iint\limits_{S_S} \frac{1}{r^3}\left(-\frac{A}{9r^9} + \frac{B}{3r^3}\right)\cdot\mathbf{r}\cdot\mathbf{n}(S_S)\,ds && \text{(C.27)}
\end{aligned}$$

By using the normals to the molecular surface $\mathbf{n}(S_M) = -\mathbf{n}(S_S)$ instead of the normals to the solvent surface, the integral can be written as

$$\begin{aligned}
\iiint\limits_{V_S} \rho_S E^{\mathrm{vdw}}(m,s)\,dv &= \rho_S \iint\limits_{S_M} \frac{1}{r^3}\left(\frac{A}{9r^9} - \frac{B}{3r^3}\right)\cdot\mathbf{r}\cdot\mathbf{n}(S_M)\,ds && \text{(C.28)}\\
&= \rho_S \iint\limits_{S_M} \left(\frac{A}{9r^{12}} - \frac{B}{3r^6}\right)\cdot\mathbf{r}\cdot\mathbf{n}(S_M)\,ds && \text{(C.29)}
\end{aligned}$$

The complete interaction energy of the solute and the solvent $E^{\mathrm{vdw}}_{M,S}$, can be obtained by summing over all atoms of the solvent and the solute.

$$E^{\mathrm{vdw}}_{M,S} = \sum_{m\in M}\sum_{s\in S} \rho_S \iint\limits_{S_M} \left(\frac{A_{m,s}}{9r^{12}} - \frac{B_{m,s}}{3r^6}\right)\cdot\mathbf{r}\cdot\mathbf{n}(S_M)\,ds \tag{C.30}$$

With this expression, it is now possible to include averaged structural information into the calculation of the solvent-solute interaction energy. In a manner of speaking, this expression describes the interaction of a molecule with a “structured continuum” surrounding it. Although the complexity of this approach is high from the mathematical point of view, the data necessary for the calculations and parts of the distribution function evaluation can be pre-computed. Hence, the slow-down to be expected from this formulation is relatively small.

D. Data Set Reference and Abbreviations

D.1. Complexes Sorted by PDB ID

PDB ID	Lectin	Ligand	n-mer	p/n
1AX0	ECorL	GalNAc	1	p
1AX1	ECorL	Lac	2	p
1AX2	ECorL	LacNAc	2	p
1AXZ	ECorL	Gal	1	p
1BQP	PSL	D-Man	1	p
1C1L	Congerin I	Lac	2	n
1C3M	Heltuba	Man(1-3)Man	2	p
1DGL	DGL	Me-3,6-di-O-(α-D-Man)-α-D-Man	3	p
1DIW	Tetanus toxin	Gal	1	n
1EHH	UDA	$(\text{GlcNAc})_3$	3	p
1EN2	UDA	$(\text{GlcNAc})_4$	4	p
1FNZ	RPbA	GlcNAc	1	p
1GIC	ConA	Me-α-D-Glc	1	p
1GLG	Chemotactic receptor	Gal	1	n
1GZC	ECL	Lac	2	p
1HKD	PSL	Me-α-D-glucopyranoside	1	p
1J4U	AIA	Me-α-D-Man	1	p
1JOT	MPA	GalNAc-Gal	2	p
1K12	AAnA	Fuc	1	n
1K7U	WGA	$(\text{GlcNAc})_2$	2	p
1KJ1	ASA	α-D-Man	1	p
1KUJ	AIA	Me-Man	1	p
1MVQ	*Cratylia mollis* lectin	Me-Man	1	p
1NL5	Eng. maltose bind. lectin	Mal	1	n
1OFS	PSL	Sucrose	2	p
1ONA	ConA	Me-3,6-di-O-(α-D-Man)-α-D-Man	3	p
1PUM	ML	Gal	1	p
1PUU	ML	Lac	2	p
1QDC	ConA	Me-6-O-(α-D-Man)-α-D-Man	2	p
1QDO	ConA	Me-3-O-(α-D-Man)-α-D-Man	2	p
1QF3	PNA	Me-β-D-Gal	1	p

Table D.1 – Continued on next page

PDB ID	Lectin	Ligand	n-mer	p/n
1RIN	PSL	Man_3	3	p
1SLT	S-lectin	LacNAc	2	n
1WBL	PTL	Me-Gal	1	p
2BQP	PSL	D-Glc	1	p
2GAL	hGal-7	Gal	1	n
2PEL	PNA	Lac	2	p
4GAL	hGal-7	Lac	2	n
5CNA	ConA	Me-α-D-Man	1	p
5GAL	hGal-7	LacNAc II	2	n

D.2. Lectin Abbreviations

Abbreviation	**Lectin**
AAnA	*Anguilla anguilla* lectin
AIA	*Artocarpus integrifolia* agglutinin
ASA	*Allium sativum* agglutinin
ConA	Concanavalin A (*Canavalia ensiformis* lectin)
DGL	*Dioclea grandiflora* lectin
ECL	*Erythrina crista-galli* lectin
ECorL	*Erythrina corallodendron* lectin
Heltuba	*Helianthus tuberosus* agglutinin
hGal-7	Human galectin-7
ML	*Viscum album* lectin
MPA	*Maclura pomifera* agglutinin
PNA	Peanut agglutinin
PSL	*Pisum sativum* lectin
PTL	*Psophocarpus tetragonolobus* lectin
UDA	*Urtica dioica* agglutinin
WGA	Wheat germ agglutinin

