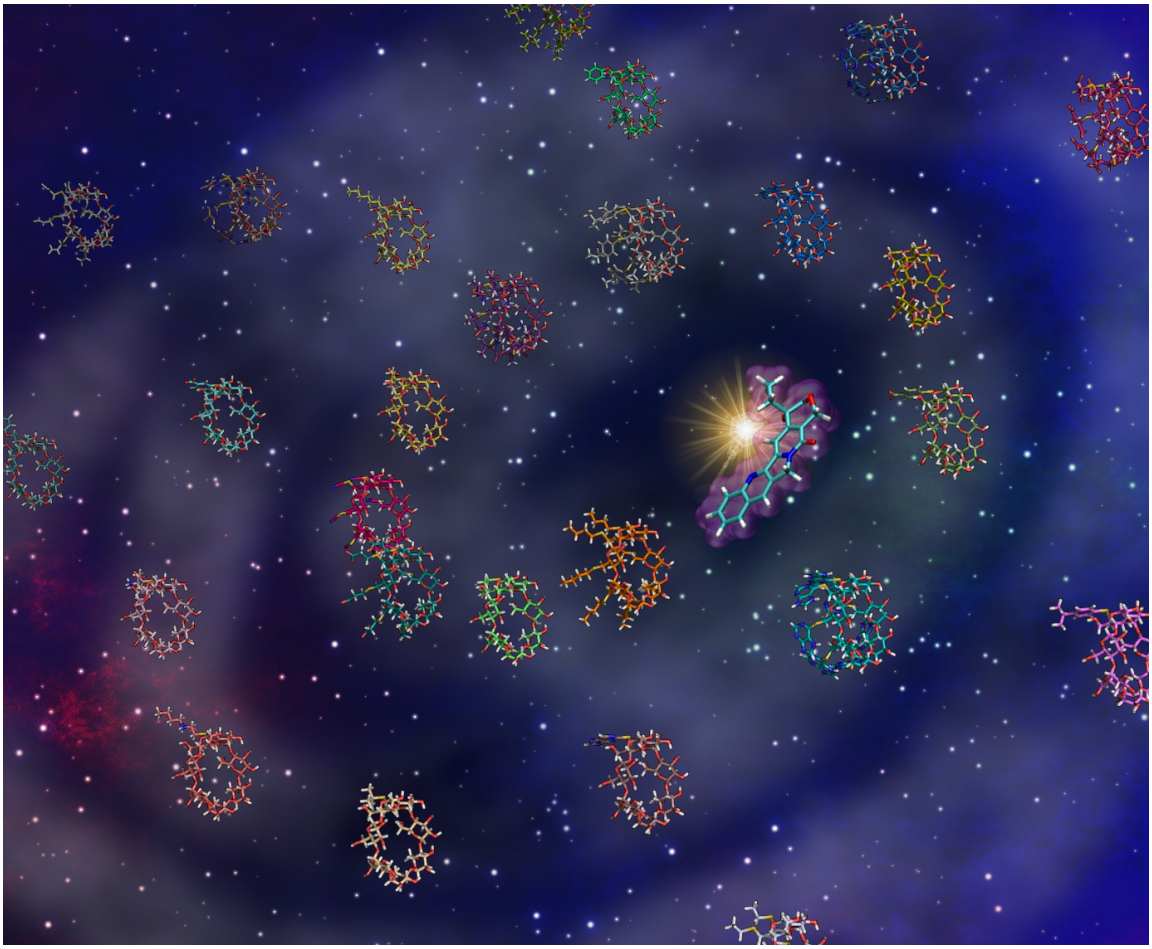
Computational Approaches in Supramolecular Chemistry with a Special Focus on Virtual Screening

Dissertation
zur Erlangung des Grades des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

Andreas Steffen
Saarbrücken 2007



Computational Approaches in Supramolecular Chemistry With a Special Focus on Virtual Screening

Dissertation
zur Erlangung des Grades des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
Andreas Steffen
Saarbrücken 2007

Tag des Kolloquiums: 16.01.2008

Dekan:	Prof. Dr. Uli Müller
Berichterstatter:	Prof. Dr. Thomas Lengauer, PhD
	Prof. Dr. Gerhard Wenz

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.
Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, der 17. Januar 2008

(Andreas Steffen)

Die vorliegende Arbeit entstand in der Zeit von Juni 2004 bis Juli 2007 am Max-Planck Institut für Informatik, Saarbrücken unter Leitung von Herrn Prof. Dr. Thomas Lengauer, PhD.

Meinen Eltern

Danksagung
Ich möchte mich bei allen bedanken, die mich bei der Erstellung der vorliegenden Arbeit unterstützt haben:

- Professor Dr. Thomas Lengauer, PhD danke ich für die Möglichkeit in seiner Arbeitsgruppe promovieren zu dürfen. Die Begleitung meiner Arbeit mit stetigem Interesse und seine Anregungen waren hilfreich und entscheidend für das Gelingen dieser Arbeit. Die gemeinsame Erstellung von Publikationen war lehrreich und prägend. Ich danke ihm ferner für die kritische Durchsicht dieser Arbeit.
- Professor Dr. Gerhard Wenz danke ich für die spannende interdisziplinäre Kooperation, in der ich viel gelernt habe. Ich danke für die Hilfe bei der Erstellung unserer gemeinsamen Publikationen, für viele gute Ratschläge, die kritische Durchsicht der vorliegenden Arbeit und für die Übernahme der zweiten Gutachtertätigkeit.
- Dr. Andreas Kämper danke ich dafür, dass er mich in die Gruppe gebracht hat, mich die Promotionszeit über betreut, begleitet und unterstützt hat und mich immer wieder ermutigt und aufgebaut hat. Ich danke ihm für die Möglichkeit mit ihm über Probleme und Fragestellungen im Projekt sprechen zu können. Ferner danke ich ihm für die hilfreichen Kommentare zur vorliegenden Arbeit.
- Dr. Joannis Apostolakis danke ich für die hervorragende Kooperation in den Cyclodextrin-Projekten. Die Zusammenarbeit war stets anregend und herausfordernd! Die Diskussionen mit Joannis haben entscheidend zum Gelingen dieser Arbeit beigetragen.
- Stephan Raub danke ich für die gute und freundschaftliche Zusammenarbeit im Rahmen unseres DFG-Projektes, sowie für die kritischen und sehr hilfreichen Kommentare zur Einleitung.
- Carolin Thiele und Dr. Christian Strassnig danke ich für die gute Zusammenarbeit und ihre Arbeit im Labor für unsere Projekte.
- Meinem Zimmerkollegen Christoph Hartmann danke ich für die gute Atmosphäre im Zimmer und die hilfreichen Programmiertipps.
- Der gesamten Arbeitsgruppe danke ich für die tolle Gemeinschaft sowie für anregende und unterhaltsame Gespräche.
- Über die Arbeit hinaus habe ich wertvolle Freunde gewonnen, mit denen ich viel erlebt habe und die meine Zeit hier sehr bereichert haben: Fidel Ramirez und seine kleine Familie, Hongbo Zhu und seine ebenso kleine Familie, Adrian Alexa, Christoph Bock, Tobias Sing, Dr. Ingolf Sommer, Jochen Maydt, Dr. Gaby Mayr, Dr. Francisco Domingues, Lars Kunert, André Altmann und Oliver Sander.
- Ruth Schneppen-Christmann und Dr. Joachim Büch danke ich für die umfassende Unterstützung.
- Der Deutschen Forschungsgemeinschaft danke ich für die Förderung dieser Arbeit durch das Projekt KA 1804/1.
- Kerstin Mahnkopf danke ich für ihre Fröhlichkeit, aber eigentlich überhaupt für alles und natürlich auch für die Bereitschaft mit mir ins ferne Saarbrücken zu ziehen.

- Meine Eltern haben mir den Weg bis hierher ermöglicht, mich immer in allen Dingen unterstützt und mich immer wieder aufgemuntert. Dafür und für alles andere bin ich ihnen sehr dankbar.

Abstract

Within this thesis novel computational tools for the rational design of synthetic host-guest complexes (SHGC) were developed and applied that employ the concepts of efficient virtual screening (VS) approaches. The first part describes the development of a fast structure prediction tool for flexible SHGC. The tool was validated on a test dataset comprising crystallographically determined SHGC. In nine of ten cases near-native solutions were generated. The tool can be applied for VS. In the second part of the thesis computational techniques were applied for designing SHGC based on β-cyclodextrins (β-CD). We performed a structure-based inverse virtual screening for identifying modified β-CDs as receptors for the anticancer drug camptothecin (CPT). Six of the proposed receptors exhibited binding affinities which were significantly higher than for any other CPT-receptor. Furthermore, we applied a combination of a similarity-based virtual screening technique with a regression model (RM) for identifying novel high affinity guest molecules of β-CD. Ten of the proposed guest molecules exhibited a binding free energy of lower than -20 kJ mol^{-1}. The last chapter describes a comparison of regression methods regarding their ability to generate predictive RM for thermodynamical parameters (ΔG, ΔH and ΔS) of β-CD-guest complexes. ΔG could be predicted in good agreement with experimental values, none of the methods led to comparably good predictive models for ΔH. ΔS appears almost unpredictable.

Kurzfassung

Im Rahmen dieser Arbeit wurden rechnergestützte Verfahren (RGV) zum gezielten Entwurf von synthetischen Wirt-Gast Komplexen (SWGK) entwickelt und eingesetzt. Dabei wurde ein Fokus auf schnelle virtuelle *Screening* (VS) Verfahren gelegt. Der erste Teil beschreibt die Entwicklung eines Programms zur schnellen Strukturvorhersage von flexiblen SWGK. Das Programm wurde auf einem Testdatensatz an kristallographisch vermessenen SWGK validiert. Für neun von zehn SWGK wurden nativ-ähnliche Lösungen gefunden. Das Programm kann für VS eingesetzt werden. Der zweite Teil der Arbeit behandelt RGV zum gezielten Entwurf von β-Cyclodextrin (β-CD) Komplexen. Mit Hilfe eines strukturbasierten inversen VS wurden sechs modifizierte β-CD-Rezeptoren für den Krebsarzneistoff Camptothecin (CPT) gefunden, die deutlich höhere Bindungsaffinitäten zu CPT aufwiesen als alle bislang bekannten CPT-Rezeptoren. Zur Identifizierung neuer hochaffiner Gäste von β-CD wurde ein ähnlichkeitsbasiertes VS Verfahren in Kombination mit einem Regressionsmodell (RM) eingesetzt. Zehn der mit Hilfe dieses Verfahrens vorgeschlagenen Moleküle wiesen eine Bindungsenergie von unter -20 kJ mol^{-1} auf. Das letzte Kapitel beschreibt einen Vergleich von drei Regressionsverfahren. Es wurde die Fähigkeit untersucht, vorhersagekräftige RM für thermodynamische Parameter (ΔG, ΔH und ΔS) von β-CD-Gast-Komplexen zu generieren. ΔG konnte mit allen Methoden sehr gut vorhergesagt werden, während ΔH nur begrenzt und ΔS unzureichend vorhersagbar war.

Contents

List of Figures

List of Tables

1

Introduction

The integration of computational methods into chemistry is of outstanding current interest and high impact. The field of medicinal chemistry, for example, has benefited enormously from computer science (Jorgensen, 2004; Bajorath, 2004). Concepts such as virtual screening nowadays take a substantial part in the drug discovery process and significantly influence the field (Kitchen et al., 2004; Lengauer et al., 2004). This thesis introduces novel computational techniques that transfer the concepts of efficient virtual screening approaches from the field of medicinal chemistry to supramolecular chemistry.

1.1 Supramolecular Chemistry

One of the major proponents of the field, Jean-Marie Lehn, described supramolecular chemistry as the "chemistry beyond the molecule" (Lehn, 1988, 1995). Other descriptions characterize supramolecular chemistry as the "chemistry of the noncovalent bond," or "the chemistry of molecular assemblies" (Steed & Atwood, 2000). An illustrative term is host-guest chemistry. This notion pinpoints the major characteristic of the molecules of interest: one molecule – the host molecule – has the potential to bind one or more molecules, termed guest molecules (see Figure 1.1). Host molecules are commonly larger than guest molecules and possess a sizable central hole or cavity (Steed & Atwood, 2000). In the context of this dissertation I use the term host molecule synonymously with synthetic receptor. The guest molecule may be a mono-atomic ion, a small molecular fragment, or a molecule such as a drug that fits into this cavity. The complex is held together by reversible interactions, for example hydrogen bonds, dispersive interactions, π-interactions, or hydrophobic effects (see Section 1.1.1). The host-guest complex formation requires a complementarity of both complex partners with respect to interactions and steric arrangement and results in a unique structural relationship in most cases.

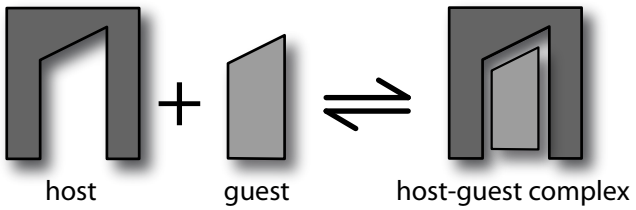


Figure 1.1. Schematic illustration of a host-guest complex. The host molecule possesses a cavity to which a guest molecule can bind. The host-guest complex formation involves a complementarity regarding shape and interaction groups.

Supramolecular chemistry dates back to the late 1950s. The origins, however, are to be found much earlier. Already in 1894, the lock-and-key principle by Emil Fischer picked up the main characteristics of host-guest chemistry and is today recognized as the major theoretical breakthrough thereof (see Section 1.1.2) (Fischer, 1894). The three chemists Charles Pedersen, Jean-Marie Lehn and Donald Cram developed the idea of imitating this principle from natural systems in a laboratory. Pedersen discovered crown ethers (see Figure 1.2 a), which are synthetic macrocycles with the ability to selectively bind cations (Pedersen, 1988). Based on this work, Lehn constructed cryptands, which are three-dimensional analogs of crown ethers and show a higher degree of preorganization (see Figure 1.2 b) (Lehn, 1988). Donald Cram contributed the so-called spherands (see Figure 1.2 c), amongst others (Cram, 1988). All three were awarded with the Nobel prize in chemistry in 1987 for their research in supramolecular chemistry.

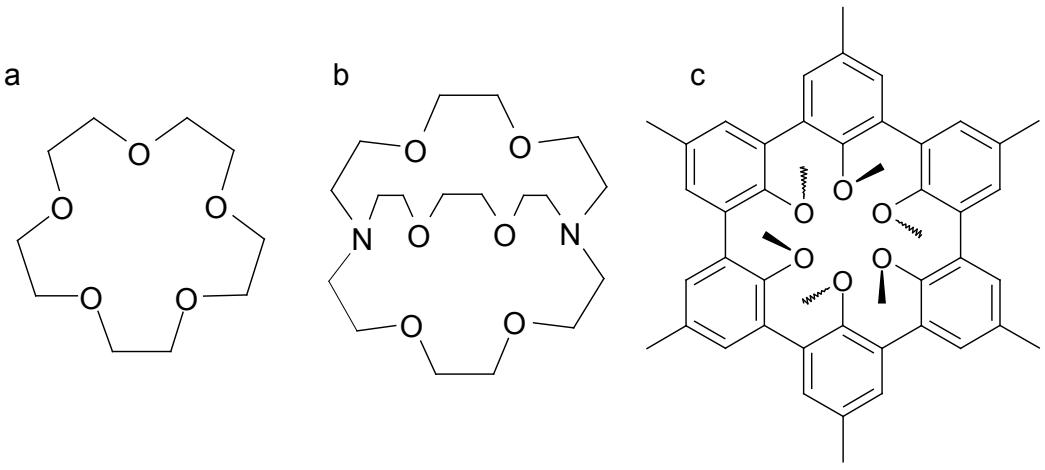


Figure 1.2. The first synthetic receptors: Crown ethers were synthesized by Pedersen (a). The three-dimensional analogs of crown ethers are called cryptands and were synthesized by Lehn. One of the first contributions of Cram to the field of supramolecular chemistry were spherands (c).

Since then, supramolecular chemistry has undergone huge developments and is already applied technically. Although synthetic receptors in general cannot rival their natural counterparts in terms of binding affinity and specificity, they do exhibit numerous advantages over natural receptors that make them interesting candidates for diagnostic (Bell et al., 1995), therapeutic (Schrader & Hamilton, 2005), analytical (Zadmard & Schrader, 2005), and separation purposes (Muderawan et al., 2006). Compared to natural receptors the three-dimensional structure of synthetic receptors is generally more stable at high temperatures and non-physiological pH conditions. Their comparably low molecular weights and their better tolerance by the human immune system predestines their use as drug-delivery agents (Davis, 2004).

Furthermore supramolecular chemistry contributed to the understanding and deciphering of molecular recognition and nowadays even impacts the structure-based design of drugs (Brooijmans & Kuntz, 2003). The following sections introduce the basic concepts of supramolecular chemistry and detail successful examples of supramolecular design.

1.1.1 Non-covalent Interactions

Traditional organic chemistry focuses on reactions that involve the formation or the break of covalent bonds. Supramolecular chemistry concentrates on non-covalent bonds (Schneider, 1991; Steed & Atwood, 2000). These types of interactions are typically much weaker than covalent bonds (a single bond between two carbons, for example, has an energy of 348 $\mathrm{kJ\,mol^{-1}}$) and their formation is therefore reversible under standard conditions. Table 1.1 gives representative examples of non-covalent interactions. In Table 1.2 the equations to calculate the strength of non-covalent interactions are shown and illustrated.

Electrostatic Interactions (Ion-Ion)

Electrostatic interactions occur between charged interaction groups and are described by Coulombs's law. Within crystals the strength of this type of interaction lies in the range of covalent bonds (100-350 $\mathrm{kJ\,mol^{-1}}$) (Steed & Atwood, 2000). In solvents the strength of electrostatic interactions decreases, being a linear function of the reciprocal of the dielectric constant. In water, for example, the ion pair $Ca^{2+}{\cdot}SO_4^{2-}$ exhibits a binding free energy ΔG of -13.2 $\mathrm{kJ\,mol^{-1}}$ (Schneider & Yatsimirsky, 2000). The strength of an electrostatic interaction is proportional to $\frac{1}{r}$.[1]

[1] with r equal to the distance between the two interacting entities

Table 1.1. Representative literature examples of non-covalent interactions.

Interaction type	Example	ΔG in kJ mol^{-1}	Medium	Lit.[1]
Ion-ion	$Ca^{2+}\cdots SO_4^{2-}$	-13.2	H_2O	a
Ion-dipole	$(\text{-}CH_2)_2O\cdots{}^{\oplus}N$ t-butyl	-3	$CHCl_3$	b
Dipole-dipole	$R_2C{=}O\cdots O{=}CR_2$ [2]	-20	calc. *in vacuo*	c
Hydrogen bond	$HOH\cdots OH_2$	-22	gas phase	a
Cation-π	$K^+\cdots$benzene	-50	gas phase	a
π-π	adenine$\cdots$benzene	-1.85 ± 0.15	$CHCl_3$	a
Dispersive interactions	-CH$\cdots$HC-	-0.2	calc. *in vacuo*	b
Hydrophobic interactions	$\text{-}CH_2\text{-}\cdots\text{-}CH_2\text{-}$	-2.3	water	b

[1] a = Schneider & Yatsimirsky (2000), b = Schneider (1991), c = Steed & Atwood (2000)
[2] parallel orientation

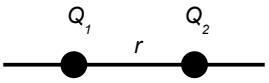
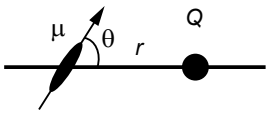
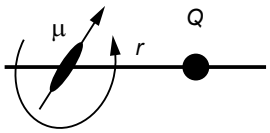
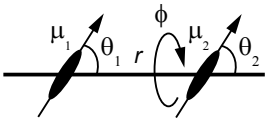
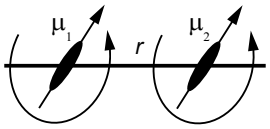
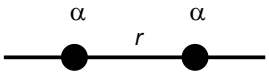
Ion-Dipole Interactions

The intermolecular interactions found in the complex of a sodium ion and a crown ether are an example for a cation (sodium ion) dipole (carbon oxygen bonds) interaction (Pedersen, 1988). This supramolecular system was in fact one of the first host-guest systems described in supramolecular chemistry. The bond energies of ion-dipole interactions vary in a range of 50-200 kJ mol^{-1} in the gas phase (Steed & Atwood, 2000). The interaction energy clearly decreases in polar solvents. The interaction energy of an ether group and tert-butylammomium in trichlormethane, for example, amounts to about -3 kJ mol^{-1} (Schneider, 1991). Ion-dipole interactions are directional interactions since they depend on the orientation of the dipole. With an increasing distance the energy of a fixed dipole-ion interaction decreases with a $\frac{1}{r^2}$ dependence, whereas the interaction energy between a freely rotating dipole and an ion decreases with a $\frac{1}{r^4}$ dependence.

Dipole-Dipole Interactions

Dipole-dipole interactions occur between molecules with permanent dipoles. A permanent dipole exists if electrons of a chemical bond are permanently distributed in a non-uniform manner. The energy of dipole-dipole interactions lies in the range of 5-50 kJ mol^{-1} in the gas phase and depends on the orientation of the dipoles to each other. In an *in vacuo* calculation, two carbonyl groups that orientate in a parallel manner show a binding free energy of -20kJ mol^{-1}. A fixed dipole-dipole interaction decreases with a $\frac{1}{r^3}$ dependence. Freely rotating dipoles

Table 1.2. Non-covalent interactions together with the equations to calculate their interaction energy. Q, electric charge; ϵ, dielectric constant; r, distance between the interacting entities; k, Boltzmann constant; T, absolute temperature; h, Planck's constant; v, electronic absorption frequency; α, electric polarizability.