Table D.2.: Lectin abbreviations

D.3. Carbohydrate Abbreviations

Abbreviation	**Carbohydrate**
Fuc	Fucose
Gal	Galactose
GalNAc	N-Acetyl-Galactosamine
Glc	Glucose
GlcNAc	N-Acetyl-Glucosamine
Lac	Lactose
LacNAc	N-Acetyl-Lactosamine
Man	Mannose
Suc	Sucrose

Table D.3.: Carbohydrate abbreviations

E. Short Curriculum Vitae

Name	Andreas Kerzmann
Date of Birth	May 13, 1974
Birthplace	Saarbrücken
Citizenship	German

Education

08/2003 – 08/2006	Ph. D. student in the Department for Simulation of Biological Systems (Prof. Kohlbacher), Eberhard-Karls-Universität Tübingen Thesis: *Protein-Carbohydrate Docking*
03/2000 – 08/2003	Ph. D. student in the Bioinformatics department (Prof. Lenhof), Universität des Saarlandes, Saarbrücken (supervisor: Oliver Kohlbacher)
03/2000	German university degree in Computer Science (Dipl.-Inform.) Thesis: *Zwangsbasierte Dynamiksimulation im R^2 unter Verwendung von Kreiskanten*
09/1994 – 03/2000	Student of Computer Science, minor: Physics, Universität des Saarlandes, Saarbrücken
07/1993	German A-level equivalent (Abitur)
1984 – 1993	Staatliches Gymnasium am Rotenbühl, Saarbrücken

Working Experience

03/2000 – 08/2006	Developer in the BALL project, a C++ molecular modelling framework Realisation of diverse computational interaction models
08/2003 – 08/2006	Teaching assistant Lectures on protein structure, drug design and molecular modelling; practical training, supervision of several student research projects
08/2003 – 08/2006	Computer administration in the Kohlbacher department, Tübingen File and backup services, cluster integration, software, support
03/2000 – 08/2003	Computer administration in the Lenhof department, Saarbrücken Design, implementation and administration of the computer infrastructure, cluster integration, backup facilities, software, support
03/2000 – 03/2002	Researcher in the GELENA project BMBF funded project on lectin-functionalised non-viral gene transfer systems
01/1997 – 03/2001	Student assistant in the computer support group (Rechnerbetriebsgruppe) of Max-Planck-Institut für Informatik, Saarbrücken
08/1993 – 09/1994	Civilian alternative service (Zivildienst)

Publications

Journals

A. Kerzmann, J. Fuhrmann, O. Kohlbacher, and D. Neumann
Protein-Carbohydrate Docking with SLICK
in preparation

A. Kerzmann, D. Neumann, and O. Kohlbacher
SLICK – Scoring and Energy Functions for Protein-Carbohydrate Interactions
J. Chem. Inf. Model, **46**:1635–1642, 2006, DOI: 10.1021/ci050422y

Conferences

A. Kerzmann, D. Neumann, and O. Kohlbacher
High-Accuracy Prediction of Protein-Carbohydrate Interactions
Trends in Glycoscience and Glycotechnology, **16** (Supplement):S26, 2004

D. Neumann, A. Kerzmann and O. Kohlbacher
Modelling the sugar-lectin interaction by computer simulated docking
Interlec 20, p. 116, 2002

G. Hotz, A. Kerzmann, C. Lennerz, R. Schmid, E. Schömer and T. Warken
Calculation of Contact Forces
ACM Symposium on Virtual Reality Software and Technology, VRST'99, p. 180-181, 1999

G. Hotz, A. Kerzmann, C. Lennerz, R. Schmid, E. Schömer and T. Warken
SiLVIA - a simulation library for virtual reality applications
Proceedings of IEEE Virtual Reality, p. 82, 1999

Awards

do-it Software Award 2005 (5th prize)
"BALLView – Ein Open-Source-Werkzeug zur Visualisierung und Modellierung von Biomolekülen"

bwcon Open Source Sonderpreis 2005
"BALLView – Ein Open-Source-Werkzeug zur Visualisierung und Modellierung von Biomolekülen"

Bibliography

[1] Pedro M. Coutinho, Michael K. Dowd, and Peter J. Reilly. Automated docking of monosaccharide substrates and analogues and methyl α-acarviosinide in the glucoamylase active site. *PROTEINS*, 27:235–248, 1997.

[2] Pedro M. Coutinho, Michael K. Dowd, and Peter J. Reilly. Automated docking of glucosyl disaccharides in the glucoamylase active site. *PROTEINS*, 28:162–173, 1997.

[3] Pedro M. Coutinho, Michael K. Dowd, and Peter J. Reilly. Automated docking of α-(1,4)- and α-(1,6)-linked glucosyl trisaccharides in the glucoamylase active site. *Ind. Eng. Chem. Res.*, 37:2148–2157, 1998.

[4] Alain Laederach, Michael K. Dowd, Pedro M. Coutinho, and Peter J. Reilly. Automated docking of maltose, 2-deoxymaltose, and maltotetraose into the soybean b-amylase active site. *PROTEINS*, 37:166–175, 1999.

[5] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662, 1998.

[6] Alain Laederach and Peter J. Reilly. Specific empirical free energy function for automated docking of carbohydrates to proteins. *J. Comput. Chem.*, 24:1748–1757, 2003.

[7] Jan Fuhrmann. Paralleles Docking zyklischer Peptide an ausgewählte Proteine mit Hilfe eines genetischen Docking-Algorithmus. Diplomarbeit, Universität des Saarlandes, 2005.

[8] Daniel B. Werz and Peter H. Seeberger. Total synthesis of antigen bacillus anthracis tetrasaccharide - creation of an anthrax vaccine candidate. *Ang. Chem. Int. Ed.*, 44:6315–6318, 2005.

[9] Peter H. Seeberger. Exploring life's sweet spot. *Nature*, 437:1239, 2005.

[10] Feng Hong, Jun Yan, Jarek T. Baran, Daniel J. Allendo rf, Richard D. Hansen, Gary R. Ostroff, Pei Xiang Xing, Nai-Kong V. Cheung, and Gordon D. Ross. Mechanism by which orally administered β-1,3-glucans enhance the tumoricidal activity of antitumor monoclonal antibodies in murine tumor models. *J. Immunol.*, 173:797–806, 2004.

[11] Andreas von Bubnoff. Sugar coating improves anticancer treatment. *news@nature.com*, pages doi:10.1038/news050418–6, 2005.

[12] Mark Staples. Carbohydrates are ubiquitous and perform critical functions in biological systems. *AAPS Newsmagazine*, June:18–21, 2003.

[13] Hans-Joachim Gabius. The how and why of protein-carbohydrate interaction: A primer to the theoretical concept and a guide to application in drug design. *Pharm. Res.*, 15:23–30, 1998.

[14] D. Solis, J. Jimenez-Barbero, H. Kaltner, A. Romero, H.-C. Siebert, C.-W. von der Lieth, and H.-J. Gabius. Towards defining the role of glycans as hardware in information storage and transfer: Basic principles, experimental approaches and recent progress. *Cells Tissues Organs*, 168:5–23, 2001.

[15] Susumu Ito. The enteric surface coat on cat intestinal microvilli. *J. Cell Biol.*, 27:475, 1965.

[16] Andreas Bohne, Elke Lang, and Claus-Wilhelm von der Lieth. W3-SWEET: Carbohydrate modeling by internet. *J. Mol. Model.*, 4:33–43, 1998.

[17] Hans-Christian Siebert, Sabine Andre, Juan Luis Asensio, Francisco Javier Canada, Xin Dong, Juan Felix Espinosa, Martin Frank, Martine Gilleron, Herbert Kaltner, Tibor Kozar, Nicolai V. Bovin, Claus-Wilhelm von der Lieth, Johannes F. G. Vliegenhart, Jesus Jimenez-Barbero, and Hans-Joachim Gabius. A new combined computational and NMR-spectroscopical strategy for the identification of additional conformational constraints of the bound ligand in an aprotic solvent. *CHEMBIOCHEM*, 1:181–195, 2000.

[18] R. A. Laine. *Glycosciences: Status and Perspectives*, pages 1–14. Chapman & Hall, 1997.

[19] Qiang Liu and J. W. Brady. Anisotropic solvent structuring in aqueous sugar solutions. *J. Am. Chem. Soc.*, 118:12276–12286, 1996.

[20] Ahmed Touhami, Barbara Hoffmann, Andrea Vaselle, Frederic A. Denis, and Yves F. Dfrene. Aggregation of yeast cells: direct measurement of discrete lectin-carbohydrate interactions. *Microbiology*, 149:2873–2878, 2003.

[21] Alessandra Cambi, Marjolein Koopman, and Carl G. Figdor. How C-type lectins detect pathogens. *Cellular Microbiology*, 7:481–488, 2005.

[22] S. Y. Dai, R. Nakagawa, A. Itoh, H. Murakami, Y. Kashio, H. Abe, S. Katoh, K. Kontani, M. Kihara, S. L. Zhang, T. Hata, T. Nakamura, A. Yamauchi, and M. Hirashima. Galectin-9 induces maturation of human monocyte-derived dendritic cells. *J. Immunol.*, 175:2974–2981, 2005.

[23] Shugo Ueda, Ichiro Kuwabara, and Fu-Tong Liu. Suppression of tumor growth by galectin-7 gene transfer. *Cancer Res.*, 64:5672, 2004.

[24] Andre Danguy, Isabelle Camby, and Robert Kriss. Galectins and cancer. *Biochim. Biophys. Acta*, 1572:285–293, 2002.

[25] Melanie Demers, Thierry Magnalso, and Yves St-Pierre. A novel function for galectin-7: Promoting tumorigenesis by up-regulating MMP-9 gene expression. *Cancer Res.*, 65:5205–5210, 2005.