Interaction type	Illustration	Interaction energy
Ion-ion		$\frac{Q_1 Q_2}{4\pi\epsilon_0 r}$
Ion-dipole (fixed)		$\frac{Q\mu\cos\theta}{4\pi\epsilon_0 r^2}$
Ion-dipole (rotating)		$\frac{Q^2\mu^2}{6(4\pi\epsilon_0)^2 kTr^4}$
Dipole-dipole (fixed)		$\frac{\mu_1\mu_2(2\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2\cos\phi)}{4\pi\epsilon_0 r^3}$
Dipole-dipole (rotating)		$\frac{\mu_1^2\mu_2^2}{3(4\pi\epsilon_0)^2 kTr^6}$
Dispersive interactions		$\frac{3hv\alpha^2}{4(4\pi\epsilon_0)^2 r^6}$
Hydrogen bond		energy roughly $\propto \frac{1}{r^2}$
Cation-π		energy roughly $\propto \frac{1}{r^n}$, with $n < 2$
π-π		complicated, short range

show a $\frac{1}{r^6}$ distance dependence. Aside from dipoles, quadrupoles and higher multipoles also show a similar type of interaction, generally of lower binding energy.

Hydrogen Bonding

A hydrogen bond is sometimes described as a particular type of a dipole-dipole interaction (Steed & Atwood, 2000). However, this description was criticized significantly, since only a very weak correlation between the strength of a hydrogen bond and the dipole moment of the hydrogen bond acceptor was found (Buckingham, 2000). Current research shows that a hydrogen bond should rather be regarded as a combination of interaction types, such as electrostatic interactions

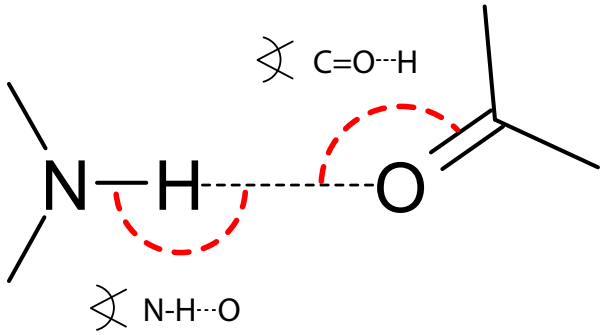


Figure 1.3. The geometry of a hydrogen bond. The distance O - N is between 2.8 and 3.2 Å. The angle N–H···O is larger than 150^o in most cases. The angle C=O···H is typically in between 100° and 180°.

(Stone, 2000), induction (Hodges et al., 1997), dispersive interactions (Wennmohs et al., 2003), repulsive and charge transfer interactions (Mo, 2006). A hydrogen bond is formed if a hydrogen atom that is attached to an electronegative atom (hydrogen bond donor) is attracted by a partially negatively charged atom (hydrogen bond acceptor). Hydrogen bonds are highly directional and restrictive regarding bond angle and length (see Figure 1.3). The binding energy of a hydrogen bond can differ enormously depending on the chemical surrounding of the interacting atoms and the medium in which it formed. It lies in a range of 4-120 kJ mol^{-1} (Steed & Atwood, 2000). It should, however, be noted that strong hydrogen bonds such as between F-H···F$^-$ already exhibit a semi-covalent behavior. Hydrogen bonds play a crucial role in biological systems such as proteins and the double-stranded DNA (Cooke & Rotellob, 2002). The hydrogen bond interaction dominated binding energy of the Watson-Crick base pairing between adenine and thymine (two hydrogen bonds) amounts to -8.5 kJ mol^{-1} in $CDCl_3$, whereas in the case of guanin and cytosin (three hydrogen bonds) it is equal to -24.5 kJ mol^{-1} (Schneider & Yatsimirsky, 2000) although there is only one additional hydrogen bond. The reason for this lies in secondary interactions, which are repulsive in the adenine-thymine pairing but favorable in the other.

Cation-π Interactions

Cations can interact with the π-face of an aromatic ring. The interaction results from an attraction of the positive charge of the cation and the π-electron system. This type of interaction highlights the role of supramolecular research for the deciphering of natural molecular recognition phenomena, since it could be investigated intensively on supramolecular model systems in the gas phase and in aqueous solutions (Dougherty, 1996). The binding energy between a potassium ion and benzene, for example, amounts to about -50 kJ mol^{-1} in the gas phase.

Dougherty (1996) found a distance dependence of the interaction energy of $\frac{1}{r^n}$, with $n < 2$. In nature, cation-π interactions occur between the aromatic amino acids phenylalanine, tyrosine, and tryptophan and cationic ligands or substrates. The binding of the positively charged acetylcholine to the enzyme acetylcholine esterase involves a cation-π interaction (Sussman et al., 1991), amongst others.

π-Stacking

The interaction between aromatic ring systems has been described as π-stacking (Hunter & Sanders, 1990). π-π interactions are presumed to be caused by an overlap of π-orbitals of aromatic systems from different molecules. The term stacking refers to the stacked arrangement of the aromatics. They are most favorable if an electron poor aromatic, such as pyridine, interacts with an electron rich aromatic (e. g. phenol). The energetic strength of a π-π interaction lies in the range 0 – 50 $\mathrm{kJ\,mol^{-1}}$ (Steed & Atwood, 2000). The binding free energy between benzene and adenine amounts to about -1.85 $\mathrm{kJ\,mol^{-1}}$ in chloroform. In nature π-π interactions occur, for example, between consecutive base pairs of the double-stranded DNA.

Dispersive Interactions

Dispersive interactions result from the attraction of two temporarily induced dipoles. They are non-directional and thus rather non-specific. The energy of dispersive interactions amounts to less than 5 $\mathrm{kJ\,mol^{-1}}$ (Steed & Atwood, 2000). The interaction energy between two C-H (carbon-hydrogen) groups, for example, is -0.2 $\mathrm{kJ\,mol^{-1}}$ in an *in vacuo* calculation (Schneider, 1991). Since the strength of this interaction decreases rapidly with the rising distance of the interacting atoms ($\propto \frac{1}{r^6}$), this type of interaction is highly dependent on the shape complementarity between the host and the guest molecules.

Hydrophobic Interactions

Hydrophobic interactions occur in polar solvents and result from a mixture of entropic and enthalpic effects (Abraham, 1982; Smithrud et al., 1990). In water and other polar solvents hydrophobic molecules aggregate and thus decrease the hydrophobic surface that is exposed to the solvent. Water molecules are thereby liberated and entropy increases. Simultaneously, novel attractive interactions are formed between the hydrophobic molecules, and new hydrogen bonds are formed between the deliberated water molecules. The interaction energy of two methylene groups (-CH_2-) amounts to -2.3 $\mathrm{kJ\,mol^{-1}}$ in water (Schneider, 1991).

1.1.2 Molecular Recognition

Two molecules that are complementary regarding shape and interaction groups exhibit the principle structural features for forming a supramolecular complex. Molecular recognition describes the selective binding process of a guest to a host molecule that accounts for particular physicochemical properties of the guest (Lehn, 1988; Schneider, 1991). The term goes back to Emil Fischer who introduced his lock-and-key principle in his breakthrough publication "Einfluss der Konfiguration auf die Wirkung der Enzyme" (Fischer, 1894). In this article, he investigated the specificity of the enzymes invertine and emulsin for the α- or the β-form of glucosides, respectively. The illustrative comparison of the complementarity of a key to a lock corresponds to the way a small molecule binds to an enzyme. While the necessity of shape complementarity is sufficiently covered by this comparison, the need for interaction complementarity is however neglected. That the isolated host molecule does not necessarily complement the shape of the guest was only discovered years later. During the binding it can undergo a so-called induced fit and adopts to the conformation leading to ideal shape complementarity (Koshland, 1994).

Molecular recognition can be understood as an information storage system on the supramolecular level (Lehn, 1988). The information is stored in the steric arrangement of the complex, i. e. the architecture of the complex, or by the thermodynamical properties of the complex. Usually molecular recognition is quantified by determining the binding free energy of the complex. The ability of a chemical entity to account for structural and chemical properties of another has received considerable attention in supramolecular research. A score of examples is detailed below.

Molecular Recognition of Length

Harada & Kataoka reported on a supramolecular system in which molecular recognition based on length was observed (Harada & Kataoka, 1999). In their study, polyanionic and polycationic block copolymers of varying chain lengths were synthesized. In aqueous solution, exclusively oppositely charged pairs of copolymers were formed that exhibited the same block length. These pairs then formed larger core-shell-type assemblies with a narrow size distribution (see Figure 1.4).

Molecular Recognition of Geometry

Synthetic receptors with the ability to selectively recognize a guest molecule of tetrahedral geometry were presented by Graf & Lehn (1975). The macrotricycles

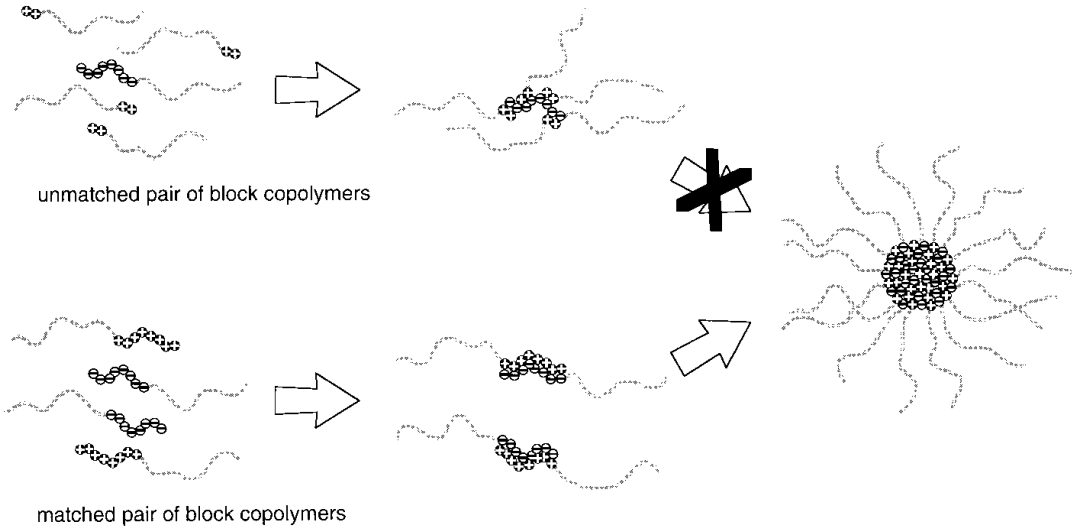


Figure 1.4. The supramolecular system recognizes molecular length. In solution only oppositely charged pairs of polymers with the same block length were found. As a result the core-shell type assemblies showed a narrow size distribution. From Harada & Kataoka (1999).

shown in Figure 1.5 form extremely strong complexes with their guest molecules, the tetrahedral ammonium ion and water, respectively.

Molecular Recognition of Thickness

Müller and Wenz described the ability of α-cyclodextrins to account for the thickness of molecular guests (Müller & Wenz, 2007). In their study, a series of bolaamphiphiles with increasing central thickness was synthesized (see Figure 1.6). The central thickness was calculated based on electron density maps created by semiempirical calculations. An indirect proportionality of molecular

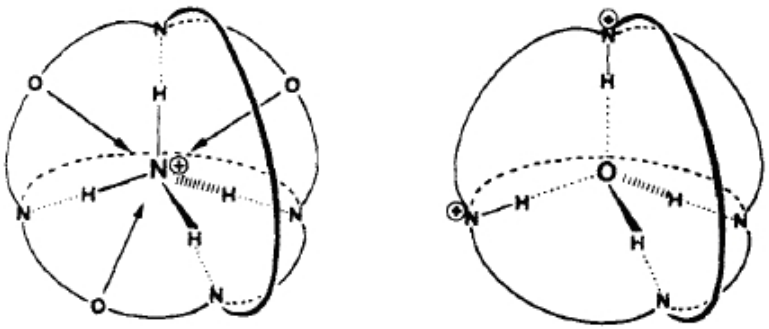


Figure 1.5. Both synthetic receptors shown in this figure account for the tetrahedral geometry of their comprised molecular guests. From Lehn (1988).

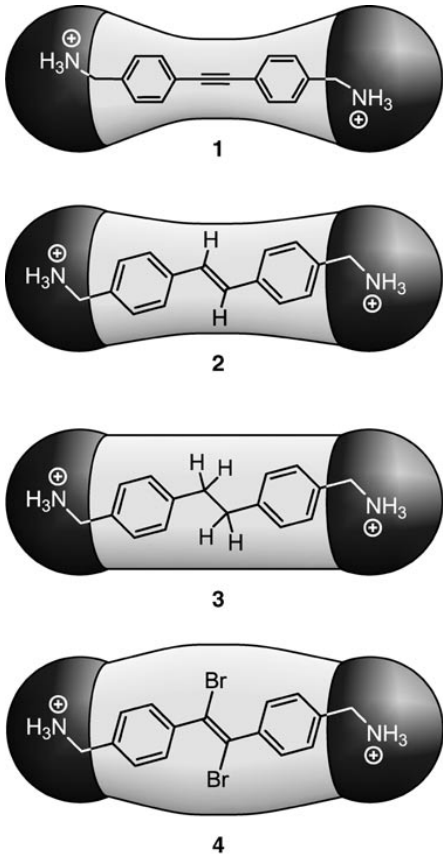


Figure 1.6. Müller & Wenz investigated the molecular recognition of thickness. The central thickness of the shown bolaamphiphiles increases from the first to the forth molecule. The molecular thickness correlates indirectly with the binding affinity to α-cyclodextrin. The fourth molecule exceeds a threshold and no complexation to the α-cyclodextrin cavity could be detected. From Müller & Wenz (2007).

thickness and the binding free energy ΔG to α-cyclodextrin was observed and the absolute value of the binding free energy decreases from the first to the third molecule ($\Delta G^0(\mathbf{1}) = -22.54\,\mathrm{kJ\,mol^{-1}} \pm 0.38$, $\Delta G^0(\mathbf{2}) = -16.28\,\mathrm{kJ\,mol^{-1}} \pm 0.10$ and $\Delta G^0(\mathbf{3}) = -10.13\,\mathrm{kJ\,mol^{-1}} \pm 0.29$). No binding was detected for the forth compound, since its thickness exceeded the diameter of the α-cyclodextrin cavity.

Molecular Recognition of Size, Shape, and Charge.

Hof et al. (2003) presented a biomimetic receptor for acetylcholine based on a cavitand (see Figure 1.7). Four negatively charged carboxylate groups were positioned at the entrance of a deep pocket. For the tetramethylammonium cation complex formation could be measured by NMR titrations ($K_a = 3800\,\mathrm{M^{-1}} \pm 600$), whereas the larger tetrapropyl- and tetrabutylammonium did not show any evidence of complex formation, presumably due to steric barriers. The positively charged molecules choline and acetylcholine were bound with high affinity

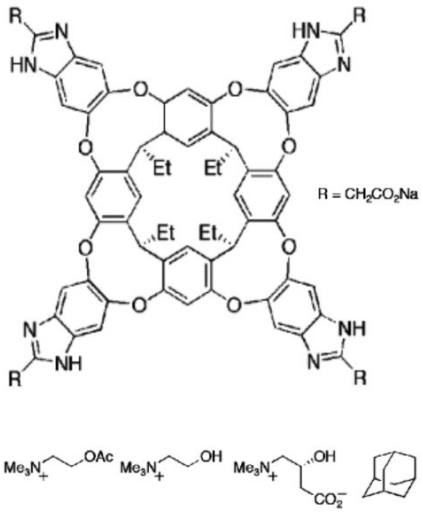


Figure 1.7. The synthetic receptor accounts for size, shape and charge. High binding affinities were found for the positively charged choline molecule (charge recognition) and adamantane, which shows a strong shape complementarity with the deep pocket (shape recognition). From Hof et al. (2003).

($K_a = 25900\,\mathrm{M}^{-1} \pm 700$ and $K_a = 14600\,\mathrm{M}^{-1} \pm 1200$), whereas the structurally closely related but zwitterionic molecule L-carnithin only exhibited a weak binding affinity ($K_a = 150\,\mathrm{M}^{-1} \pm 10$). Adamantane forms a 1:1 complex to the receptor as detected by ^{1}H NMR, suggesting a strong shape complementarity with the deep pocket of the receptor.

1.1.3 Host Design

Until recently, the design of supramolecular systems has often been the result of longtime work, experience, time-consuming trial and error approaches or even serendipity. Often, numerous optimization steps are accomplished before a sufficient binding affinity of a synthetic receptor to a particular guest molecule is achieved. The work of Schrader et al. towards a synthetic adrenaline receptor offers insights into the steps involved in the rational design of a synthetic host (Schrader, 1996, 1998; Herm & Schrader, 2000; Herm et al., 2001).

Over the years the Schrader group has presented successive generations of synthetic adrenaline receptors, that aim to mimic the architecture of the native archetype, the β-adrenergic receptor. Figure 1.8 illustrates the presumed binding mode of the neurotransmitter noradrenaline to the β-adrenergic receptor (Schrader, 1996). The ammonium functionality of noradrenaline is bound via Coulombic interactions and hydrogen bonds to an aspartate carboxylate group of the receptor. Furthermore, the ammonium protons form out cation-π inter-

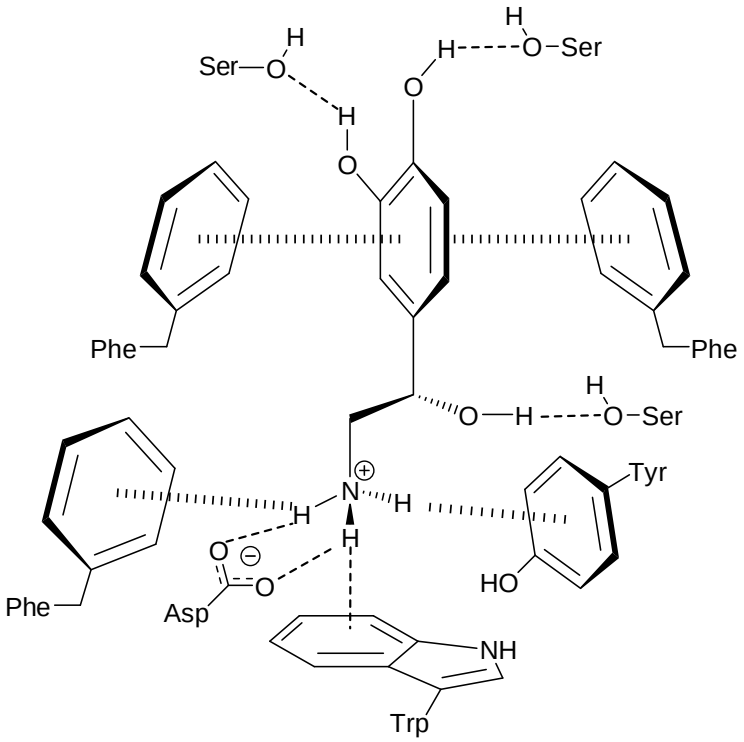


Figure 1.8. The interaction of noradrenaline with the binding site of the β-adrenergic receptor. Within this complex three types of interactions are found: π-stacking, cation-π interaction and hydrogen bonding. Adopted from Schrader (1996).

actions with three electron-rich aromatic residues (Phe, Tyr, Trp). Each of the hydroxyl groups of noradrenaline interacts through the formation of a hydrogen bond to one of the serine hydroxyl groups. The aromatic catechol ring is buried between two phenylalanine aromatic rings so that a double π-stacking is presumed (Trumpp-Kallmeyer et al., 1992).

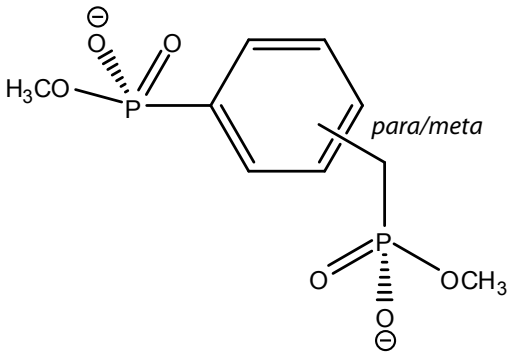


Figure 1.9. The first structural motif for a synthetic receptor that mimics the interaction properties of the β-adrenergic receptor. The receptor is based on a bisphophonate motif in meta- or para-position. Adopted from Schrader (1996).