[26] Henri Debray, Dominique Decout, Gerard Strecker, Genevieve Spik, and Jean Montreuil. Specificity of twelve lectins towards oligosaccharides and glycopeptides related to N-glycosylproteins. *Eur J Biochem.*, 117:41–55, 1981.

[27] M. Wirth and A. Fuchs. Lectin-mediated drug targeting: Preparation, binding characteristics and antiproliferative activity of wheat germ agglutinin conjugated doxorubicin in Caco-2 cells. *Pharm. Res.*, 15(7):1031–1037, 1998.

[28] M. Wirth, G. Hamilton, and F. Gabor. Lectin-mediated drug targeting: Quantification of binding and internalization of wheat germ agglutinin and solanum tuberosum lectin using Caco-2 and HT-29 cells. *J. Drug Targeting*, 6(2):95–104, 1998.

[29] A. Clark, M. H. Hirst, B. and A. Jepson, M.˙ Lectin-mediated mucosal delivery of drugs and microparticles. *Adv. Drug Delivery Rev.*, 43:207–223, 2000.

[30] W. Yin and P.-W. Cheng. Lectin conjugate-directed gene transfer to airway epithelial cells. *Biochem. Biophys. Res. Commun.*, 205(1):826–833, 1994.

[31] M. Yamazaki, S. Kojima, V. Bovin, N. S. André, S. Gabius, and H.-J. Gabius. Endogenous lectins as targets for drug delivery. *Adv. Drug Delivery Rev.*, 43:225–244, 2000.

[32] Andrew R. Leach. *Molecular modelling: Principles and applications.* Prentice Hall, 2002.

[33] Chiara Taroni, Susan Jones, and Janet M. Thornton. Analysis and prediction of carbohydrate binding sites. *Prot. Eng.*, 13:89–98, 2000.

[34] A. Pusztai. *Plant Lectins.* Cambridge University Press, Cambridge, UK, 1991. See pages 78–95.

[35] A. Pusztai and S. Bardocz. Biological effects of plant lectins on the gastrointestinal tract: Metabolic consequences and applications. *Trends in Glycoscience and Glycotechnology*, 8:149–165, 1996.

[36] Matthew D. Eldridge, Christopher W. Murray, Timothy R. Auton, Gaia V. Paolini, and Roger P. Mee. 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.*, 11:425–445, 1997.

[37] Michael Levitt and Max. F. Perutz. Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.*, 201:751–754, 1988.

[38] Michiro Muraki. The importance of CH/π interactions to the function of carbohydrate binding proteins. *ProteinPeptLett*, 9:195–209, 2002.

[39] M. Muraki, M. Ishimura, and K. Harata. Interactions of wheat-germ agglutinin with GlcNAcβ1,6Gal sequence. *Biochim. Biophys. Acta*, 1569:10–20, 2002.

[40] K. Harata and M. Muraki. Crystal structures of Urtica dioica agglutinin and its complex with tri-N-acetylchitotriose. *J. Mol. Biol.*, 297(3):673–681, 2000.

[41] Motohiro Nishio, Minoru Hirota, and Yoji Umezawa. *The CH/π Interaction.* John Wiley and Sons, New York, 1998.

[42] Yoji Umezawa and Motohiro Nishio. CH/π interactions in the crystal structure of class I MHC antigens and their complexes with peptides. *Bioorg. Med. Chem.*, 6:493–504, 1998.

[43] Yoji Umezawa and Motohiro Nishio. CH/π interactions as demonstrated in the crystal structure of guanine-nucleotide binding proteins, src homology-2 domains and human growth hormone in complex with their specific ligands. *Bioorg. Med. Chem.*, 6:2507–2515, 1998.

[44] Maria Brandl, Manfred S. Weiss, Andreas Jabs, Jürgen Sühnel, and Rolf Hilgenfeld. CH$\cdots\pi$-interactions in proteins. *J. Mol. Biol.*, 307:357–377, 2001.

[45] D. Chandler. Hydrophobicity: Two faces of water. *Nature*, 417:491, 2002.

[46] Mary C. Chervenak and Eric J. Toone. A direct measure of the contribution of solvent reorganization to the enthalpy of ligand binding. *J. Am. Chem. Soc.*, 116:10533–10539, 1994.

[47] Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[48] Gareth Jones, Peter Willett, Robert C. Glen, Andrew R. Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

[49] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[50] Michal Vieth, Jonathan D. Hirst, Andrzej Kolinski, and Charles L. Brooks III. Assessing energy functions for flexible docking. *J. Comput. Chem.*, 19:1612–1622, 1998.

[51] Michal Vieth, Jonathan D. Hirst, Brian N. Dominy, Heidi Daigler, and Charles L. Brooks III. Assessing search strategies for flexible docking. *J. Comput. Chem.*, 19:1623–1631, 1998.

[52] David J. Diller and Christophe L. M. J. Verlinde. A critical evaluation of several global optimization algorithms for the purpose of molecular docking. *J. Comput. Chem.*, 20:1740–1751, 1999.

[53] R. D. Taylor, P. J. Jewsbury, and J. W. Essex. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.*, 16:151–166, 2002.