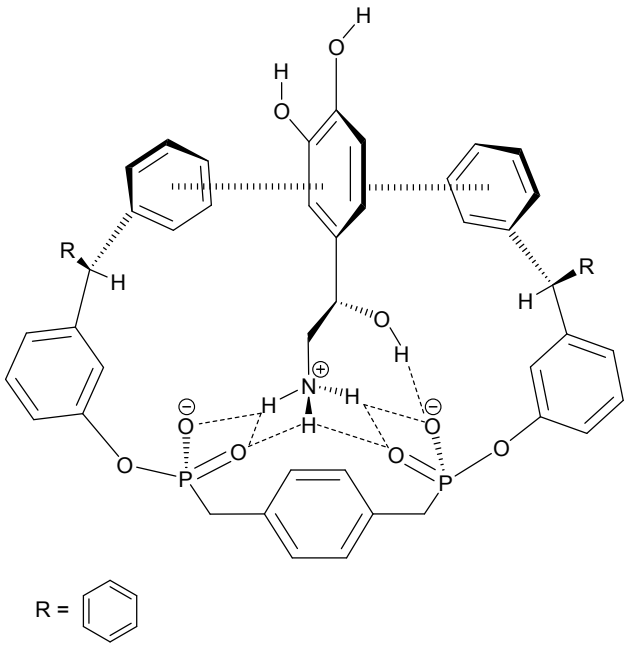


Figure 1.10. The second generation of synthetic receptors that mimic β-adrenergic receptors. To allow for π-stacking of the receptor with the catechol moiety of noradrenaline aromatic groups were added. Adopted from Schrader (1998).

Schrader started with a basic structural motif based on two phosphonate groups and an aromatic ring (see Figure 1.9) (Schrader, 1996). The para-form of the receptor exhibited strong binding affinities to 1,2-amino alcohols (R-propanolol $K_a = 66000\,\mathrm{M}^{-1}$ in DMSO). The major driving force for complex generation is the formation of hydrogen bonds and electrostatic interactions between the amino and hydroxyl group of the guest molecules and the phosphonate oxygens of the receptor. These interactions imitate the ammonium-aspartate interaction in the native β-adrenergic receptor. Cation-π interaction – although structurally plausible – could not, however, be experimentally verified.

The second generation of synthetic adrenaline receptors focused on the mimicry of the double sandwich-type π-stacking in the natural counterpart (Schrader, 1998). The aim was to retain both phosphonate groups for selective binding of the amino alcohol moiety, but also to provide a hydrophobic binding epitope for the aromatic catechol ring of noradrenaline. Based on the first structural motif, phosphonate esters with aromatic rings were synthesized (see Figure 1.10). For the receptor shown in Figure 1.10, π-stacking was observed with D-(-)-threo-2-amino-1-(4-nitrophenyl)-1,3-propanediol. However, the association constant was clearly lower compared to the ones obtained for the first receptor ($K_a = 18900\,\mathrm{M}^{-1}$ in DMSO). Reasons for this were mainly seen in entropic disadvantages due to an increase of rotational degrees of freedom.

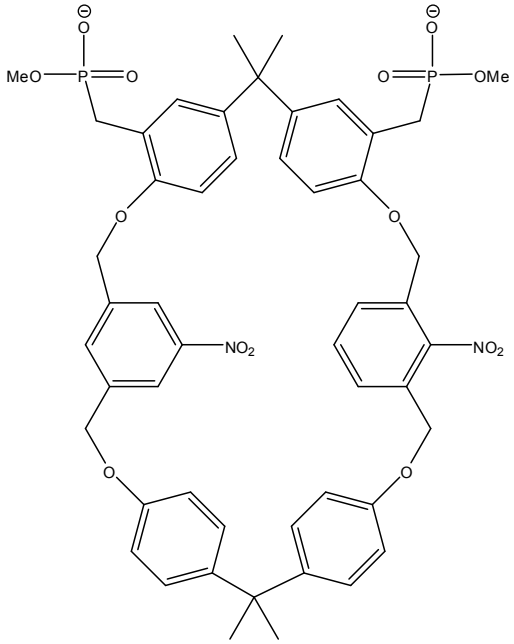


Figure 1.11. The third generation of synthetic receptors that mimic β-adrenergic receptors. To increase the preorganization of the receptor a macrocycle was introduced. Adopted from Herm & Schrader (2000).

This was tackled in the third generation. A new approach was introduced which is based on the preorganization concept of the host molecule (Herm & Schrader, 2000). The novel synthetic host shown in Figure 1.11 consists of a macrocycle based on aromatic building blocks and two attached phosphonate groups. In order to obtain electron poor aromatics, two of the aromatic building blocks had nitro groups as substituents. High binding affinities were measured for 1,2-amino alcohols (e.g. R-propanolol $K_a = 22,500\,\mathrm{M}^{-1}$ in DMSO, $K_a = 1,250\,\mathrm{M}^{-1}$ in methanol) but no π-stacking could be detected for catecholamines. Herm & Schrader (2000) assumed that the foreseen aromatic units could perhaps not orientate in the appropriate manner due to steric barriers.

The fourth generation reverted to the original structural motif of the first receptor but held on to the preorganization concept of the third (Herm et al., 2001). The novel macrocycle asserted an orientation of the aromatic parts for which a π-π stack to the catechol ring of noradrenaline was observed (see Figure 1.12). Furthermore, an isophtalamide group was inserted resulting in additional hydrogen bonds to a phenolic hydroxyl group of the guest molecule. This synthetic receptor finally was able to bind dopamine and noradrenaline in water with comparably high affinity ($K_a = 246\,\mathrm{M}^{-1} \pm 38\%$ and respectively $K_a = 215\,\mathrm{M}^{-1} \pm 12\%$).

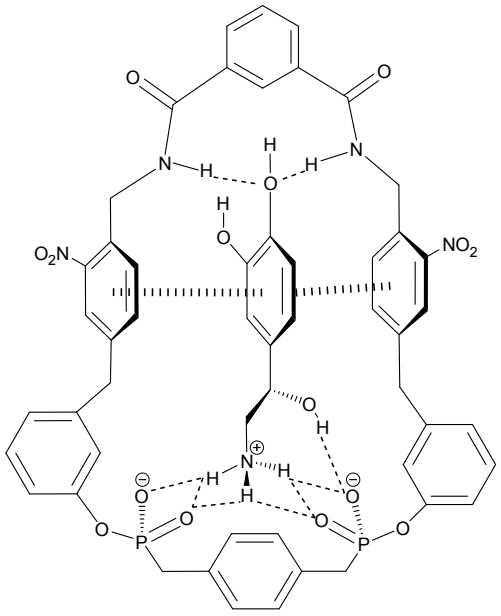


Figure 1.12. The fourth generation of synthetic receptors that mimic β-adrenergic receptors. This receptor reverts to the original bisphosphonate motif and retains the macrocycle. Adopted from Herm et al. (2001).

1.1.4 Computational Approaches in Supramolecular Chemistry

Today, computational chemistry plays an important role in the development of synthetic host-guest complexes. It comprises a variety of computational methods. Some of these simulate the dynamical behavior of chemical systems (e. g. molecular dynamics). Others aim at the prediction of thermodynamical properties (e. g. quantitative structure property relationships). Sometimes computational methods (e. g. *ab initio* calculations) provide precise access to properties that might be inaccessible for experimental procedures (Connors, 1997; Lamb & Jorgensen, 1997; Lipkowitz, 1998). In the following sections representative computational approaches are summarized together with their applications in supramolecular chemistry.

1.1.4.1 Quantitative Structure Property Relationship

Quantitative structure property relationship (QSPR) techniques try to correlate a property of a molecule or a molecular system to calculated structural descriptors (Hansch & Fujita, 1964; Yang & Huang, 2006). As a result, statistical models are obtained that are potentially able to predict a molecular property of interest on the basis of molecular descriptors. Beyond this, the generated statistical models can sometimes be interpreted and thus help to reveal the importance of particular descriptors for the predicted property. A large variety of computable molecular

descriptors is available for application in QSPR studies (Todeschini & Consonni, 2000). These descriptors account for example for simple one-dimensional properties such as molecular weight, or the occurrence of a defined fragment, for two-dimensional properties such as the topology of a molecule, for three-dimensional properties that describe molecular shape, and for elaborate quantum chemical properties.

To generate a QSPR model, a set of molecules is required for which the considered property is already known, for example, from experimental measurements. This set serves as the training basis for which a regression method generates the QSPR model. Usually, the generation of a QSPR model involves a descriptor selection that aims at reducing the number of descriptors that do not contribute to the model generation, thus enhancing the interpretability and generalizability of the model (Blum, 1997).

Katritzky et al. (2004) conducted a QSPR study, aiming at a QSPR model for the prediction of binding free energies of host-guest complexes between β-cyclodextrin and various guest molecules. Experimental binding free energies of 218 guest molecules served as the training set. QSPR regression models were built on the basis of two different molecular descriptors derived from two different tools, namely CODESSA-PRO [www.codessa-pro.com], which comprise a large variety of descriptors, and TRAIL (Solov'ev et al., 2000), which are fragment based descriptors. In tenfold cross-validation, cross-validated squared linear correlation coefficients r^2_{cv} of 0.78 and 0.85, respectively were obtained for the predicted and the experimental binding free energies. For the TRAIL descriptors, the training set had to be reduced to 195 molecules since the remaining molecules exhibited fragments of rare occurrence. Thus, the comparison appears unfair.

1.1.4.2 Docking

Computational docking addresses the questions whether two molecules (e. g. a protein and a ligand) form a complex, how strong this complex is, and what the complex looks like (Rarey et al., 2007). In principle, the docking algorithm consists of two parts: first, the pose-generation algorithm, and second, the scoring function, which facilitates a prediction of the binding free energy ΔG of the complex. In contrast to time-consuming molecular dynamics simulations (see Section 1.1.4.5), docking tools are designed to provide solutions in the range of seconds to minutes. This efficiency allows for so-called virtual screenings (see Section 1.2). Therefore efficient algorithms are needed for both, the docking and the energy assessment.

The pose generation comprises the search for an optimal translation, orientation and conformation of the ligand within the binding site of the protein. Different algorithmic approaches have been proposed (Brooijmans & Kuntz, 2003). Some generate conformational ensembles for the ligand before the placement and then dock each single representative independently into the binding site (Halgren et al., 2004; McGann et al., 2003). Others rely on the fragmentation of the ligand. From the generated fragments, the ligand is reconstructed within the binding site (Kuntz et al., 1982; Rarey et al., 1996a). Furthermore, evolutionary algorithms have been proposed (Jones et al., 1997; Morris et al., 1998). Here, the conformation, translation, and orientation of the ligand is coded into a virtual chromosome. Genetic operations such as mutation and crossover (i.e. recombination) act on these virtual chromosome and thus cast the search for an optimal pose as an evolutionary process. During the binding process the protein is presumed to remain rigid. However, several attempts were made to incorporate flexibility on the receptor side (see Section 2.1.1) (Claussen et al., 2001; Wei et al., 2004; Osterberg et al., 2002).

Basically, three different types of scoring functions are applied in current protein–ligand docking tools (Gohlke & Klebe, 2002): 1) Empirical scoring functions are derived from experimental data by regression (Böhm, 1994). Here, a set of protein-ligand complexes with a known structure and experimentally determined binding affinities is exploited by regression analysis. As a result, the binding free energies are decomposed into additive energy increments that are assigned to chemical subgraphs. 2) Knowledge-based scoring functions follow the inverse Boltzmann principle. Here, a geometrical relation between two atoms is considered as energetically more favorable the more frequently it was observed in crystal structures. For this purpose, crystal structures of protein-ligand complexes are analyzed to derive distance distributions for a defined set of atom-atom interactions. The locations of the maxima are considered as optimal distances (Muegge & Martin, 1999; Gohlke et al., 2000). Force field related scoring functions directly apply the force field expressions (see Section 1.1.4.4) to assess the binding free energies (Weiner et al., 1984; MacKerell et al., 1998). Only recently, a novel type of scoring functions was proposed by Raha & Merz (2005). They report the application of semi empirical quantum mechanics to estimate electrostatic interactions and the solvation free energy during complexation. However, the performance of this scoring function in a large scale virtual screening setting (see Section 1.2) has still to be tested.

Recently, two studies were published in which docking was applied in supramolecular chemistry to identify optimally interacting host-guest systems. Both

studies demonstrated the potential and the possible impact of docking methods for the optimization of synthetic host-guest systems (de Jong et al., 2002; Corbellini et al., 2004). Details are given in Section 3.1.

1.1.4.3 De Novo Design

In contrast to docking tools, *de novo* design tools virtually construct an entirely new guest molecule, following defined rules for the virtual synthesis (Schneider & Fechner, 2005). In the case of guest molecules for protein binding sites, the guest molecule is constructed under consideration of the properties of the protein binding site. The respective algorithm aims at constructing guest molecules with high binding affinity to the protein.

De novo tools differ in the approach for generating the molecules. Some tools begin with a fixed fragment and then construct the molecule incrementally (growing strategy) (Degen & Rarey, 2006), others place fragments independently into the binding site and subsequently link these fragments together (linking strategy) (Eisen et al., 1994), and sometimes both types are combined (Böhm, 1992; Wang et al., 2000). The advantage of *de novo* tools over docking tools is that molecules can be generated that cannot be found in standard databases. Thus, potentially new and patentable structures emerge. The major drawback is the problem of synthetic feasibility. It often occurs that a generated molecule is not synthesizable (Schneider & Fechner, 2005).

Recently, the *de novo* tool ConCept was introduced which is particularly tailored to the *de novo* design of small synthetic receptors for given guest molecules (Chen & Gilson, 2007). ConCept employs a design strategy based on an evolutionary algorithm. Mainly, the method comprises two parts. In the first part, a relatively simple energy function is used to guide an evolutionary algorithm, in which a large number of candidate receptors is generated around a single rigid conformation of a targeted guest molecule. Here both, a linking and a growing strategy, are implemented for receptor construction. In the second part, the more sophisticated M2 energy model (Gilson et al., 1997; Chang & Gilson, 2003) is used to energetically assess the most promising candidates from the first part. ConCept was applied for designing water-soluble synthetic receptors for adenine. Four synthetic receptors were presented, all of which contain interaction groups found in the native β-adrenergic receptor (see Figure 1.13) However, the synthetic accessibility was not assessed.

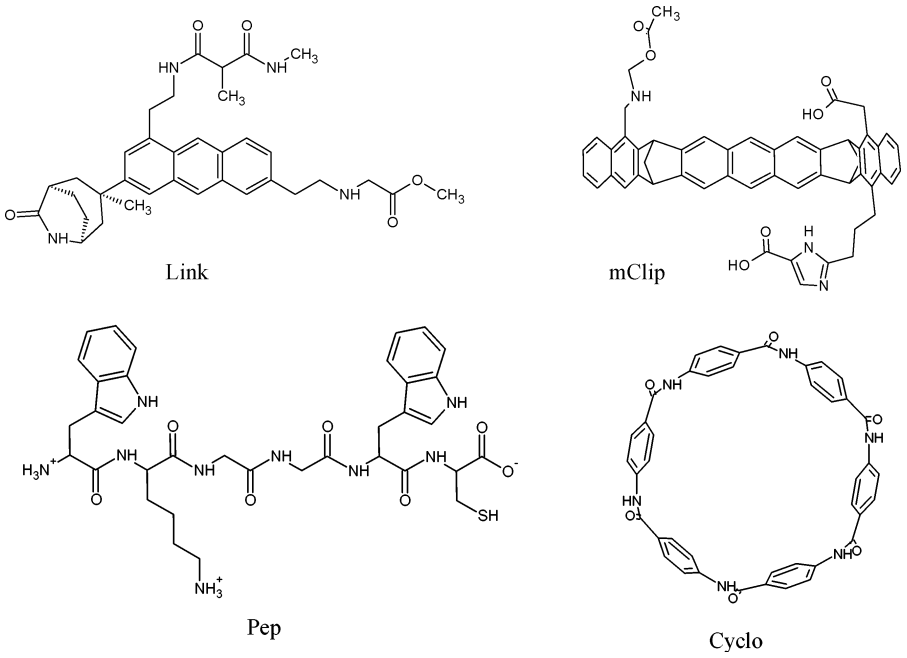


Figure 1.13. The four synthetic receptors were generated by means of the *de novo* tool ConCept for adenine as the guest molecule. From (Chen & Gilson, 2007).

1.1.4.4 Energy Minimization

During energy minimization, the energy of a molecule or a complex of molecules is calculated by means of a force field, which is an analytical function that depends on the atom coordinates and the bonds of the considered molecules. This function has a number of additive terms that represent potentials for assessing energetic contributions of pairs of bonded and non-bonded atoms. Force fields are based on classical mechanics where atoms are treated as rigid spheres and bonds as Hookian springs. The principal assumption is that molecules tend to have standardized values for bond angles and bond lengths. Force fields differ in the complexity of the underlying parameter set and the type of interactions that are considered. So-called cross terms can account for the coupling of different interactions. The parameter sets are often derived from calibrations to experimental measurements, but also from quantum chemical calculations. A frequently applied force field for the optimization of small molecules is the MMFF94 force field (Halgren, 1998, 1999a,c,b), whereas the Amber force field is widely used for proteins and nucleic acids (Weiner et al., 1984).

A force field minimization aims at finding minimum energy conformations of a molecular system. The result of a force field minimization depends largely on the choice of the minimization algorithm and the selection of an appropriate input

geometry. If the input geometry is far away from the global optimum and local minima are in between, no significant result can be expected.

1.1.4.5 Molecular Dynamics

Methods from the field of molecular dynamics are applied for simulating dynamical properties of molecular systems. Forces that act on the atoms of the system are calculated by means of a force field. The application of Newton's second law of motion allows for calculation of the changes of atom velocities from these forces and thus atom movements can be deduced.

$$\mathbf{F} = m \cdot \mathbf{a} = m \cdot \frac{\partial^2 \mathbf{r}}{\partial t^2} \tag{1.1}$$

where $\mathbf{F}$ is the force, m is the mass, $\mathbf{a}$ is the acceleration, $\mathbf{r}$ is the atom position and t is the time. A trajectory of the dynamical behavior of a molecular system is obtained by integrating discrete time-steps of the system.

Koehler et al. simulated the dynamics of α-cyclodextrin in aqueous solutions and under crystalline conditions (Koehler et al., 1988). Their study gave detailed insights into the conformational differences of both states.

1.1.4.6 Quantum Chemistry

In computational quantum chemistry, methods and principles from quantum mechanics are applied to molecules, their interactions, and reactions. The state of a system, such as a molecule, is described by means of a wave function, which is the eigenfunction of the Schrödinger equation (Schrödinger, 1926b,d,a,c). One of the central postulates of quantum chemistry states that the properties of a system can be unambiguously derived from its wave function. So-called *ab initio* methods are the most accepted methods in quantum chemistry. In these methods, only natural constants are used. Thus, chemical bonds, atom types, hybridizations of atoms, charges etc. are the direct result of *ab initio* calculations and require no further fitted parameters. However, these calculations are very time-consuming and generally only applicable for systems with at most 100 atoms. A slightly different approach is described in density functional theory. Here, the properties of a system are represented as a functional of the density (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). In principle, this approach is exact, but the forms of the functionals are unknown so that empirical or semi-empirical functionals are needed. In many cases these assumptions allow for reliable calculations of molecular or intermolecular properties in a clearly shorter range of time compared to *ab initio* methods.

Quantum chemical calculations can, for instance, provide insights into systems that might be inaccessible for experimental procedures. To provide just one example, Raub & Marian (2007) studied the strength of hydrogen bonds on a representative set of chemical subgraphs on the basis of second order Møller-Plesset perturbation theory calculations (MP2). Furthermore, they split the entire binding energy into donor and acceptor atom contributions. These were implemented into the scoring function of FlexX. As a result, the scoring function performed clearly better than the original scoring function of FlexX with respect to the correlation of predicted to experimentally determined binding energies (Raub et al., 2007) of a dataset consisting of 800 protein-ligand complexes.

1.2 Virtual Screening

In the late 1980s, much effort was spent for the development of novel experimental methods to overcome the lead structure discovery bottleneck in drug discovery (Klebe, 2006). The idea to perform large-scale automation of drug screening originated from the advent of modern techniques such as computer-controlled robots, miniaturization and highly sensitive electronic detection devices. A result was the concept of high-throughput screening (HTS) (Smith, 2002).

In a typical high-throughput screen, miniaturized biological assays automatically test hundreds of thousands of small molecular compounds for their affinity towards a particular protein target. It was anticipated that this technique could lead to an unprecedented number of novel lead structures, (Klebe, 2006); effectively, however, the truth points in the opposite direction (Bolten & DeGregorio, 2002; Lahana, 1999) and a decline of productivity in drug development can be observed (Booth & Zemmel, 2004). Several reasons for this were identified. The gain in quantity and speed has the drawback of reduced accuracy (Bajorath, 2002). Aggregation and optical absorbance of the compounds, a reactivity, or poor solubility can cause false-positive detection signals and result in low true-hit rates (Jenkins et al., 2003). Furthermore, so-called frequent hitters, molecular compounds that show up as hits in many different biological assays, considerably perturb hit identification projects (Roche et al., 2002). Additionally, the costs of HTS are relatively high. On average, the realization of one single HTS costs approximately $75,000, not including the expenses for developing an appropriate biological assay (Bajorath, 2002). Then, due to the high number of false positive hits, additional laborious verification experiments are essential that increase costs and time. Thus, the need for alternative techniques becomes evident.