[54] Badry D. Bursulaya, Maxim Totrov, Ruben Abagyan, and Charles L. Brooks III. Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.*, 17:755–763, 2003.

[55] S. Ha, R. Andreani, A. Robbins, and I. Muegge. Evaluation of docking/scoring approaches: A comparative study based on MMP3 inhibitors. *J. Comput. Aided Mol. Des.*, 14:435–448, 2000.

[56] Hans-Joachim Böhm. 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.*, 8:243–256, 1994.

[57] Hans-Joachim Böhm. 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.*, 12:309–323, 1998.

[58] Didier Rognan, Sanne Lise Lauemller, Arne Holm, Sren Buus, and Vincenzo Tschinke. Predicting binding affinities of protein ligands from three-dimensional models: Application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.*, 42:4650–4658, 1999.

[59] Anna Maria Ferrari, Binqing Q. Wei, Luca Costantino, and Brian K. Shoichet. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.*, 47:5076–5084, 2004.

[60] Andreas Hildebrandt, Ralf Blossey, Sergej Rjasanow, Oliver Kohlbacher, and Hans-Peter Lenhof. Novel formulation of nonlocal electrostatics. *Phys. Rev. Lett.*, 93:108104, 2004.

[61] H. H. Uhlig. The solubilities of gases and surface tension. *J. Phys. Chem.*, 41(9):1215–1225, 1937.

[62] M. L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.

[63] H. Reiss, H. L. Frisch, and J. L. Lebowitz. Statistical mechanics for rigid spheres. *J. Chem. Phys.*, 31:369, 1959.

[64] H. Reiss, H. L. Frisch, and J. L . Lebowitz. Aspects of the staistical thermodynamics or real fluids. *J. Chem. Phys.*, 32:119–124, 1960.

[65] H. Reiss and S. W. Mayer. Theory of the surface tension of molten salts. *J. Chem. Phys.*, 34:2001–2003, 1961.

[66] J. L. Lebowitz, E. Helfand, and E. Praestgaard. Scaled particle theory of fluid mixtures. *J. Chem. Phys.*, 43:774–779, 1965.

[67] H. Reiss and D. M. Tully-Smith. Further development of scaled particle theory for rigid spheres: Application of the statistical thermodynamics of curved surfaces. *J. Chem. Phys.*, 55:1674–1689, 1971.

[68] H. Reiss and R. V. Casberg. Radial distribution function for hard spheres from scaled particle theory, and an improved equation of state. *J. Chem. Phys.*, 61:1107–1114, 1974.

[69] R. A. Pierotti. A scaled particle theory of aqueous and nonaqueous solutions. *Chem. Rev.*, 76:717–726, 1976.

[70] J. Langlet, P. Claverie, J. Caillet, and A. Pullmann. Improvements of the continumm model. 1. application to the calculation of the vaporization thernodynamic quantities of nonassociated liquids. *J. Phys. Chem.*, 92:1617–1631, 1988.

[71] Marie-Jose Huron and Pierre Claverie. Calculation of the interaction energy of one molecule with ist whole surrounding. I. Method and application to pure nonpolar compounds. *J. Phys. Chem.*, 76:2123–2133, 1972.

[72] A. I. Kitaygorodski. *Tetrahedron*, 14:230, 1961.

[73] John David Jackson. *Classical Electrodynamics, Third Ed.* John Wiley & Sons Inc., 1998.

[74] Richard M. Jackson and Michael J. E. Sternberg. A continuum model for protein protein interactions: Application to the docking problem. *J. Mol. Biol.*, 250:258–275, 1995.

[75] Zhongxiang Zhou, Philip Payne, Max Vasquez, Nat Kuhn, and Michael Levitt. Finite-difference solution of the poisson-boltzmann equation: Complete elimination of self-energy. *J. Comp. Chem.*, 17:1344–1351, 1996.

[76] Robert E. Bruccoleri, Jiri Novotny, Malcolm E. Davis, and Kim A. Sharp. Finite difference poisson-boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. *J. Comp. Chem.*, 18:268–276, 1997.

[77] Max Born. Volumen und Hydratationswärme der Ionen. *Z. Phys.*, 1:45–48, 1920.

[78] W. Clark Still, Anna Tempczyk, Ronald C. Hawley, and Thomas Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[79] Alexey Onufriev, Donald Bashford, and David A. Case. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem.*, 104:3712–3720, 2000.

[80] Vickie Tsui and David A. Case. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.*, 122:2489–2498, 2000.

[81] Vickie Tsui and David A. Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopoly.*, 56:275–291, 2001.

[82] Alexey Onufriev, Donald Bashford, and David A. Case. Exploring protein native states and large-scale conformational changes with a modifed generalized born model. *Proteins*, 55:383–394, 2004.

[83] Jayashree Srinivasan, Megan W. Trevathan, Paul Beroza, and David A. Case. Application of a pairwise generalized born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Acc.*, 101:426–434, 1999.

[84] Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, 100:19824–19839, 1996.

[85] Michael S. Lee, Freddie R. Salsbury, and Charles L. Brooks III. Novel generalized born methods. *J. Chem. Phys.*, 116:10606–10614, 2002.

[86] C. Satheesan Babu and Carmay Lim. Incorporating nonlinear solvent response in continuum dielectric models using a two-sphere description of the born radius. *J. Phys. Chem. A*, 105:5030–5036, 2001.

[87] M. Schaefer and C. Froemmel. A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.*, 216:1045, 1990.

[88] Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. Parametrized model for aqueous free energies of solvation using geometry-dependent atomic surface tensions with implicit electrostatics. *J. Phys. Chem.*, 101:7147–7157, 1997.

[89] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.

[90] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz Jr., David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

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

[92] William L. Jorgensen and Julian Tirado-Rives. The OPLS potential functions for proteins. energz minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. soc.*, 110:1657–1666, 1988.

[93] Serhat Saydam. Vergleich von Methoden zur Berechnung der molekularen Elektrostatik, 2005. Studienarbeit, Universität Tübingen.

[94] http://commons.wikimedia.org/wiki/Image:Protein_Crystal_Growth_Porcine_Elastase.jpg.

[95] http://en.wikipedia.org/wiki/Image:Myoglobindiffraction.png.

[96] Tarun K. Dam and C. Fred Brewer. Thermodynamic studies of lectin-carbohydrate interactions by isothermal titration calorimetry. *Chem. Rev.*, 102:387–429, 2002.

[97] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acid Research*, 28:235–242, 2000.

[98] R. J. Woods, R. A. Dwek, C. J. Edge, and B. Fraser-Reid. Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. GLYCAM_93 parameter development. *J. Phys. Chem.*, 99(11):3832–3846, 1995.

[99] Matthias Rarey, Bernd Kramer, and Thomas Lengauer. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins*, 34:17–28, 1999.

[100] Marcel L. Verdonk, Gianni Chessari, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor. Modeling water molecules in protein-ligand docking using gold. *J. Med. Chem.*, 48:6504–6515, 2005.

[101] Chris de Graaf, Chris Oostenbrink, Peter H. J. Keizers, Tushar van der Wijst, Aldo Jongejan, and Nico P. E. Vermeulen. Catalytic site prediction and virtual screening of cytochrome p450 2d6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.*, 49:2417–2430, 2006.

[102] N. P. Boghossian, O. Kohlbacher, and H.-P. Lenhof. BALL: Biochemical Algorithms Library. In J. Vitter and C. Zaroliagis, editors, *Algorithm engineering, 3rd international workshop, WAE'99*, volume 1668 of *Lecture Notes in Computer Science (LNCS)*, pages 330–344. Springer, 1999.

[103] Oliver Kohlbacher and Hans-Peter Lenhof. BALL - rapid software prototyping in computational molecular biology. *Bioinformatics*, 16:815–824, 2000.

[104] The Chemical Computing Group. *The Molecular Operating Environment.* The Chemical Computing Group, Montreal, Canada. http://www.chemcomp.com/.

[105] A. Bondi. Van der Waals volumes and radii. *J. Phys. Chem.*, 68:441–451, 1964.

[106] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery. General atomic and molecular electronic structure system. *J. Comput. Chem.*, 14:1347–1363, 1993.

[107] J. V. Pratap, A. A. Jeyaprakash, P. G. Rani, K. Sekar, A. Surolia, and M. Vijayan. Crystal structures of Artocarpin, a Moraceae lectin with mannose specificity, and its complex with methyl-α-D-mannose: Implications to the generation of carbohydrate specificity. *J. Mol. Biol.*, 317:237–247, 2002.

[108] P. G. Rani, K. Bachwhawat, S. Misquith, and A. Surolia. Thermodynamic studies of saccharide binding to artocarpin, a b-cell mitogen, reveals the extended nature of its interaction with mannotriose. *J. Biol. Chem.*, 274(42):29694–29698, 1999.

[109] J. H. Naismith, C. Emmerich, J. Habash, S. J. Harrop, J. R. Helliwell, W. N. Hunter, J. Raftery, A. J. Kalb, and J. Yariv. Refined structure of concanavalin-A complexed with methyl α-D-mannopyranoside at 2.0 Angstrom resolution and comparison with the saccharide-free structure. *Acta Cryst.*, D50:847–858, 1994.

[110] M. C. Chervenak and E. J. Toone. Calorimetric analysis of the binding of lectins with overlapping carbohydrate-binding specifities. *Biochemistry*, 34(16):5685–5695, 1995.

[111] S. J. Harrop, J. R. Helliwell, T. C. M. Wan, A. J. Kalb, L. Tong, and J. Yariv. Structure solution of a cubic crystal of concanavalin A complexed with methyl α-D-glucopyranoside. *Acta Cryst.*, D52:143–155, 1996.

[112] J. Bouckaert, T. Hamelryck, L. Wyns, and R. Loris. The crystal structures of Man(α1-3)Man(α1-O)Me and Man(α1-6)Man(α1-O)Me in complex with concanavalin A. *J. Biol. Chem*, 274:29188–29195, 1999.