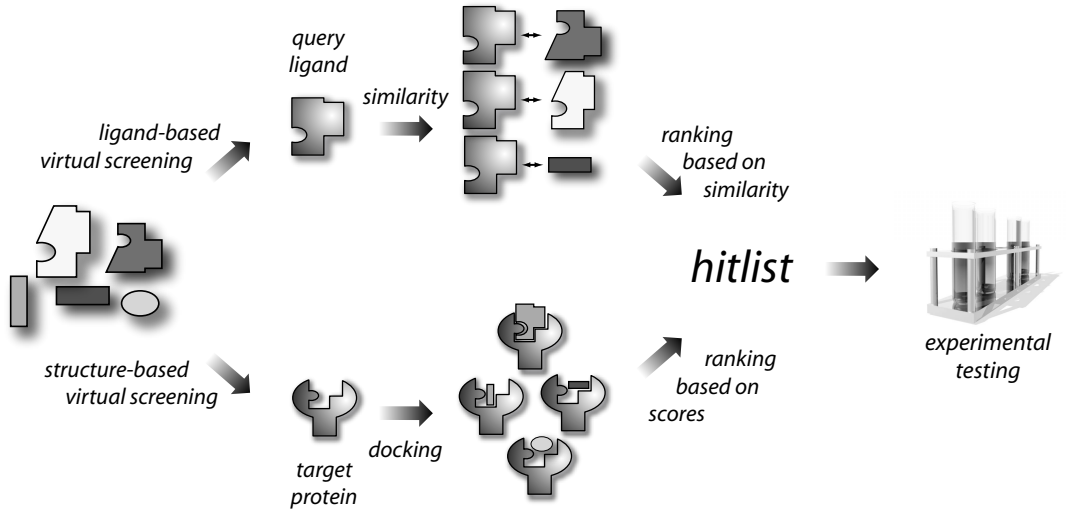


Figure 1.14. Ligand- and structure-based virtual screening. If one or more active drugs are known, a ligand-based virtual screening is possible. Structure-based virtual screening needs the structure of the protein target. In the first case the ranking is based on the similarities, whereas in docking ranking is based on the docking scores. The most promising candidates of the ranking lists are submitted to experimental testing.

Virtual screening (VS) is one such attempt and can be understood as a virtual analog to experimental high-throughput screening. The term VS encompasses a score of computational techniques, each of which aims at reducing a huge virtual library of potential drug candidates to a more manageable size (Walters & Murcko, 2002). Basically, these methods can be categorized as being either ligand based (also known as similarity based) or structure based (Lengauer et al., 2004). In both cases VS is knowledge-driven and its outcome largely depends on the amount and the quality of available data (Klebe, 2006). Whereas ligand-based VS techniques require known active compounds as a starting point (see Section 1.2.1), structure-based methods depend on the availability of three-dimensional protein structures (see Section 1.2.2). The compound libraries used for VS can comprise molecules that do not necessarily exist. Thus, the considered chemical space can be significantly larger than in HTS. Moreover the "virtual testing" does not directly consume valuable substance material (Klebe, 2006). Methods from both categories of VS techniques have produced success stories and retrieved hits that led to lead candidates for the development of novel drugs (Klebe, 2006; Kämper et al., 2007). Nevertheless, virtual screening methods often suffer from crude assumptions or approximations.

In some recent studies, direct comparisons between VS and HTS were made (Doman et al., 2002; Paiva et al., 2001; Polgár et al., 2005). In all three cases the hit rates found for the VS methods were considerably higher than for HTS. However, it is widely agreed that HTS and VS should be regarded as complementary

rather than as alternatives. In this manner both methods mutually reduce their drawbacks. In fact, VS can help to lower the number of false positives of HTS (Jenkins et al., 2003) and limit the costs by focusing the experimental testing to promising candidate molecules.

1.2.1 Ligand-Based Virtual Screening

Ligand-based VS techniques rely on the assumption that structurally similar molecules exhibit similar binding properties with respect to a given target (Martin et al., 2002). The methods depends on the availability of one or more compounds (query structures) that are known to bind to the considered target protein. For all compounds of a database, similarities to each of the query structures are calculated. From these similarities ranking lists are derived where the most similar structures are supposed to be the most promising candidates. Different approaches for describing similarity between molecules have been proposed (Sheridan & Kerarley, 2002; Lengauer et al., 2004): some rely on the comparison of molecular graph topology (Rarey & Stahl, 2001) (see Section 4.1.2), other tools compare the three-dimensional shape of molecules (Grant et al., 1996; Lemmen et al., 1998a) (see Section 4.1.3), and some generate so-called molecular fingerprints (see Section 4.1.1). In general, ligand-based methods are significantly faster than structure-based screening methods. Today, ligand-based VS is a standard tool for the identification of new drugs for a given protein target (Kämper et al., 2007).

1.2.2 Structure-Based Virtual Screening

Structure-based VS exploits the three-dimensional structure of the target protein. Commonly, docking methods are applied that generate complexes for all molecules of a defined virtual database and the binding site of the considered target protein (see Section 1.1.4.2). Ranking lists are derived from the estimated scores. The compounds with the best scores are supposed to be the most promising candidates for experimental verification. Aside from docking, structure-based ligand design also comprises, for example, structure-based pharmacophore searches. Here, a pharmacophore is derived from so-called interaction hot spots in the binding site. Pharmacophore tools determine for each compound of a database whether the compound fits into the pharmacophore model of the binding site [unity, www.tripos.com]. Until recently, about 50 protein targets were used for structure-based virtual screenings (Klebe, 2006). Considerable success studies were reported in both academia and industry (Bajorath, 2002; Klebe, 2006).

1.3 Goals and Outline of this Thesis

The aim of this thesis was to develop and to validate computational techniques that assist experimentalists in the development of novel host-guest systems. A major focus thereby was to transfer the concept of virtual screening to supramolecular chemistry.

In Part I (Chapter 2) I will describe the development of a tool for the fast and reliable structure prediction of hydrogen bond-based synthetic host-guest complexes. The tool is based on the algorithms and data structures of the efficient protein-ligand docking tool FlexX (Rarey et al., 1996a). FlexX considers molecular flexibility only for the ligand and treats the receptor as rigid. However, for a number of synthetic receptors, this assumption cannot be carried over, since they can exhibit a high degree of flexibility. Thus, novel algorithms were developed that account for the molecular flexibility of both complex partners. The algorithms were integrated into the tool FlexR. FlexR was tested on a set of experimentally derived crystal structures of synthetic host-guest complexes regarding its ability to generate near-native complex structures. Beyond this, the tool was designed to be efficient enough to allow for virtual screenings.

Part II of this thesis (Chapters 3, 4, and 5) concentrates on β-cyclodextrin and derivatives as host molecules. In Chapter 3, I describe a novel protocol for the identification of tailored synthetic receptors for a given guest molecule based on molecular docking. We chose camptothecin as a guest molecule and generated a library of β-cyclodextrin based synthetic receptors that were sequentially docked onto camptothecin. Selected receptors with promising properties regarding the complexation of camptothecin were experimentally verified. The major driving force of the complex formation between the β-cyclodextrin host and a guest molecule are hydrophobic interactions (Wenz, 1994; Connors, 1997; Rekharsky & Inoue, 1998). Since FlexR was particularly tailored towards hydrogen bond based host-guest complexes, it could not be applied for β-cyclodextrin inclusion complexes. Hence, two docking tools, AutoDock and GlamDock both with a known capability of handling hydrophobic interactions were used. Chapter 4 focuses on a virtual screening technique that uses information solely from guest molecules. The method combines a ligand-based virtual screening technique with a quantitative structure property relationship (QSPR) model that predicts the binding free energy of the complex between the guest molecule to the synthetic host. I applied this method for identifying novel guest molecules for β-cyclodextrin. In Chapter 5 I evaluate three different regression methods, namely principal component regression (PCR), partial least squares regression (PLSR)

and support vector machine regression (SVMR) regarding their applicability for the prediction of thermodynamical properties of complexes between various guest molecules and β-cyclodextrin as the host. I identify reasons for differences in the predictability of thermodynamical quantities.

Part I

Development of a Virtual Screening Tool for Synthetic Host-Guest Complexes

2

Flexible Docking of Guest Molecules into Synthetic Receptors Using a Two-Sided Incremental Construction Algorithm

This chapter describes a novel computational method which tackles the problem of docking flexible guest molecules into flexible synthetic hosts. The developed algorithms were implemented into the tool FlexR (Steffen et al., 2006). The conformational sampling of the guest and the host molecule is handled by applying a new approach: the conformational spaces of both molecules are explored simultaneously using information from the respective counter-molecule.

In Section 2.2 we introduce the algorithms of our method. We validate the method by means of a test set, consisting of crystallographically determined host-guest complexes in Section 2.3. Furthermore an application of the tool was performed, in which we identified potentially competing guest molecules of the selective creatinine receptor of Bell et al. (1995) from a database of molecular compounds (see Section 2.4).

2.1 Introduction

During the development of a synthetic receptor, one of the major goals is to achieve a high binding affinity to a considered guest molecule. In some cases it is even more important to secure high selectivity of the receptor to this particular guest. Thus, additional efforts often have to be expended until the receptor is specific enough to discriminate between different guest molecules. Due to the industrial relevance of synthetic receptors (see Section 1.1), the need for methods that allow for faster, more rational development becomes apparent. This need mirrors a recent trend in medicinal chemistry. In earlier times the discovery of drugs was often a question of extensive work and sometimes luck (Kubinyi, 1999). Nowadays, a wide range of diverse methods and tools are available that assist in the fast and rational development of novel drugs. One of the methods which a medicinal chemist has at his disposal employs so-called docking tools (see Section 1.1.4.2 and 1.2.2).

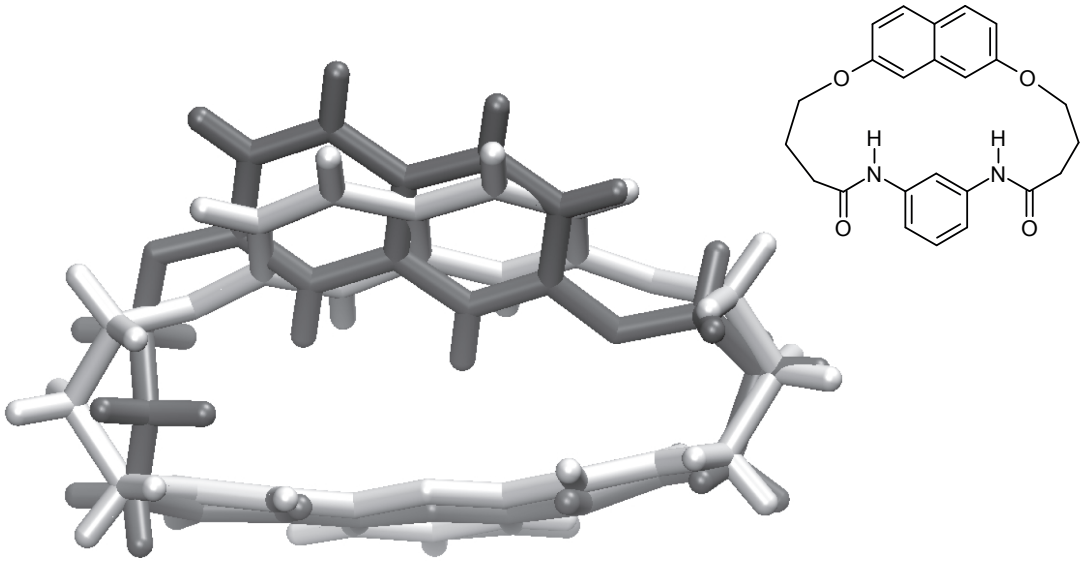


Figure 2.1. Two conformations of a synthetic receptor in presence and absence of a guest molecule and the structural formula. The conformation shown in dark grey represents the holo-form. The apo-form is shown in white. The significant conformational change is induced by the guest molecule. Adopted from Kämper et al. (2006).

Current state-of-the-art tools for docking are all able to handle the flexibility of the ligand. The efficient and reliable modeling of the protein's flexibility, however, still remains a challenging task (see Section 2.1.1). Despite the fairly rough assumption of a rigid receptor, numerous success stories have been published in which docking tools have led to novel drug candidates (Kämper et al., 2007). This supports the applicability of the docking approach for the rational design of novel drug molecules and thus makes docking to be an interesting candidate as a method for the rational design of synthetic host-guest complexes. However, for a number of synthetic receptors, the assumption of a rigid receptor cannot be applied, since these molecules can exhibit a high degree of flexibility similar to the guest molecules (Otto, 2006). In many cases not only the guest molecule but also the synthetic receptor adapts its conformation during complex formation. This process is often referred to as induced fit (Hamilton & van Engen, 1987). As an example, consider the synthetic receptor for nucleotides of Hamilton & van Engen shown in Figure 2.1 (Hamilton & van Engen, 1987). Here, two experimentally observed conformations of the receptor are superimposed to each other. The dark grey representation shows the apo-form (non-complexed) (CSD-ID: FODTEX), the light grey conformation is taken from the crystal structure of the complex to a guest molecule (CSD-ID: FODTIB). A clear structural change induced by the guest molecule can be seen in the naphthalene moiety, which undergoes a sig-

nificant turn during complex formation. In the complexed state (holo-form) the naphthalene part is oriented approximately parallel to the pyridine plane (161.6°), whereas in the uncomplexed state (apo-form) the angle between the two planes changes to 127.5°. The conformational changes that a synthetic receptor can undergo in presence of a guest molecule are huge. The example shows that the conformation seen in a particular complex is only one of many possibilities the host can adopt with only a small difference in conformational energy.

In the following text I describe computational tools that are in some respects related to the method presented in this chapter. In Section 2.1.1 I detail attempts to implement receptor flexibility into protein-ligand docking tools. Section 2.1.2 summarizes recent computational tools that are particularly tailored for the structure prediction of synthetic host-guest complexes.

2.1.1 Modeling of Receptor Flexibility in Protein-Ligand Docking

The integration of receptor flexibility into protein-ligand docking tools is the subject of current research (Carlson, 2002; Brooijmans & Kuntz, 2003; Murray et al., 1999). The analysis of experimentally determined protein-ligand complexes of the same protein but different ligands reveals distinct types of flexibility that can be observed on the protein side when different ligands are bound (Krebs et al., 2003; Gerstein & Krebs, 1998). The flexibility of proteins can, for example, consist of local side-chain rotations, smaller adaptations of single loops or large movements of complete domains. The complexity of the protein's flexibility and the concurrent need for efficiency of docking tools require the use of simplifications in the implementation. In relevant literature, approaches were proposed that tackle the problem of receptor flexibility in protein-ligand docking with different assumptions. Representative approaches are detailed in the following sections.

Use of a Soft Scoring Function

One of the simplest ways to implement a limited receptor flexibility is to soften the criterion for the steric fit of a ligand and a receptor. In this manner, overlaps of the van-der-Waals radii of two atoms are penalized less (Jiang & Kim, 1991; Ferrari et al., 2004). In a retrospective virtual screening scenario Ferrari et al. (2004) studied the performance of the standard version of DOCK (Lorber & Shoichet, 1998; Shoichet et al., 1999) in comparison with a modified version employing a soft scoring function. The soft scoring function was obtained by diminishing the repulsive term in the Lennard-Jones potential, allowing for close contacts between the ligand and the protein.[1] Compared to the original scoring

[1] Ferrari et al. replaced the original Lennard-Jones 12-6 potential with a 9-6 potential

function, the soft scoring approach performed significantly better with respect to the enrichment of known active compounds on top ranks of a derived ranking list. This is due to the fact that larger active ligands could not be docked into the binding site because of steric clashes that are more strongly penalized when the original scoring function was used. However, the flexibility represented by this approach is reduced and therefore only slight conformational adaptations are considered. Thus, the outcome of a "soft docking" largely depends on the degree of flexibility which a considered protein can undergo.

Integration of Side-Chain Flexibility

Some protein-ligand docking tools integrate the handling of side-chain flexibility. The docking tool GOLD presented by Jones et al. allows for the rotation of hydroxyl groups within the binding site (Jones et al., 1995, 1997). Leach presented an algorithm based on the A^*-algorithm with dead-end-elimination (Leach, 1994). For a given orientation of a ligand, i. e. the translation and rotation, the algorithm finds the combination of the side-chain and the ligand conformations with the lowest energy with respect to the energy function. Possible conformations for a side-chain are represented by discrete low-energy states from a rotamer library. The latest version (V. 4.0) of AUTODOCK (http://autodock.scripps.edu/, 2007) integrates the optimization of side-chain torsion angles into the genetic algorithm which optimized the ligand conformation in previous versions. However, the number of considered torsions is limited due to the increasing complexity of the conformational search space.

Integration of Hinge Movements

Sandak et al. presented an approach to handle the movement along hinges within proteins (Sandak et al., 1998), in which hinges have to be manually assigned either to the ligand or to the protein. Due to the limited number of allowed hinges, the approach is of reduced applicability because alternative side-chain conformations are not considered, for example.

Use of Multiple Protein Structures

Another approach for the consideration of receptor flexibility in docking is to use multiple conformations of a protein, so-called ensembles. These ensembles of protein conformations can be derived from available crystal structures, NMR structures, molecular dynamic simulations, or homology models. In principle, each docking program can be used to perform so-called cross-dockings (Kramer et al., 1999; Osterberg et al., 2002; Murray et al., 1999). In cross-docking, the ligands are

subsequently docked into all available conformations of the protein binding site. This technique is rather time-consuming since the docking time increases with each considered protein. Furthermore, only the available conformations of a protein are considered and no new conformations are generated. To circumvent these drawbacks concepts for combining multiple protein structures were recently introduced: Osterberg et al. (2002) attempted to combine the generated energy grids of AutoDock for different superimposed protein conformations. They studied four methods of combining multiple target structures within a single grid. The combined grids derived by energy-based averaging turned out to be the best with respect to docking and energy prediction accuracy. The problem of this method is that the ligand cannot differentiate between two different states of a side chain and possibly, interactions with both can occur simultaneously. Claussen et al. (2001) introduced the computationally more advanced technique FlexE, which is an extension for FlexX. In FlexE, superimposed protein structures are combined to an ensemble representation. Similar parts (instances) of the integrated proteins are clustered together whereas dissimilar parts remain as alternatives. During the docking of a ligand, compatible instances of the protein are selected to interact with the ligand. A so-called incompatibility graph evaluates whether selected instances of the protein ensemble are compatible with each other. FlexE is significantly faster than parallel docking into the integrated protein conformations. Furthermore, the implemented algorithm allows for the recombination of protein parts to entirely new but still plausible conformations. However, a recent study showed that in a virtual screening setting the way receptor flexibility is modeled in FlexE has the drawback of reduced specificity. As a result, the performance in a retrospective virtual screening study was worse than with the standard FlexX docking (Polgár & Keserü, 2006).

Docking Based on Molecular Dynamics Simulation

Simulation methods, such as molecular dynamics simulation (see Section 1.1.4.5), can be used to predict protein-ligand interactions while at the same time considering flexibility for both molecules. Molecular dynamics simulations are generally very time-consuming. Today, protein trajectories of only 10 to 100 nanoseconds can be simulated. Simulating the diffusion of a ligand into a protein's binding site would require significantly more time. Hence, the high computational costs of the method require the introduction of simplifications. Mangoni et al. (1999) introduced a simulation method that tackles the docking problem. They applied a modified temperature coupling scheme to the ligand. Therein, a very high temperature is used for the translational moves, whereas the internal degrees of freedom

are coupled with a regularly low temperature (300 K). Protein flexibility was sampled only for binding site atoms by using the low temperature. As a test case, phosphocholine was docked into the immunoglobulin McPC603 in the presence of water molecules. In comparison to the crystal structure of the complex of these molecules, a near-native solution was found. The computation time required for a 100 ps simulation remained in the range of hours. Thus, an application of this method in a virtual screening scenario appears inappropriate at the moment.

2.1.2 Structure Prediction Tools for Synthetic Host-Guest Complexes

In general, the chemical building blocks of synthetic receptors are clearly more diverse than those of proteins. The problem of receptor flexibility can thus not be handled with the same assumptions as those made for the proteins. In respective literature, some approaches have been proposed that tackle the problem of structure prediction for synthetic host-guest complexes. Examples are detailed below.

Corina

Corina is today known as one of the leading fast structure-generation tools for druglike molecules (Sadowski & Gasteiger, 1993). A rather unknown feature of Corina is the ability to predict the structure of synthetic host-guest complexes based on cyclic or polycyclic synthetic hosts (Sadowski et al., 1992). The structure generation works as follows: two-dimensional connection tables with annotated stereochemistry are required as input for both molecules of the complex. Corina starts with the generation of a reasonable low-energy conformation of the guest molecule. It assigns standard values to all bond lengths and angles. Torsion angles along rotatable bonds are set to preferred low-energy states. Corina detects ring systems and assigns a standard conformation taken from a comprehensive database of ring systems, to each of them. If atom-atom overlaps exist within the generated conformation, they are removed by systematically modifying rotatable torsion angles until no more overlaps exists. The essentially cyclic or polycyclic host molecule is processed in a different manner. Initially, Corina identifies a so-called superstructure of the host molecule. This superstructure represents the topology of the molecule and is scaled according to the number and types of bonds within the host. Then, Corina constructs the structure of the host molecule along the edges of the superstructure (see Figure 2.1.2). A reduced force field serves for a preliminary structure optimization. The complex structure is obtained by putting the center of mass of the guest molecule onto the center of mass of the host. Subsequently, Corina rotates the guest molecule in steps of 120° around

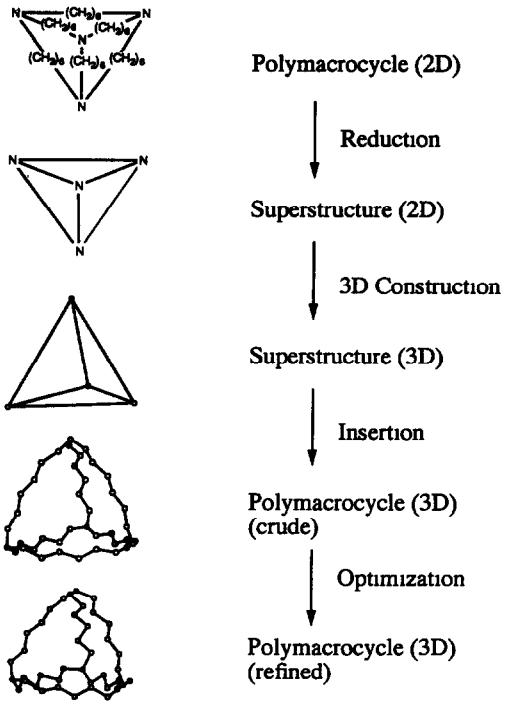


Figure 2.2. The superstructure of the cryptand is a pyramid. The molecule is constructed along the edges of this pyramid. A structure optimization is the performed by means of a reduced force field. From Sadowski et al. (1992).

each the x-, y- and z-axes, resulting in 27 different positions. Each of these initial configurations is submitted to a coarse orientation optimization, in which the conformations of both molecules are kept rigid. Of all resulting structures, the complete force field finally optimizes the one with the lowest energy by means of a full geometry optimization. A SUN SPARC Station IPC CPU required 51 minutes for the computation of the complex structure shown in Figure 2.3. The approach has a major problem: it is only applicable for symmetric cyclic host molecules.