[113] R. Loris, D. Maes, F. Poortmans, L. Wyns, and J. Bouckaert. A structure of the complex between concanavalin A and methyl-3,6-di-O-(α-D-mannopyranosyl)-α-D-mannopyranoside reveals two binding modes. *J Biol Chem*, 271:30614–30618, 1996.

[114] D. A. Rozwarski, B. M. Swami, C. F. Brewer, and J. C. Sacchettini. Crystal structure of the lectin from Dioclea grandiflora complexed with core trimannoside of asparagine-linked carbohydrates. *J. Biol. Chem.*, 273:32818–32825, 1998.

[115] S. Elgavish and B Shaanan. Structures of the Erythrina corallodendron lectin and of its complexes with mono- and disaccharides. *J. Mol. Biol.*, 277:917–932, 1998.

[116] A. Surolia, N. Sharon, and F. P. Schwarz. Thermodynamics of monosaccharide and disaccharide binding to Erythrina corallodendron lectin. *J. Biol. Chem.*, 271(30):17697–17703, 1996.

[117] V. Z. Pletnev, S. N. Ruzheinikov, I. N. Tsygannik, I. Mikhailova Yu, W. Duax, D. Ghosh, and W. Pangborn. The structure of pea lectin-D-glucopyranose complex at a 1.9 A resolution. *Russ. J. Bioorg. Chem.*, 23:469–, 1997.

[118] F. P. Schwarz, K. D. Puri, et al. Thermodynamics of monosaccharide binding to concanavalin A, pea (Pisum sativum) lectin and lentil (Lens culinaris) lectin. *J. Biol. Chem.*, 268(11):7668–7677, 1993.

[119] S. N. Ruzheinikov, I. Y. Mikhailova, I. N. Tsygannik, W. Pangborn, W. Duax, and V. Z. Pletnev. The structure of the pea lectin-D-mannopyranose complex at a 2.1 A resolution. *Russ. J. Bioorg. Chem.*, 24:277–, 1998.

[120] R. Ravishankar, K. Suguna, A. Surolia, and M. Vijayan. Structures of the complexes of peanut lectin with methyl-β-galactose and N-acetyllactosamine and a comparative study of carbohydrate binding in Gal/Galnac-specific legume lectins. *Acta Cryst.*, D55:1375–1382, 1999.

[121] K. J. Neurohr, N. M. Young, and H. H. Mantsch. Determination of the carbohydrate-binding properties of peanut agglutinin by ultraviolett difference spectroscopy. *J. Biol. Chem.*, 255(19):9205–9209, 1980.

[122] R. Banerjee, K. Das, R. Ravishankar, K. Suguna, A. Surolia, and M. Vijayan. Conformation, protein-carbohydrate interactions and a novel subunit association in the refined structure of peanut lectin-lactose complex. *J. Mol. Biol.*, 259:281–296, 1996.

[123] R. T. Lee, H. J. Gabius, and Y. C. Lee. Thermodynamic parameters of the interaction of Urtica dioica agglutinin with N-acetylglucosamine and its oligomers. *Glycoconj. J.*, 15(7):649–655, 1998.

[124] F. A. Saul, P. Rovira, G. Boulot, E. J. M. Van Damme, W. J. Peumans, P. Truffa-Bachi, and G. A. Bentley. Crystal structure of Urtica dioica agglutinin, a superantigen

presented by MHC molecules of class I and class II. *Structure Fold Des.*, 8:593–603, 2000.

[125] G. Bains, R. T. Lee, et al. Microcalorimetric study of wheat germ agglutinin binding to N-acetylglucosamine and its oligomers. *Biochemistry*, 31(50):12624–12628, 1992.

[126] Demetrios D. Leonidas, Efstratia H. Vatzaki, Henrik Vorum, Julio E. Celis, Peder Madsen, and K. Ravi Acharya. Structural basis for the recognition of carbohydrates by human galectin-7. *Biochemistry*, 37:13930–13940, 1998.

[127] C. Fred Brewer. Thermodynamic binding studies of galectin-1, -3 and -7. *Glycoconjugate Journal*, 19:459–463, 2004.

[128] G. Ramachandraiah, N. R. Chandra, A. Surolia, and M. Vijayan. Re-refinement using reprocessed data to improve the quality of the structure: A case study involving garlic lectin. *Acta Crystallogr. D Biol. Crystallogr.*, 58:414–420, 2002.

[129] Y. Bourne, C. H. Astoul, V. Zamboni, W. J. Peumans, L. Menu-Bouaouiche, E. J. Van Damme, A. Barre, and P. Rouge. Structural basis for the unusual carbohydrate-binding specificity of jacalin towards galactose and mannose. *Biochem. J.*, 364:173–180, 2002.

[130] G. A. De Souza, P. S. Oliveira, S. Trapani, A. C. Santos, J. C. Rosa, Laure H. J, V. M. Faca, M. T. Correia, G. A. Tavares, G. Oliva, L. C. Coelho, and L. J. Greene LJ. Amino acid sequence and tertiary structure of cratylia mollis seed lectin. *Glycobiology*, 13:961–972, 2003.

[131] M. M. Prabu, R. Sankaranarayanan, K. D. Puri, V. Sharma, A. Surolia, M. Vijayan, and K. Suguna. Carbohydrate specificity and quaternary association in basic winged bean lectin: X-ray analysis of the lectin at 2.5 A resolution. *J. Mol. Biol.*, 276:787–796, 1998.

[132] A. Rabijns, C. Verboven, P. Rouge, A. Barre, E. J. Van Damme, W. J. Peumans, and C. J. De Ranter. Structure of a legume lectin from the bark of robinia pseudoacacia and its complex with N-acetylgalactosamine. *Proteins*, 44:470–478, 2001.

[133] Y. Bourne, V. Zamboni, A. Barre, W. J. Peumans, E. J. Van Damme, and P. Rouge. Helianthus tuberosus lectin reveals a widespread scaffold for mannose-binding lectins. *Structure*, 7:1473–1482, 1999.

[134] C. Svensson, S. Teneberg, C. L. Nilsson, A. Kjellberg, F. P. Schwarz, N. Sharon, and U. Krengel. High-resolution crystal structures of erythrina cristagalli lectin in complex with lactose and 2'-α-l-fucosyllactose and correlation with thermodynamic binding data. *J. Mol. Biol.*, 321:69–83, 2002.

[135] X. Lee, A. Thompson, Z. Zhang, H. Ton-that, J. Biesterfeldt, C. Ogata, L. Xu, R. A. Johnston, and N. M. Young. Structure of the complex of Maclura pomifera agglutinin and the t-antigen disaccharide, galβ1,3galnac. *J. Biol. Chem.*, 273:6312–6318, 1998.

[136] R. Mikeska, R. Wacker, R. Arni, T. P. Singh, A. Mikhailov, A. Gabdoulkhakov, W. Voelter, and C. Betzel. Mistletoe lectin i in complex with galactose and lactose reveals distinct sugar-binding properties. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, 61:17–25, 2005.

[137] M. B. Shevtsov and I. N. Tsygannik. Structure of pea lectin in complex with α-methyl-D-glucopyranoside. to be published, 2003.

[138] J. M. Rini, K. D. Hardman, H. Einspahr, F. L. Suddath, and J. P. Carver. X-ray crystal structure of a pea lectin-trimannoside complex at 2.6 A resolution. *J. Biol. Chem.*, 268:10126–10132, 1993.

[139] P. Emsley, C. Fotinou, I. Black, N. F. Fairweather, I.G. Charles, C. Watts, E. Hewitt, and N. W. Isaacs. The structures of the h(c) fragment of tetanus toxin with carbohydrate subunit complexes provide insight into ganglioside binding. *J. Biol. Chem.*, 275:8889–8894, 2000.

[140] M. N. Vyas, N. K. Vyas, and F. A. Quiocho. Crystallographic analysis of the epimeric and anomeric specificity of the periplasmic transport/chemotactic protein receptor for d-glucose and d-galactose. to be published, 1993.

[141] M. A. Bianchet, E. W. Odom, G. R. Vasta, and L. M. Amzel. A novel fucose recognition fold involved in innate immunity. *Nat. Struct. Biol.*, 9:628–634, 2002.

[142] P. G. Telmer and B. H. Shilton. nsights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. *J. Biol. Chem.*, 278:34555–34567, 2003.

[143] T. Shirai, C. Mitsuyama, Y. Niwa, Y. Matsui, H. Hotta, T. Yamane, H. Kamiya, C. Ishii, T. Ogawa, and K. Muramoto. High-resolution structure of the conger eel galectin, congerin I, in lactose-liganded and ligand-free forms: Emergence of a new structure class by accelerated evolution. *Structure*, 7:1223–1233, 1999.

[144] D. I. Liao, G. Kapadia, H. Ahmed, G. R. Vasta, and O. Herzberg. Structure of s-lectin, a developmentally regulated vertebrate β-galactoside-binding protein. *Proc. Natl. Acad. Sci.*, 91:1428–1432, 1994.

[145] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[146] Walter Moreira and Gregory R. Warnes. *RPy (R from Python).* http://rpy.sf.net/.

[147] Florian Kirchner. Lecxplorer. ein Bindungstaschenfinder für Lektine, 2006. Studienarbeit, Universität Tübingen.

[148] J. Ruppert, W. Welch, and A. N. Jain. Automatic identification and representation of protein binding sites for molecular docking. *Prot. Sci.*, 6:524–533, 1997.

[149] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3:449–462, 1996.

[150] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.

[151] A. J. Hopfinger. *Conformational Properties of Macromolecules.* Academic Press, 1973.

[152] Anki Gustafsson, Anna Hultberg, Rolf Sjöström, Imre Kacskovics, Michael E Breimer, Thomas Borén, Lennart Hammarström, , and Jan Holgersson. Carbohydrate-dependent inhibition of helicobacter pylori colonisation using porcine milk. *Glycobiology*, in press, 2005.

[153] The Free Software Foundation. http://www.gnu.org/licenses/lgpl.html.

[154] M. Ostrogradsky. Note sur les intégrales definies. *Mém. Acad. Sci. St. Petersbourg. Sér. 6, Sci. Math. Phys. et Naturelles*, 1, 1831.