Tork

Tork predicts low-energy conformations of single organic molecules as well as bimolecular host-guest complexes. It is based on a normal-mode analysis in bond-angle-torsion coordinates. It focuses on a key subset of torsional coordinates to identify natural molecular motions that lead the initial conformation to new energy minima. New conformations are generated via distortion along these modes and their paired combinations, followed by an energy minimization. For complexes, special treatment is accorded to the six coordinates that specify the position and orientation of one molecule relative to the other. Tork is efficient in

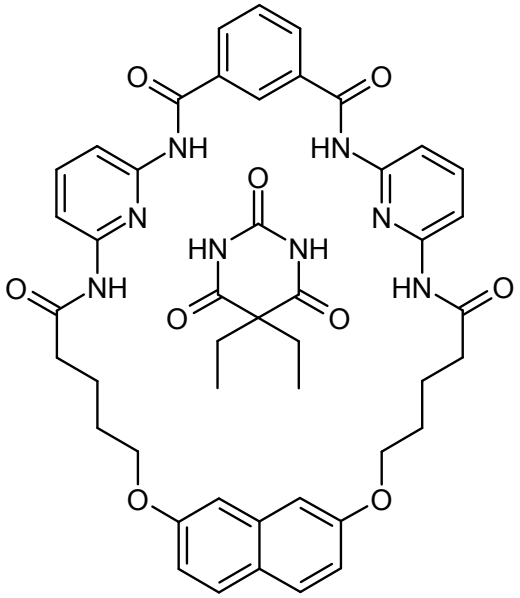


Figure 2.3. The complex between diethyl barbiturate and a synthetic barbiturate receptor. CORINA predicted the complex structure within about 51 minutes.

the prediction of single molecule conformations. The generation of bimolecular complexes is comparably slow. On a current state-of-the art PC (CPU: Pentium III 733-MHz), the structure prediction for a complex between a cyclophane host and menthol took approximately ten hours.

MOMO

The Egert group has developed the software suite MOMO which provides functionality for the conformational optimization of chemical structures (Gemmel et al., 1999). Therein, the SUPRA-module (Söntgen, 2003) focuses on the structure prediction of complexes between two molecules and functions as follows. First, conformations are generated for both molecules independently. This is either done stochastically or systematically by generating all possible discrete conformers of a molecule, in which the torsion angles are systematically varied by rotation in defined step size. Each of the derived conformations is energetically assessed by means of the MOMO force field. The energetically most favorable conformer is selected for each molecule. Minimal enclosing cuboids are constructed around the selected conformers. Subsequently, initial complexes are generated. This is done by testing all possible configurations on a virtual grid with adjustable step size in which the two cuboids are in direct contact with each other. Finally, a complex force field minimization is performed. The complexes with the lowest energies are presented as the solution set. MOMO is apparently inefficient. The generation of

the complex structure between the two rigid molecules cytosine and guanine took 1 hour and 20 minutes when default values were applied (CPU: Intel Pentium III, 350 MHz).

FlexR *V 1.0*

The first version of FlexR was developed by Kämper et al. (2006) It transfers the concept of the protein-ligand docking tool FlexX (Rarey et al., 1996a) to synthetic host-guest complexes (Kämper et al., 2006). FlexX is extended in the sense that receptor flexibility is tackled. A preparatory step generates a set of representative conformations of the synthetic receptor. In order to diminish the complexity of the conformational sampling, the number of torsion angles are reduced to sets of preferred values. Those are derived from the Mimumba database (Klebe & Mietzner, 1994). The conformational space of ring systems within the molecule is sampled by means of Corina (Sadowski & Gasteiger, 1993). Conformers with intramolecular short contacts are discarded. The complex structure is obtained by subsequently docking the guest molecule into each of the conformers of the host molecule. At this point, the standard FlexX algorithms are applied (see Section A.1). In principle, the roles of the synthetic host and the guest molecules can be exchanged. In fact, two docking directions are allowed, namely forward and inverse docking. In forward docking the guest molecule is flexibly docked into the generated receptor conformations, whereas in inverse docking the synthetic receptor is sequentially docked around the conformations of the guest molecule. This approach was successfully validated on a set of experimentally determined complex structures which could be reproduced in a similar manner. However, in cases in which both molecules of the complex exhibited a high degree of flexibility the docking time increased dramatically. Nevertheless this study proved the transferability of the FlexX concepts and hence served as a profound basis for the work presented in this chapter.

2.2 Methodology

The new docking strategy for synthetic receptors extends our previously described method (Kämper et al., 2006). The overall principle of the algorithm relies on the incremental construction of the complex from fragments (Rarey et al., 1996a). To model the molecular flexibility a discrete set of preferred torsional angles is used for each rotatable bond (Klebe & Mietzner, 1994). These torsional angles have been derived from a statistical analysis of the Cambridge Structural Database (CSD) (Allen, 2002). In contrast to the first version of FlexR (Kämper et al.,

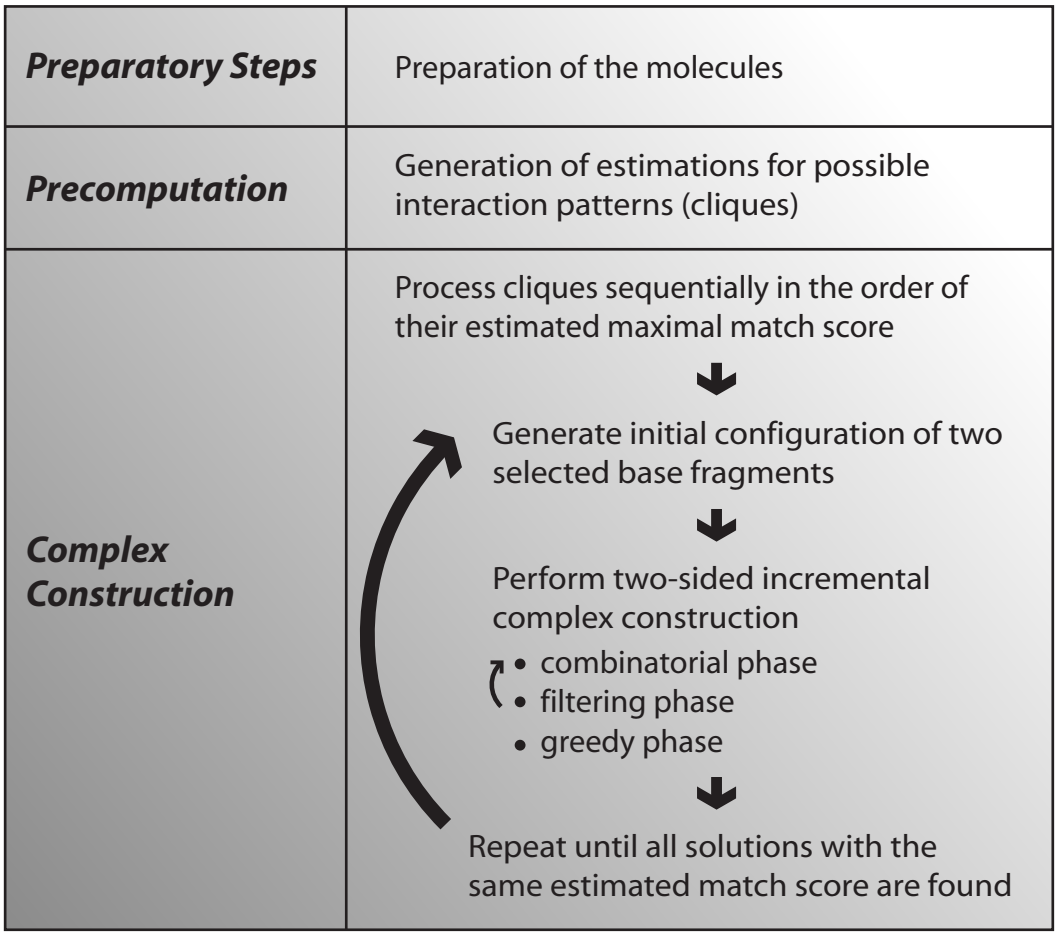


Figure 2.4. Flow chart of the algorithm of FlexR.

2006) the structures of both molecules - the guest and the host - are built up incrementally during docking time. Since the conformations of both molecules are unknown initially, we cannot use the conformation of one molecule to direct the generation of the conformation of the other molecule. Due to this fact a precomputation phase is introduced that determines putative interaction patterns between the two interacting molecules. These interaction patterns direct the subsequent phase of the complex construction. A schematic flow diagram of the algorithm is shown in Figure 2.4 and summarized below.

2.2.1 Details of the Chemical Model

The implementation of FlexR is based on the chemical model of FlexX (see Section A.1). The major principles of the chemical modeling are detailed below.

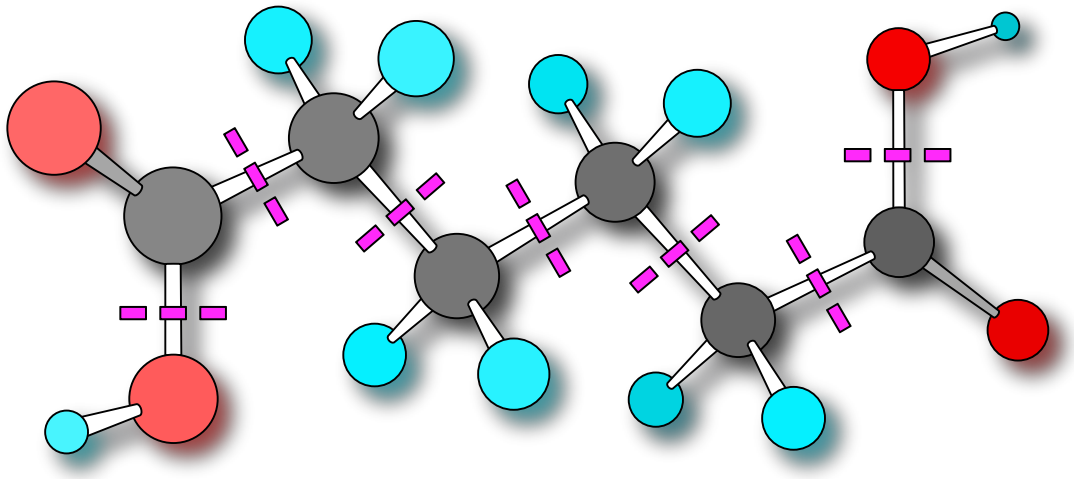


Figure 2.5. The molecules are cut at each acyclic single bond between non-hydrogen atoms (cutting is denoted by pink dashed lines). In this way fragments are obtained.

2.2.1.1 Fragmentation Principle

Initially, the molecules are severed at each acyclic single bond between non-hydrogen atoms and thus molecular fragments are obtained (see Figure 2.5). Torsion angles at double bonds, bond lengths and bond angles are taken from the input structure. The conformations of small ring systems up to a ring size of ten atoms are computed with the program CORINA (Sadowski & Gasteiger, 1993)

2.2.1.2 Interaction Model

The interaction model used in FLEXX and FLEXR has been adapted from LUDI (Böhm, 1992). Each interacting group in the molecules is described by an interaction center and interaction surface (see Figure 2.6). Depending on the type and the neighborhood of an interacting group an appropriate interaction surface and an interaction center is assigned (see Table 2.1). Two interaction groups of different molecules enter into an interaction if their interaction types are compatible, e. g. an hydrogen bond donor and an hydrogen bond acceptor, and their surfaces mutually lie on the center of the counter group (see Figure 2.7). For computational reasons the surfaces are represented by discrete point sets (see Figure 2.8) (Rarey et al., 1996a).

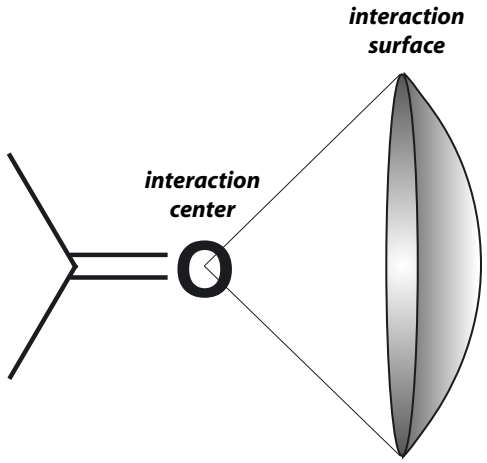


Figure 2.6. An interaction is represented by an interaction center and an interaction surface.

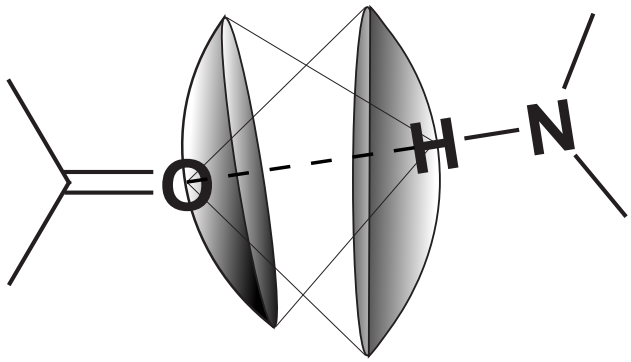


Figure 2.7. The FLEXX interaction scheme. Two interaction groups enter into an interaction if their surfaces mutually lie on the center of the counter group.

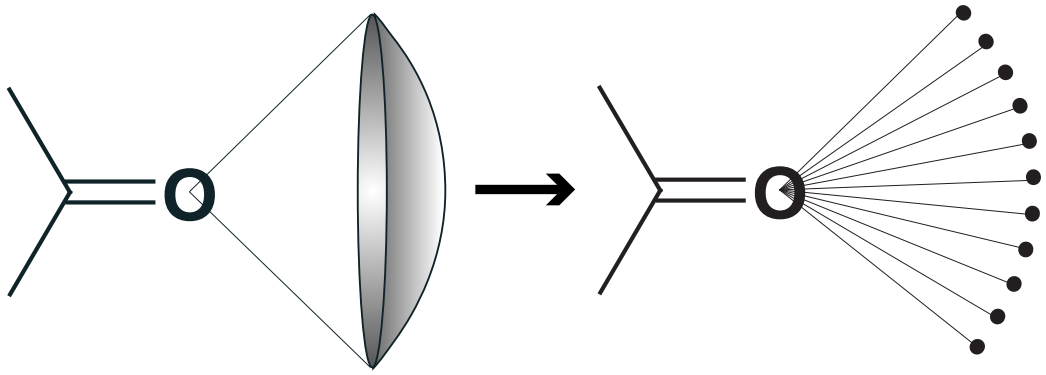
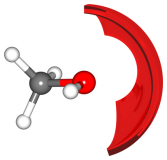
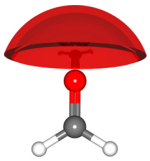
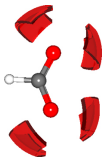
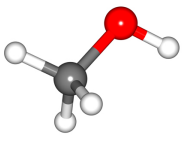
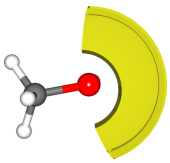
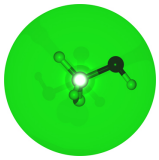
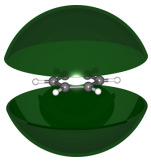


Figure 2.8. The interaction surfaces are represented by discrete point sets.

Table 2.1. The interaction geometries of FlexX and FlexR. The figures are provided by Stephan Raub.

	interaction type	charge	ΔG in kJ mol^{-1}	r_0 in Å
	H-bond acceptor	neutral ionic	−2.35 −4.15	1.9
	sp^2-H-bond-acceptor	neutral ionic	−2.35 −4.15	1.9
	COO^--H-bond-acceptor	ionic	−4.7	1.9
	H-bond donor	neutral ionic	−2.35 −4.15	1.9
	metal-acceptor		−2.35	2.0
	lipophilic contact	neutral	−0.35	4.0 – 4.8
	aromatic interaction	neutral	−0.35	4.5

2.2.1.3 Scoring Function

A scoring function based on the work of Böhm is used for fast energy evaluation throughout the algorithm (Böhm, 1994; Rarey et al., 1996a).

$$\begin{aligned}\Delta G = \Delta G_0 &+ \Delta G_{rot} N_{rot} \\ &+ \Delta G_{hb} \sum_{HBonds} f\left(\Delta r, \Delta\alpha\right) \\ &+ \Delta G_{io} \sum_{ion.int.} f(\Delta r, \Delta\alpha) \\ &+ \Delta G_{aro} \sum_{arom.int.} f(\Delta r, \Delta\alpha) \\ &+ \Delta G_{lipo} \sum_{lipo} f^*(\Delta r) \end{aligned} \tag{2.1}$$

Six terms are included that describe neutral hydrogen bonds (ΔG_{hb}), ionic interactions (ΔG_{io}), aromatic interactions (ΔG_{aro}), lipophilic contributions (ΔG_{lipo}) and entropic costs ($\Delta G_{rot} * N$). Penalty functions (f and f^*) are used that penalize deviations from ideal interaction distances (Δr) and for directional interactions also deviations from ideal angles ($\Delta\alpha$) (for details see Rarey et al. (1996a)). In this scoring scheme interactions are considered as being independent from each other, such that the scores of the single interactions within a complex are summed up to obtain the score of the entire complex. This allows for scoring partial solutions.

2.2.2 Algorithmical Details of the Structure Generation

The complex construction phase consists of two main steps that are sequentially traversed:

- Precomputation
- Construction of the complex structure

2.2.2.1 Precomputation Phase

In the precomputation phase we collect putative interactions between host and guest molecule which can be realized simultaneously. The search is performed by first identifying putative interaction pairs between the molecules, then generating a docking graph which comprises all identified putative interaction pairs, and finally executing a clique search for finding maximal sets of interactions that can be realized simultaneously.

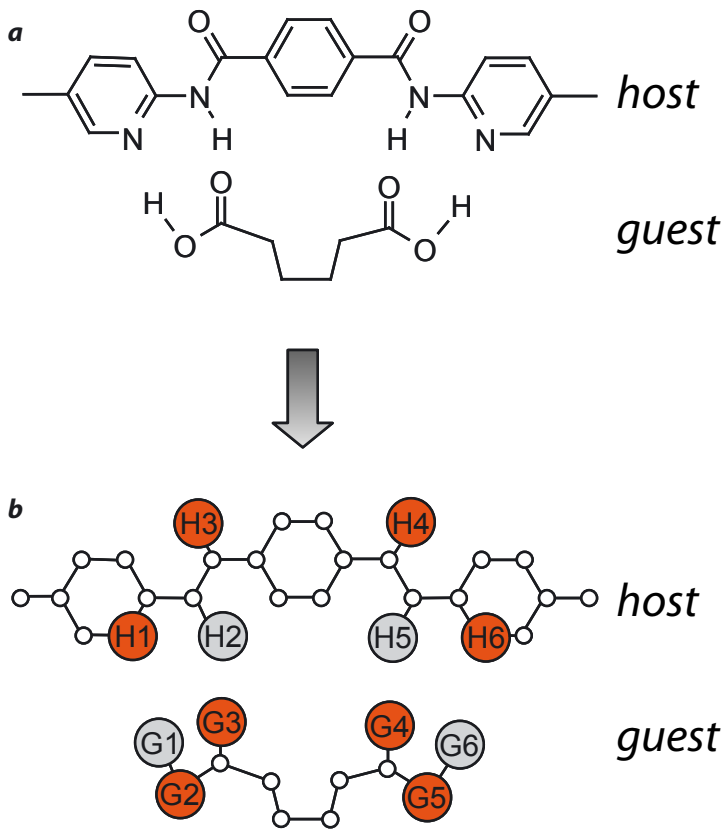


Figure 2.9. (a) Structural formula of the host-guest complex by Garcia-Tellado et al. (1990). Only polar hydrogen atoms are shown for clarity. (b) Molecular graph representation of the complex. Nodes (circles) represent atoms; edges represent chemical bonds between two atoms. Centers of directional short-range interactions are labeled with unique identifiers. Atoms that do not exhibit directional interactions are depicted in white. Hydrogen atoms at hydrogen donor sites and hydrogen-bond acceptors are shown in gray and red, respectively. Apolar hydrogen atoms are not shown for clarity.

Generation of the Docking Graph

To illustrate the generation of the docking graph consider the complex shown in Figure 2.9. From the molecular structures, first molecular graphs are derived in which atoms are represented by nodes and edges denote covalent bonds between atoms (see Figure 2.9 b). We define centers of directional short-range interactions as atoms that can form hydrogen bonds or salt bridges. In the case of hydrogen-donor interaction groups we regard the hydrogen atoms themselves as interaction centers. Center of directional short-range interactions in the two molecules are identified, labeled and colored according to their functionality (see Figure 2.9b).

From the molecular graphs of the host and the guest a docking graph is derived (see Figure 2.10) similarly to the approach used by the program in DOCK (Kuntz et al., 1982). A node represents a possible directional short-range inter-

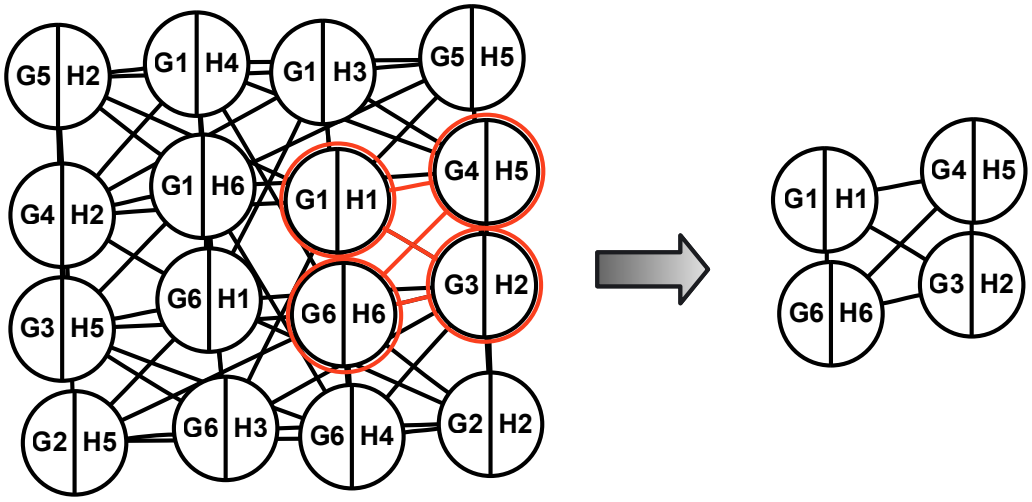


Figure 2.10. Generated docking graph of the complex by Garcia-Tellado et al. (1990) (left side). Nodes are generated for each possible interaction pair. Two nodes share an edge if they are compatible, i.e., the corresponding interactions can be realized simultaneously. The clique exemplifies one possible complex interaction pattern (right side).

action that can be formed between an atom of the host molecule and an atom of the guest molecule. The node is labeled with the identifiers of both atoms (see Figure 2.10). Edges between two nodes are inserted only, if two interactions can be realized simultaneously. Two interactions can be realized simultaneously, if the two corresponding interaction centers in the receptor are at about the same distance as the two corresponding interaction surfaces in the guest molecule. In this case, the two nodes representing both interactions are connected with an edge in the docking graph (see Figure 2.11). The respective distance property is checked by computing bounds on the maximal and minimal distances that can occur between two centers of directional short-range interactions and their corresponding interaction surfaces, respectively, within the conformational space of the molecules (see Paragraph *Distance Range Estimation*). The resulting graph is then submitted to a maximal clique search using the Bron-Kerbosch algorithm (Bron & Kerbosch, 1973). A clique is a subset of nodes (V') within an undirected graph ($G = (V, E)$), in which all nodes are connected to every other node of the subset V'. A maximal clique is a clique to which no more nodes of the graph can be added, such that the first proposition is still valid. For our work a clique represents sets of interactions that can be realized simultaneously within the same complex structure (see Figure 2.10, right side). However, it is important to note that the generated cliques have to be further validated in the complex construction step because, at this point, only simple distance constraints are con-

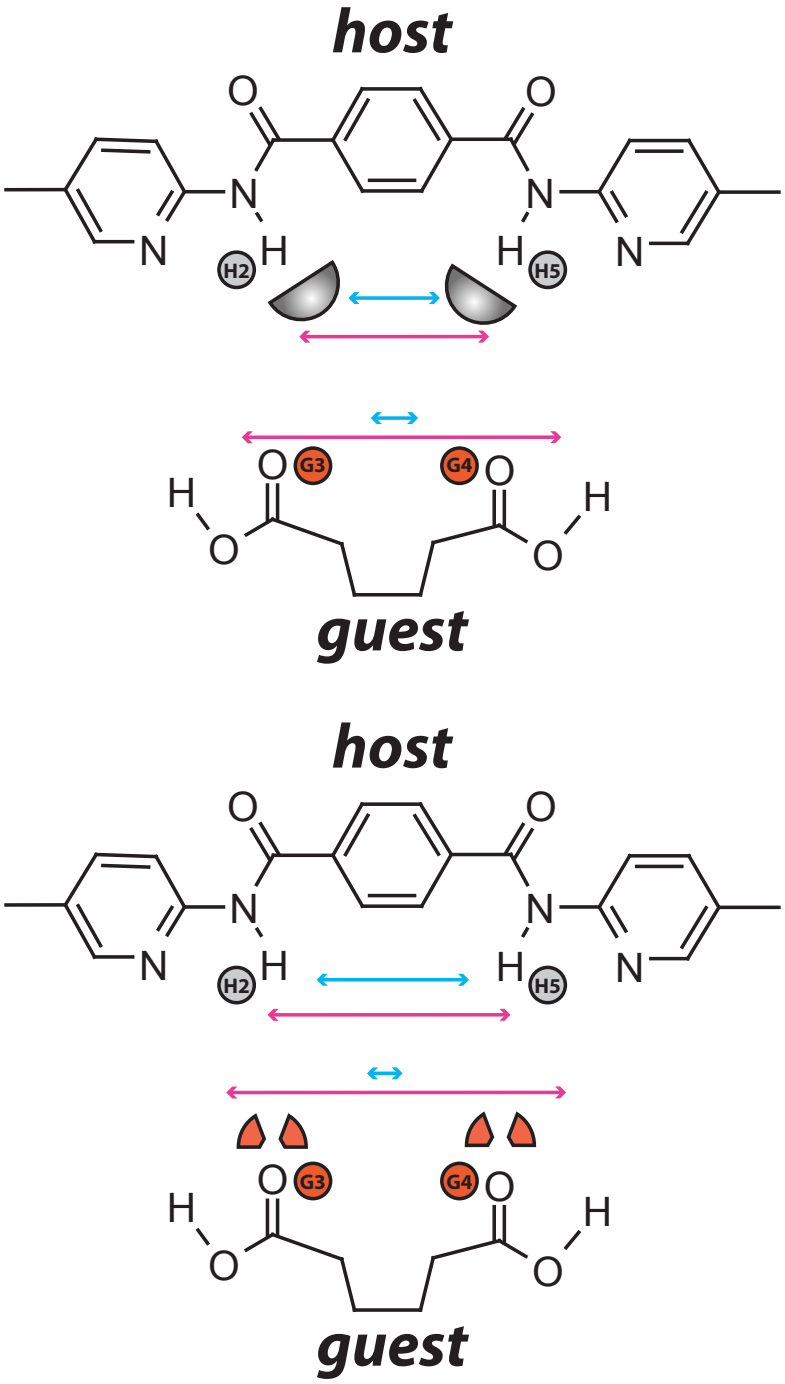


Figure 2.11. The distance range between the interaction surfaces of the host atoms H2 and H5 exhibits an overlap with the distance range between the interaction centers G3 and G4 of the guest molecule (left). The same applies for the interaction centers H2 and H5 of the host molecule and the interaction surfaces G3 and G4 of the guest molecule (right). This results in an edge between the corresponding nodes in the docking graph.

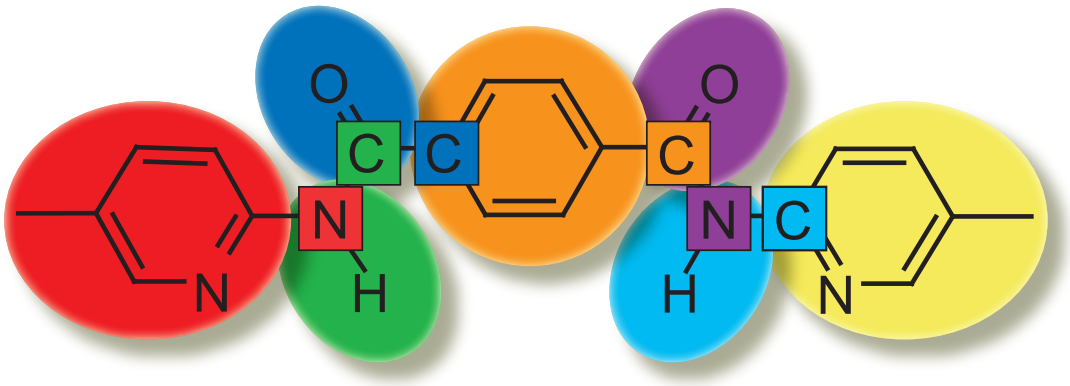


Figure 2.12. Given a fragmentation (ellipsoids) and a base fragment (red) all outgoing atoms (squares) are defined. An outgoing atom is placed with the fragment of the same color.

sidered. Additional constraints, such as the exclusion of possible atom overlaps are not taken into account. In cases in which the molecules exhibit a high degree of flexibility degenerate cliques may occur. A degenerate clique is a clique comprising a subclique of interactions that can be realized simultaneously plus several interactions that cannot be realized. In order to deal with these cases not only the maximal cliques that are contained in the docking graph but also their subcliques are added to the clique list. Whether the generated cliques can yield a valid complex structure is assessed in the complex construction step (see Section 2.2.2.2).

So far, in docking algorithms based on the incremental construction principle (Rarey et al., 1996a; Kämper et al., 2006), only one molecule is constructed incrementally. The alternative docking strategy presented here builds up both molecules incrementally, guided by the calculated cliques.

Distance Range Estimation

Throughout the method intramolecular distances are used. For the generation of the docking graph, distance ranges (i.e. the minimal and maximal distances) between all pairs of centers of directional short-range interactions have to be computed, as well as distance ranges between the corresponding interaction surfaces. The complex construction phase requires distance information as a basis for filtering out inappropriate solutions. Here, we compute the distance range between the so-called outgoing atoms and the next targeted center of a directional short-range interaction. An outgoing atom of a fragment is an atom that is placed together with the fragment but actually belongs to an adjacent fragment in the fragment tree (see Figure 2.12). These outgoing atoms are used as anchor points for the adjustment of the appropriate torsion angles of a child fragment.

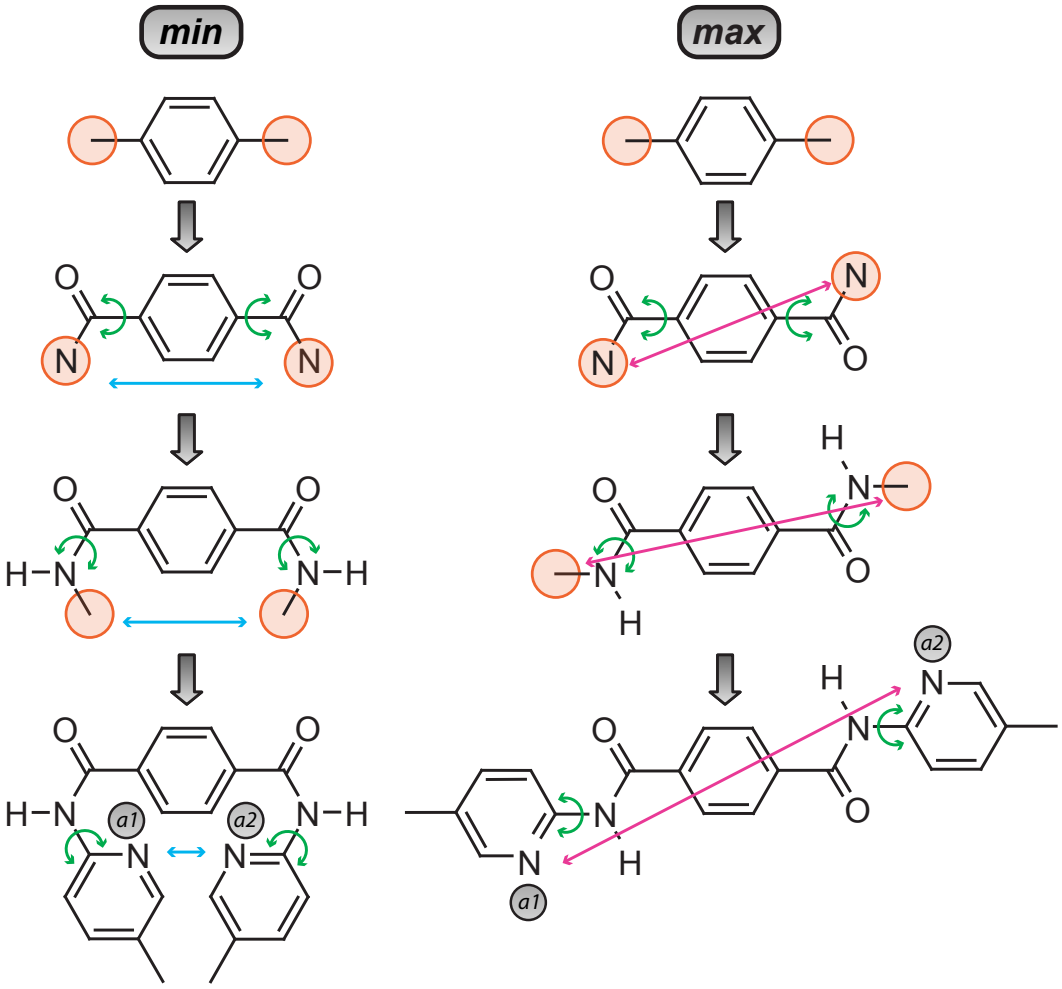


Figure 2.13. Scheme of the distance range estimation algorithm. In the first two steps the marked torsion angles (green arrows) are chosen in such a way that the distances between the outgoing atoms (red circles) are minimal (left) or maximal (right), respectively. In the last step the torsion angles of the last added fragments are set so that the distances between the queried atoms ($a1$ and $a2$) are minimal or maximal, respectively.

The heuristic for the calculation of distance ranges is illustrated in Figure 2.13. Let two atoms $a1$ and $a2$ be given for which the distance range has to be determined. First, the fragments to which the atoms belong are identified and the shortest path in the fragment tree between these fragments is determined. In the case that the number of fragments along this path is odd we start with the central fragment and iteratively add the next two connected fragments until the two outer fragments to which the atoms $a1$ and $a2$ belong are reached. Two different procedures are applied for determining the minimal and the maximal distance, respectively. For generating the minimal distance in each step the torsion angles of the newly added two fragments are chosen for which the distance between

the two corresponding outgoing atoms is minimal and no internal clashes exist. For determining the maximal distance, here, the distance is set to the maximally possible value. Note that transitivity holds for neither minimal nor maximal distances. Thus we cannot guarantee not to lose optimal solutions in this step. This is repeated until the fragments of the atoms $a1$ and $a2$ are reached. In this last step the torsion angles of the fragments containing the atoms $a1$ and $a2$ are set such that the atoms are minimally or maximally distant, respectively.

The determination of the minimal and maximal distance between two corresponding interaction surfaces of the considered atoms $a1$ and $a2$ differs from the above mentioned algorithm only in the last step. When the last two fragments are added, the distance between the midpoints of the corresponding surfaces are adjusted for being minimal or maximal, respectively. Then both surfaces are discretized to point sets and all distances between the discrete interaction points are calculated. The resulting distance ranges are stored in a distance matrix.

In the case that the examined fragment list has an even number of fragments the distance analysis starts with the two central fragments. The corresponding torsion angle between them is adjusted such that the appropriate outgoing atoms are set to their minimal or maximal distance, respectively. With the exception of small flexible rings all single fragments are treated as rigid and thus offer only one conformation. If the two targeted atoms belong to the same fragment no conformational analysis has to be performed and the distance between the atoms and the surface can be calculated directly. In the case of a fragment consisting of a small flexible ring several low-energy conformations are computed with CORINA (Sadowski & Gasteiger, 1993). All of them are considered iteratively within our procedure.

2.2.2.2 Complex Construction

The precomputation phase results in a set of cliques each of which represents a putative interaction pattern for the molecular complex between host and guest molecule. We denote the interactions represented by a single clique as targeted interactions. Complex construction essentially amounts to a heuristic optimization of the complex configuration with respect to a specific score. Scores are heuristic estimates of binding energies that mostly account for contributions by directional interactions. Thus scores are minimized during the optimization and the lowest score corresponds to the predicted most favorable solution. At first, we determine for each clique a simple lower bound on the best attainable score. This lower bound is calculated by simply adding the optimal scores for all participating interactions. In the complex construction phase the cliques are processed

sequentially in increasing order of this score. First, an initial configuration is generated (see Paragraph *Generation of initial configurations*). In the second step the adaptive two-sided incremental complex construction algorithm (see Paragraph *Adaptive two-sided incremental complex construction*) is applied. The algorithm terminates if a solution is found that fulfills all targeted interactions of a clique. If other cliques exist that exhibit the same estimated maximal interaction energy the algorithm proceeds until all cliques with the same estimated score are processed. From each of the cliques that lead to a valid complex structure the ten best-scoring solutions are included in the solution set.

Generation of Initial Configurations

The docking process starts with the generation of initial configurations of two selected fragments, one fragment from each molecule. The precomputation of cliques facilitates the targeted generation of an initial configuration. For each single clique of the clique list at first the two fragments are selected that accomplish as many of the targeted interactions as possible. In the following these two fragments are called base fragments.

In the case that only one directional short-range interaction between the selected base fragments is targeted, the one-point base fragment placement algorithm is used as described in Kämper et al. (2006) (see Figure 2.14). In this algorithm a number of discrete placements are generated by just using geometric information of the two participating interaction surfaces. Therefore each vector between the interaction center and the interaction points of one fragment is superimposed onto each vector between the interaction center and the corresponding interaction points of the counter fragment. Then one fragment is rotated around the superimposed vector in discrete steps of 30°. If the two selected base fragments exhibit more than a single interaction, one of the possible interactions is selected for generating different sterically possible configurations with the one-point placement algorithm. Then, only those configurations are retained for the next step in which all targeted interactions between the base fragments are realized, i.e. the interaction criterion is fulfilled (see Figure 2.7) and no atom overlaps between the fragments exist. All complexes are assessed by means of the scoring function (see Section 2.2.1.3). In order to reduce the number of highly similar structures for all generated complexes a clustering procedure is applied (see 2.2.2.2). The remaining initial configurations are then submitted to the subsequent adaptive two-sided incremental construction algorithm. All of them exhibit the same directional short range interaction pattern, namely the targeted interactions between the two base fragments.

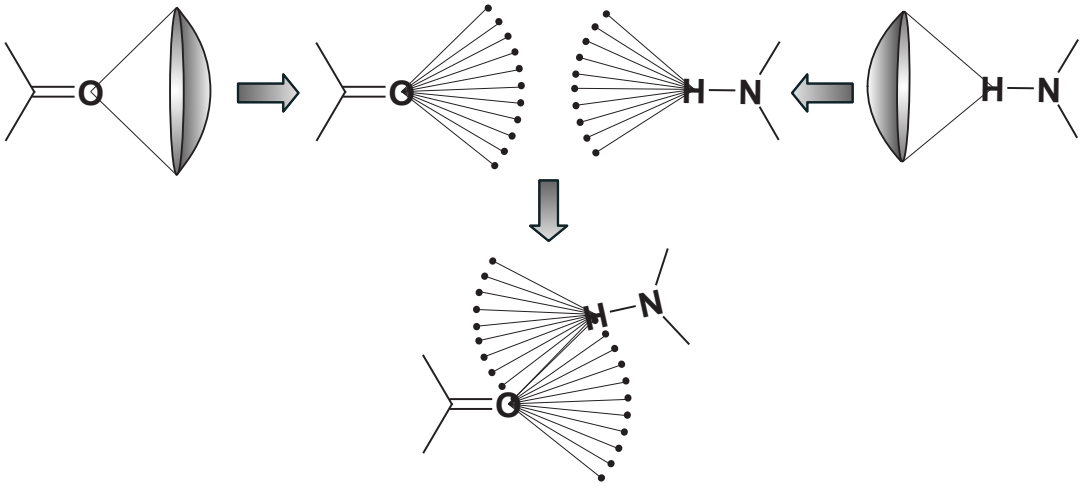


Figure 2.14. Illustrations of the one-point placement algorithm. Each vector between the interaction center and the interaction points of one fragment is superimposed onto each vector between the interaction center and the corresponding interaction points of the counter fragment. Then one fragment is rotated around the superimposed vector in discrete steps of 30°.

Computing the Fragment Order

After the initial configuration of the two base fragments has been computed, we determine the build-up order for the remaining fragments of each molecule. During the adaptive two-sided incremental construction phase the remaining filters of the clique are targeted that have not been achieved in the initial configuration. Therefore the fragments of both molecules have to be added in a synchronized manner.

The order in which the fragments are processed is computed with a greedy heuristic. Given are the two fragment trees of the molecules and the list of interactions that are targeted in the subsequent incremental construction phase (see Figure 2.15). At first, the fragments to which the remaining targeted interactions belong to are determined. The algorithm starts from the two root nodes a and A and marks these nodes as visited. The distances for all remaining interactions are computed. Here, we define the distance of an interaction as the sum of node distances from the corresponding nodes to the nearest nodes of the fragment trees that have been visited already. For instance, the distance of the interaction g-C to the interaction a-A is in Figure 2.15 equal to 7 (five fragments for the host - b, c, e, f, and g - and two fragments for the guest - B and C). The distance of the interaction e-E to a-A is equal to 6. The interaction with the smallest distance is selected as the next interaction. For the example of Figure 2.15 the interaction e-E is selected prior to g-C. All unvisited nodes that are on the path from the interaction nodes to the root nodes in their fragment trees are marked as visited.

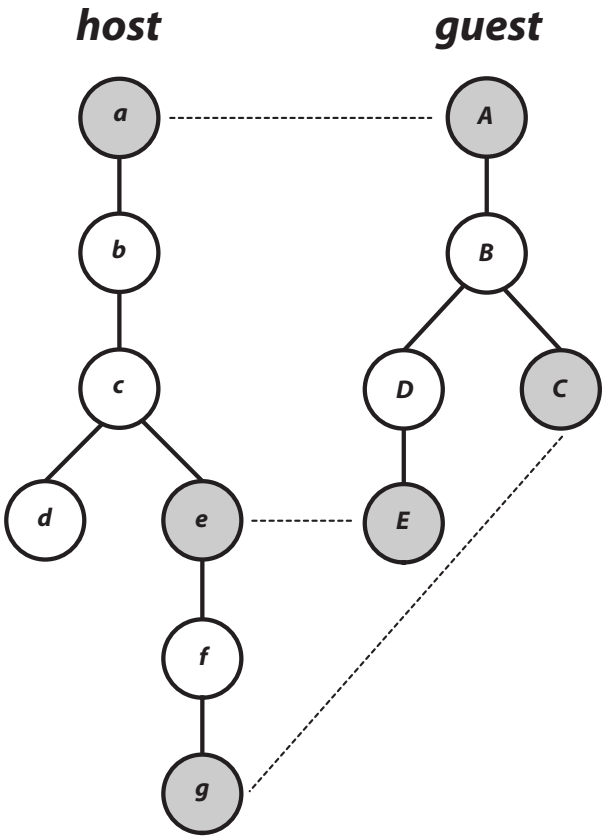


Figure 2.15. Example for the computation of the fragment order. Shown are the fragment tree representations of two interacting molecules. The nodes represent fragments; edges denote two covalently bound fragments. Nodes depicted in grey comprise centers of directional short-range interactions. Dashed lines represent the targeted interactions. The applied algorithm determines the order of execution of both fragment trees, such that all targeted interactions are reached synchronously (see text).

The fragments that have been visited in this step are stored in the fragment order list in the reverse order of their visit. In the first step of our example the fragment order for the receptor is computed to *a*-*b*-*c*-*e* and the order of the guest is *A*-*B*-*D*-*E*. This procedure is repeated until no more interactions remain. In the case there are still unmarked nodes in the fragment trees, they are visited in an order where the terminal nodes are reached as early as possible.

Adaptive Two-Sided Incremental Complex Construction

In this phase both molecules are constructed, starting with the initial complexes, in order to complete the whole complex structure. Here, we apply an iterative procedure which consists of three repetitive steps (Figure 2.4): a combinatorial step, a filtering step and a greedy step. At the beginning of the two-sided incremental construction the following information has already been computed:

- the initial base fragment configurations
- the list of the remaining targeted interactions that have to be realized in the final complex
- the order in which the remaining fragments have to be added to each of the two molecules (see Paragraph *Computing the fragment order*)
- the distance ranges from any outgoing atom to its respective next targeted interaction atom (see Paragraph *Distance range estimation method*).

For each of the remaining interactions the following procedure is iteratively repeated. In the combinatorial step the molecule which requires fewer fragments to reach the next targeted interaction is expanded combinatorially in torsion space until the next targeted interaction group is reached. Partial solutions with inter- or intramolecular clashes are discarded. The scoring function is applied for obtaining an estimation of the energy of each partial solution. In order to reduce the number of highly similar structures a clustering is performed (see Paragraph *Clustering*).

If the molecule has reached the next targeted interaction group the algorithm proceeds to the filtering step and the other molecule is incrementally built up in torsion space until the targeted interaction is realized. Here, two kinds of filters are applied (see Paragraph *Applied filters in the complex construction phase*). If none of the (partial) solutions corresponds to the constraints of the currently processed clique this particular clique is skipped as the interaction pattern represented by the clique does not lead to a valid complex structure. In this case, the algorithm proceeds to the next clique. All (partial) solutions that fulfill the mandatory targeted interactions are submitted to the clustering procedure again.

After all interactions of the processed clique are realized and still fragments remain, the algorithm proceeds to the greedy step. In this step the molecules are expanded alternately. Here, maximally the 100 best-scoring partial solutions are taken into the next construction iteration regardless of the absolute value of the scores. This cycle is repeated until the entire complex is completely built up. Finally all solutions are ranked by their scores.

Applied Filters in the Complex Construction Phase

In the filtering step of the complex construction, filters are applied in order to direct the complex construction to solutions that exhibit the targeted interaction pattern of the currently processed clique. These two types of filters are a distance filter and an interaction pattern filter.

The distance filter (see Figure 2.16) is applied whenever the position of only one interaction group of the next targeted interaction is known whereas the

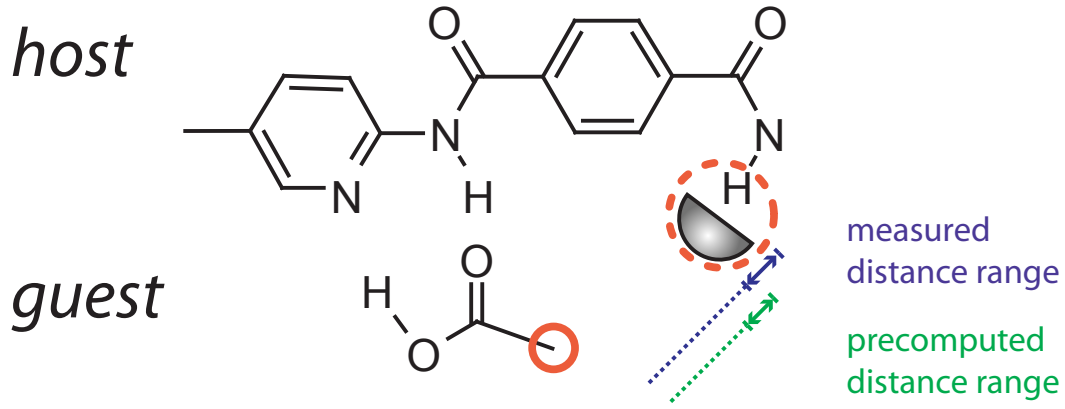


Figure 2.16. Distance filter. The distance from the current outgoing atom (red circle) to the targeted interaction (dashed red circle) is calculated (blue). If this distance exhibits an overlap with the precomputed distance range (green) the partial solution exhibits the potential to lead to the targeted interaction pattern prescribed by the clique.

counter group of this particular interaction has not been placed yet. In such a case the algorithm compares the current distances between the outgoing atom (see Section 2.2.2.1) and the discrete surface points of the counter group with the corresponding precomputed distance range. If the two distance ranges do not overlap, the solution is discarded. Otherwise, the current construction state of the complex exhibits the potential to fulfill all targeted interactions and is retained.

The interaction pattern filter verifies whether in a given (partial) solution all targeted directional short-range interactions considered so far are realized at this particular construction step. The filter is applied only when a fragment is placed that comprises an interaction group, which should form an interaction with its already placed counter group from the clique. A (partial) solution that does not meet all requirements is discarded.

Optimization of Matches

The first interaction of the complex that is formed between the two base fragments is in advantage compared to all following interactions. It is in the nature of the one-point interaction placement algorithm that the geometry of the first interaction is ideal with respect to the scoring function. All following interactions that are found during the complex construction are discriminated in this regard, as their arrangement is not necessarily ideal. Whenever a new directional interaction is found one molecule is reoriented towards the other. In this step, for each existing directional interaction - including the last one - four points are considered (see Figure 2.17): the two corresponding interaction centers and one

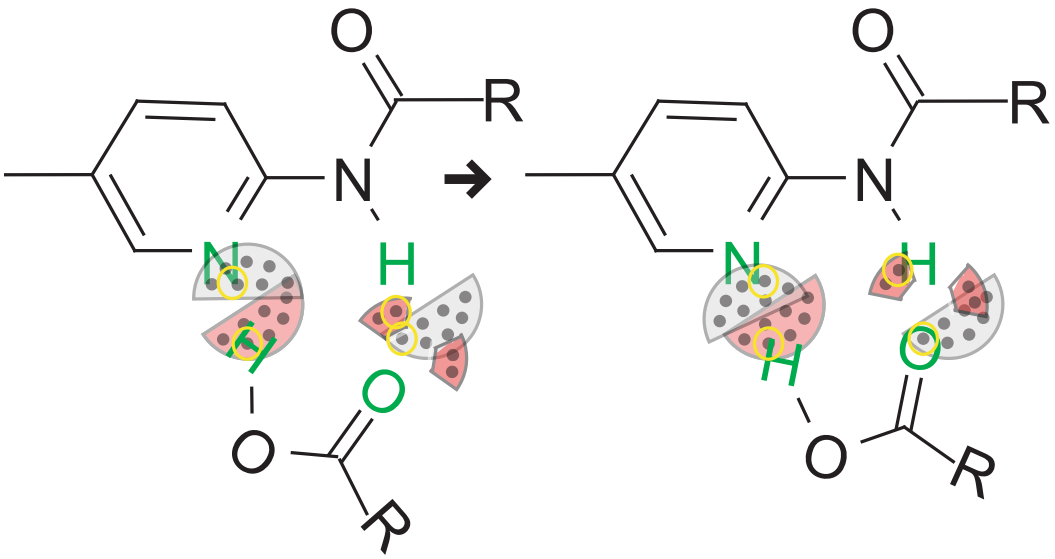


Figure 2.17. Optimization of matches using the interaction centers (green) and the closest discrete interaction points (yellow circle) on the interaction surfaces (left). Whenever a new directional interaction is found these points are used for an optimization of the placement (right).

point from each of the two interaction surfaces (yellow circles) that is the closest to the counter interaction center. Then the selected interaction centers are superimposed onto the selected discrete interaction points. This step is combined with an additional clash test. If a clash occurs the new position is disregarded and the old position is kept.

Clustering

Highly similar (partial) solutions can be clustered to one representative solution. This reduces the number of (partial) solutions while ensuring that no important structural information is lost. After each construction step a complete-linkage clustering algorithm is performed (Rarey et al., 1996a). Here, the distance between two (partial) solutions is measured by means of the root-mean-square deviation (RMSD) for all atoms that are placed in the current state. The RMSD threshold has been set to 0.8 Å. For the purpose of the new algorithm an additional clustering criterion is implemented. If only one interaction group of a targeted interaction has been placed the two partial solutions are not clustered if the distance between the two atoms exceeds a threshold. This guarantees that partial solutions with different properties regarding the targeted interactions are retained. For this distance the threshold has been set to 0.4 Å.

2.3 Validation by Means of Redocking

A common test for the evaluation of a docking tool is to assess whether it is able to reproduce native crystal structures. This test is referenced as *redocking test*. For this purpose, crystallographically determined host-guest complex structures are taken from a structural database. The crystal coordinates are used as reference coordinates. From the crystal structure single MOL2 files are derived for the guest and the host molecule. The docking tool obtains the single host and guest molecules as input. For docking no information regarding the configuration of the complex structure and the conformations of the single molecules is used from the crystal. All generated solutions were scored by means of the integrated scoring function of FLEXR and sorted in increasing order of their scores. The best-scoring solution, i.e. the one with lowest score, is on rank 1.

To assess whether a redocking run is successful we compare the generated hydrogen-bond pattern of the predicted complex structure to the crystal structure and furthermore we compute the RMSD of all predicted atom coordinates to the reference atom coordinates. The RMSD value between two different molecules is defined as follows:

$$RMSD(\theta_1, \theta_2) = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \tag{2.2}$$

where θ_1 is the set of coordinates of complex **1**, θ_2 is the set of coordinates of complex 2, n is the number of coordinates, $x_{1,i}$ is the ith coordinate of complex 1 and $x_{2,i}$ is the ith coordinate of complex **2**. An RMSD value of below 2 Å is commonly considered as a near-native prediction.

2.3.1 Test Dataset

In order to evaluate our docking strategy we assembled a set of ten experimentally determined crystal structures of synthetic receptors and their comprised guest molecules (see Tables 2.2 and 2.3) from the Cambridge Structural Database (CSD) (Allen, 2002).

The Complexes

Within all of the ten selected test complexes hydrogen bonds are the main driving force of complex formation. The host molecules differ in their degree of flexibility and thus challenge our method in different ways. The most flexible host, i. e. the host molecule of complex **7**, consists of nine rotatable bonds. The host molecules

of complexes **6** and **10** show the smallest degree of flexibility, each of which has two rotatable bonds. The guest molecules include small heterocyclic rings or ring systems and aliphatic carbonic acids as well as small cationic and anionic molecules. All molecules are shown in Tables 2.2 and 2.3.

*Glutaric acid receptor (complex **1**).* The receptor for glutaric acid was synthesized by Garcia-Tellado et al. (1990). The complex formation was measured in $CDCl_3$. Within the complex four hydrogen bonds occur between host and guest molecule.

*Ammonium ion receptor (complex **2**).* Chin et al. presented a synthetic receptor for the ammonium ion (Chin et al., 1999). The receptor can potentially be applied as a ammonium sensing unit, which might for example be useful for the detection of ammonia in air. In contrast to crown ethers that complexate ammonium ions as well as potassium ions, this receptor shows a high selectivity towards the ammonium ion. This is due to the formation of three charged hydrogen bonds to the guest molecule. The complex formation was measured in CD_2Cl_2.

*Tricarboxylic acid receptor (complex **3**).* A tripodal receptor based on a amidopyridine motif was presented by Ballester et al. (2001). This receptor forms a strong 1:1 complex with cis-1,3,5-cyclohexane tricarboxylic acid in 20% $THF/CHCl_3$. The complex of the same receptor to the trans-isomer of the guest molecule is clearly less stable. Altogether six hydrogens bonds occur within the complex.

*Two-point binding receptor (complex **4**).* Pascal & Ho (1994) presented the diacidic two-point binding receptor for pyrazine. The two carboxyl groups form hydrogen bonds to the nitrogens of pyrazine. The complex formation was detected in $CHCl_3$.

*Barbiturate receptor (complex **5**).* Berl et al. (1999) presented a synthetic receptor for barbiturates that was obtained by means of a dynamic combinatorial synthesis scheme. The receptor is based on a dihydrazone motif. In the presence of dibutylbarbiturate the (Z/Z) dihydrazone isomer was the main product of the combinatorial synthesis. Dibutylbarbiturate forms two single and two bifurcated hydrogen bonds to host molecule. The complex formation was detected in $CDCl_3$.

*Creatinine receptor (complex **6**).* The selective receptor for creatinine was presented by Bell et al. (1995). The complex formation involves a chromogenic response that is caused by a proton transfer. This property enables the receptor to be used a sensing unit for creatinine. The receptor extracts creatinine

from water into chlorocarbon solvents. Three hydrogen bonds exist between the two molecules out of which two are of charged nature.

*Bisguanidinium receptor for phenyl phosphate (complex **7**).* The host molecule of complex **7** consists of two guanidinium moieties separated by a hexahydrodicyclopentapyridine spacer (Kneeland et al., 1993). The receptor was designed to mimic the interactions of the binding site of the staphylococcal nuclease in which phosphoesters are complexated by four hydrogen bonds. The complex to the guest molecule phenyl phosphate was detected in aqueous DMSO.

*Bis(guanidinium) receptor for sulphate (complex **8**).* The host molecule of complex **8** is structurally related to the host in complex **7** and is also based on a bisguanidinium motif. The receptor complexates sulphate by charged hydrogen bonds to each of the oxygen atoms of the guest. The complex was formed out in aqueous hydrochlorid acid.

*Caffeine receptor (complex **9**).* Waldvogel et al. (2000) presented a selective caffeine receptor. The receptor exhibits a triphenylenketal unit with three ureyl side chains. Caffeine is bound in CD_2Cl_2 solution via bifurcate hydrogen bonds to each of the ureyl units.

*Receptor for guanidinium derivates (complex **10**).* The synthetic receptor of Bell et al. (2002) complexates guanidinium derivatives in aqueous solution. Four charged hydrogen bonds determine the structure of the complex.

Preparation of the Molecules

The molecules were prepared as follows: First, we extracted the structures of the complexes from the CSD in the MOL2 file format. The protonation states of the molecules were taken from of the respective input structures and checked according to the corresponding paper. The atom and bond types as well as the formal charges were assigned automatically by a rule-based heuristic and checked manually. We replaced long aliphatic chains that do not contribute to the structure of the docked complex by methyl groups (denoted as R-groups in Tables 2.2 and 2.3). In the case of the host molecule of complex **5** only one half of the dimeric receptor was used in docking. The other half was replaced by a methyl group (denoted as an R-group in Table 2.2). Finally, the molecules were energy-minimized with the TRIPOS force field (Clark et al., 1989). This was done in order to obtain suitable bond lengths and angles. Thereby the aromatic ring system of the receptor in complex **9** was set to be rigid in order to avoid out-of-plane distortions.

Table 2.2. Structural formulas of the complexes **1** to **5**.

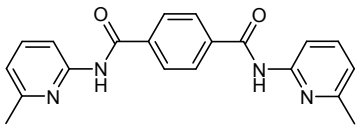
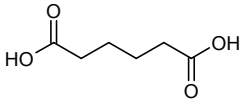
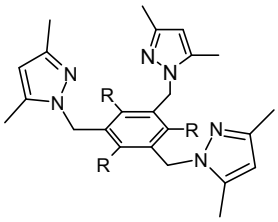
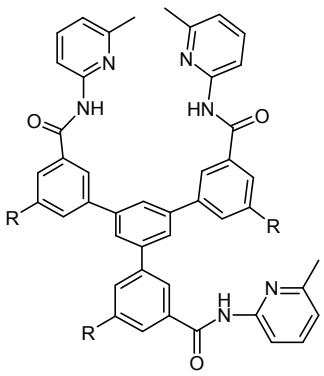
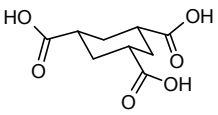
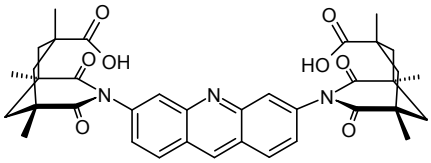
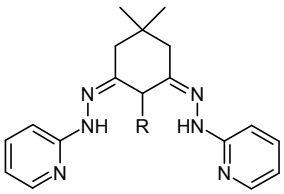
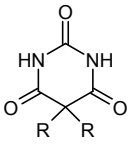
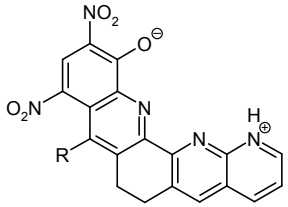
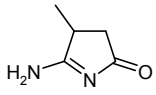
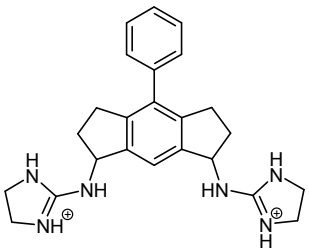
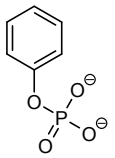
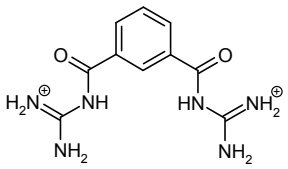
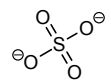
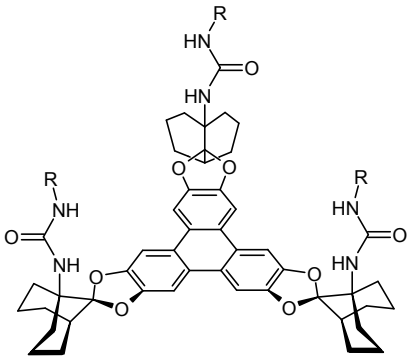
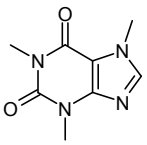
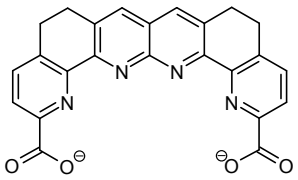
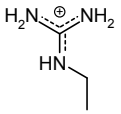
ID	Host	Guest	CSD-ID
1			JEWNUU
2			CUKTUX
3			QIJPEE
4			POLFUR
5			DAQVAS

Table 2.3. Structural formulas of the complexes **6** to **10**.

ID	Host	Guest	CSD-ID
6			ZESFEI
7			HASWUT
8			QAFVAV
9			WEWTEX
10			DAMQUD

2.3.2 Results

The results obtained in the redocking experiment are summarized in Table 2.4. As mentioned previously, in protein-ligand docking a commonly used criterion for the evaluation of docking results is the RMSD value of a predicted complex structure compared to an experimentally determined crystal structure. Here, typically an RMSD of below 2 Å is considered as a successful docking result. However, the RMSD of the best-scoring solution is not necessarily the best criterion for an assessment of docking solutions since it is highly dependent on the applied scoring function. Sometimes only slight deviations between the scores of two solutions prevent a solution with low RMSD at high rank (i.e. rank with a low rank number). For this reason we also report the lowest RMSD within the ten best-scoring solutions. Although we can demonstrate that our scoring function directed the docking algorithm to reasonable solutions, the particular scores are not tabulated since it is a known fact that they do not provide a reliable estimate of the binding affinity in most cases. The scores for the solutions of a particular clique do not vary much since all of them exhibit the same interaction pattern.

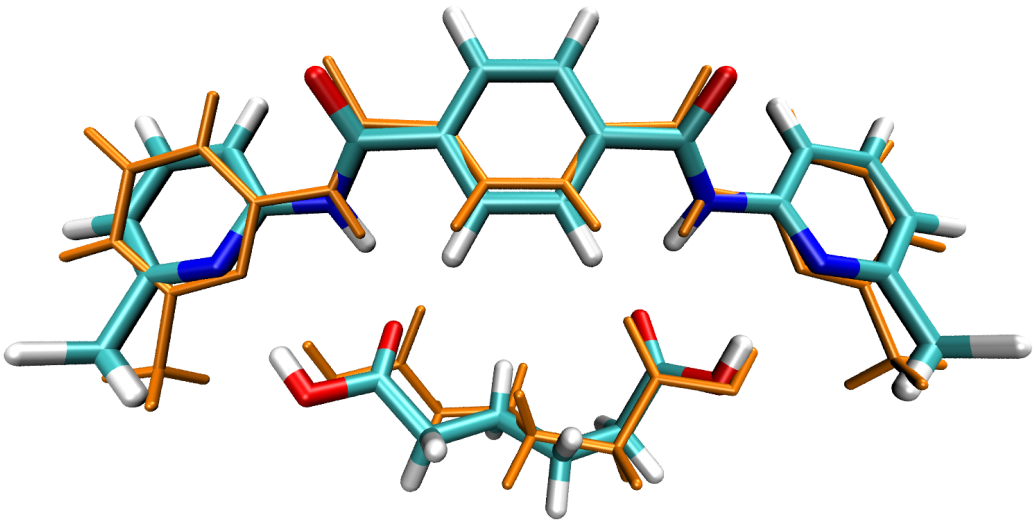


Figure 2.18. Docking result of complex **1** (rank 1, atom coloring) superimposed on the X-ray structure (orange). FlexR was able to reproduce the native crystal structure. A solution with an RMSD of 0.93 Å was found within less than 4 minutes on the first rank. The hydrogen-bond pattern corresponds to the crystal structure.

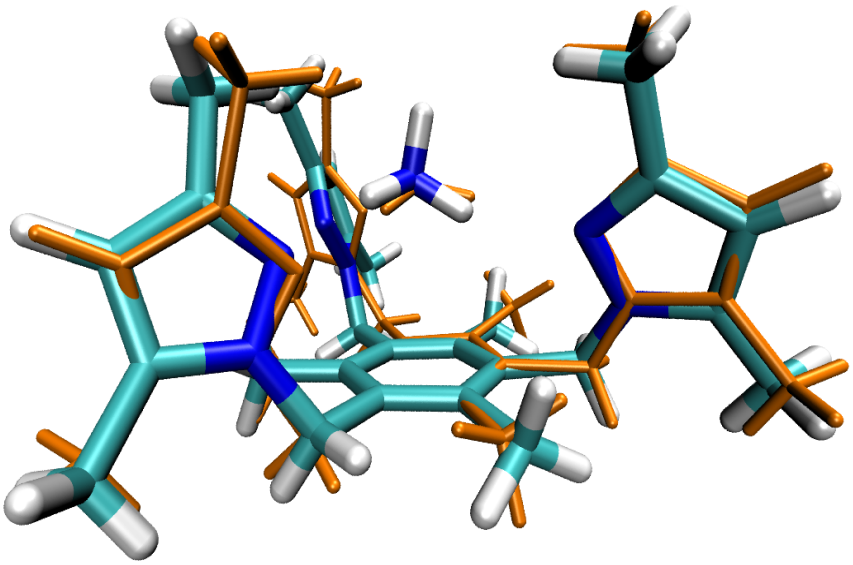


Figure 2.19. Docking result of complex **2** (rank 1, atom coloring) superimposed on the X-ray structure (orange). The predicted structure of complex **2** exhibits the native hydrogen-bond pattern as found in the crystal structure. Only marginal deviations are found within the receptor structure. The RMSD of 1.17 Å lies within the acceptable range of 2 Å. The symmetry of the system causes many identical cliques that all have to be processed.

Table 2.4. Docking results. For each complex, we list the root-mean-square deviation (RMSD) of the best-scoring solution, the best RMSD within the first ten best-scoring solutions and the CPU time. CPU times are obtained on an Intel P4 Xeon 3.06 GHz.

ID	RMSD [Å][1]	Min. RMSD [Å][2]	CPU time [MM:SS]
1	0.93	0.76	03:30
2	1.17	1.12	01:27
3	3.86	1.04	49:01
4	1.07	0.96	00:31
5	1.08	0.94	00:06
6	0.60	0.60	00:01
7	0.56	0.56	00:01
8	0.73	0.73	01:17
9	0.70	0.70	00:19
10	0.63	0.30	00:03

[1] best-scoring solution
[2] within first ten best-scoring solutions

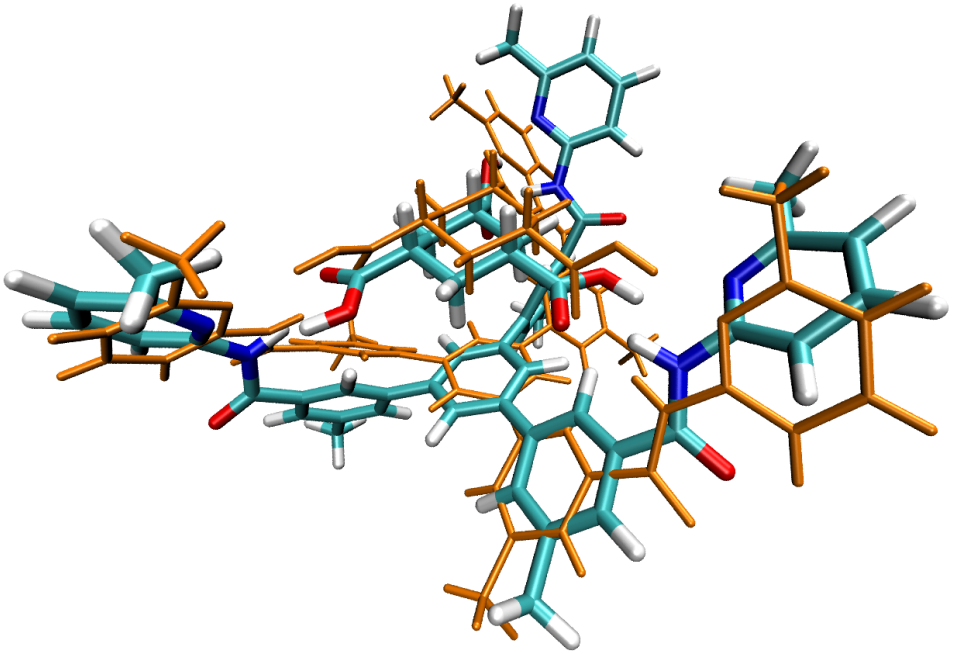


Figure 2.20. Docking result of complex **3** (rank 1, atom coloring) superimposed on the X-ray structure (orange). The solution found at rank 1 of complex 3 exhibits an RMSD that exceeds the defined threshold of 2 Å, although the exact hydrogen-bond pattern of the crystal structure was reproduced. Due to the high degree of flexibility and the symmetry of the system the computation took relatively long with about 50 minutes.

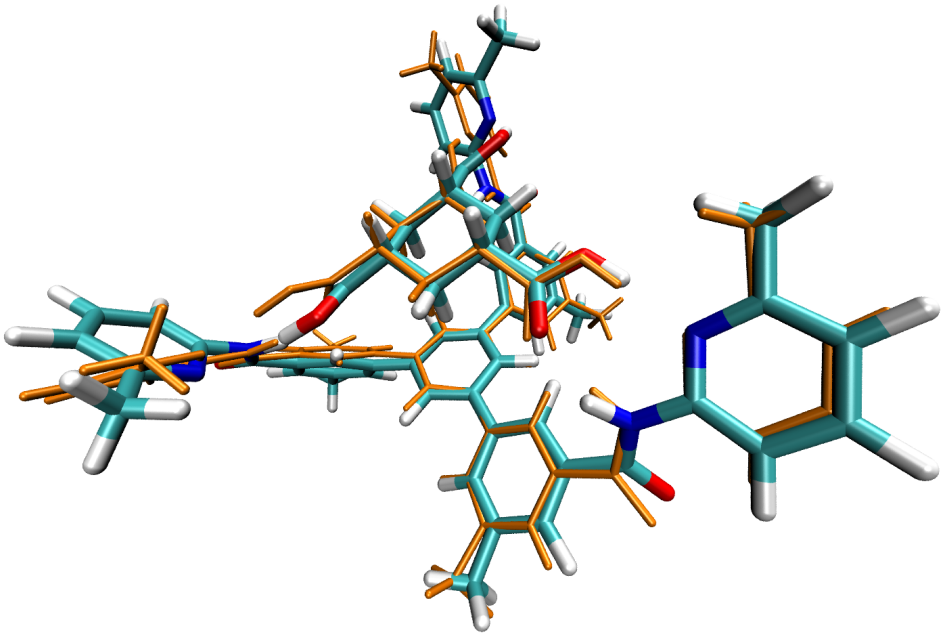


Figure 2.21. Docking result of complex **3** (rank 7, atom coloring) superimposed on the X-ray structure (orange). The solution found at rank 7 shows a good agreement with the crystal structure and exhibits an RMSD of 1.04 Å.

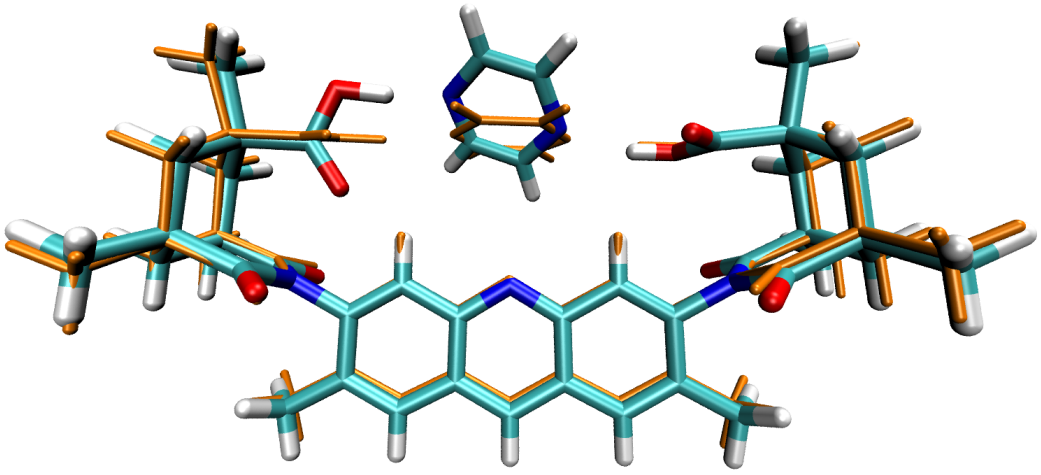


Figure 2.22. Docking result of complex **4** (rank 1, atom coloring) superimposed on the X-ray structure (orange). For complex **4** a near-native solution was found on the top rank (RMSD 1.07 Å) that has the same hydrogen-bond pattern as observed in the crystal structure. The deviation of the orientation of the guest molecule is due to packing effects in the crystal structure where π stacking is observed with other pyrazine molecules.

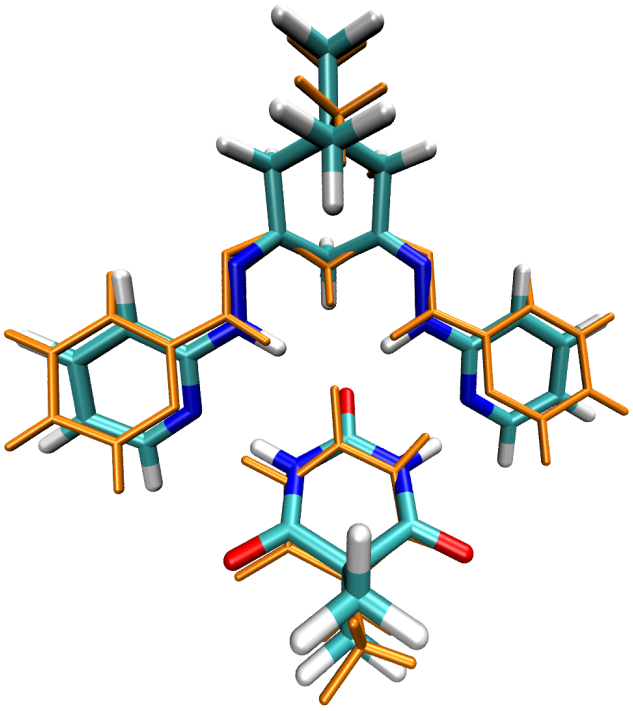


Figure 2.23. Docking result of complex **5** (rank 1, atom coloring) superimposed on the X-ray structure (orange). Complex **5** was predicted with an RMSD of 1.25 Å for the top-ranking solution. The exact hydrogen-bond pattern of the crystal structure with two single hydrogen bonds and a bifurcate hydrogen bond was found. A slight difference exists in the orientation of the two pyridine heterocycles of the receptor. The structure was predicted within 6 seconds.

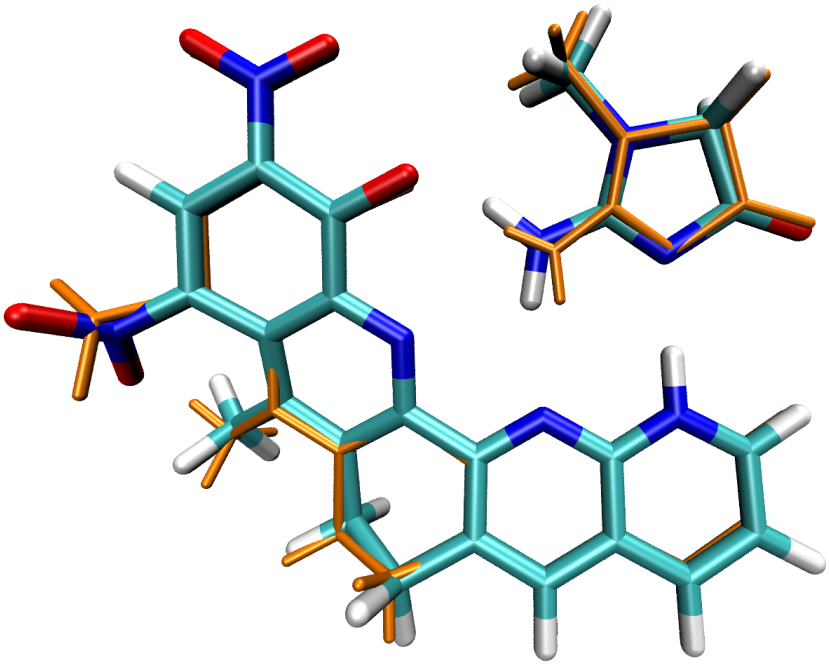


Figure 2.24. Docking result of complex **6** (rank 1, atom coloring) superimposed on the X-ray structure (orange). No significant difference is found in the predicted structure of complex **6**. The predicted hydrogen bonds are in exact agreement with the crystal structure. Only the carboxylate group stands approximately perpendicular to the orientation in the crystal structure. As there is no interaction to this group in the single unit cell, this difference should not be considered a prediction error. The structure was predicted within a second.

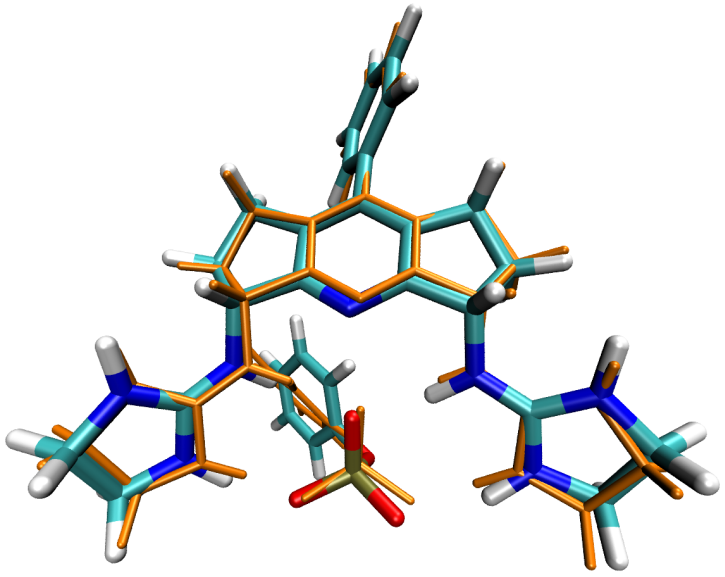


Figure 2.25. Docking result of complex **7** (rank 1, atom coloring) superimposed on the X-ray structure (orange). The predicted complex structure agrees almost perfectly with the crystal structure and the RMSD is 0.56 Å. The computation took 1 second.

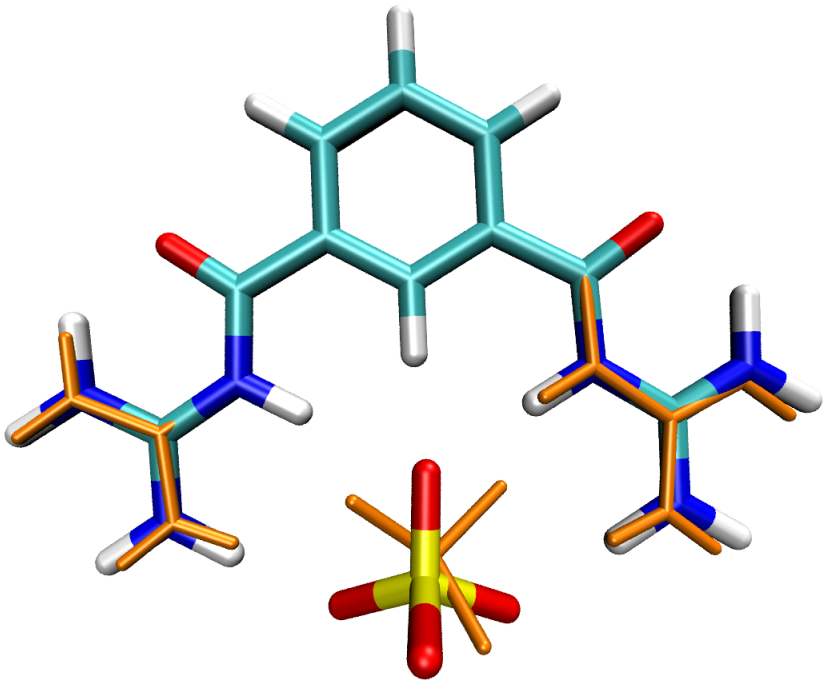


Figure 2.26. Docking result of complex **8** (rank 1, atom coloring) superimposed on the X-ray structure (orange). Although the RMSD of the top ranking solution of complex **8** falls into the threshold of 2 Å the predicted salt bridges do not correspond to the crystal structure. The structure prediction took 1 minute 17 seconds.

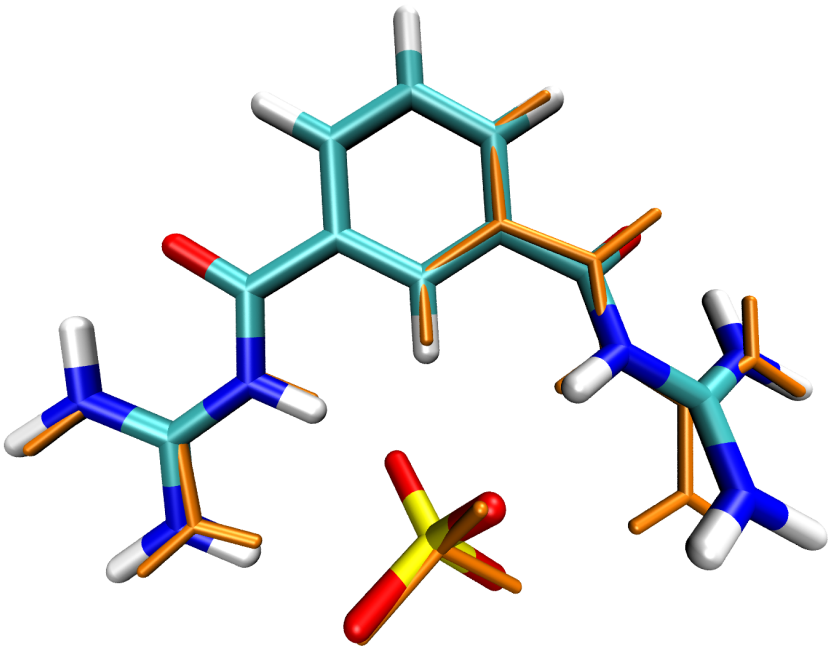


Figure 2.27. Docking result of complex **8** (rank 263, atom coloring) superimposed on the X-ray structure (orange). The figures shows that in principle FlexRis able to find a near-native complex structure but was not able to score it appropriately.

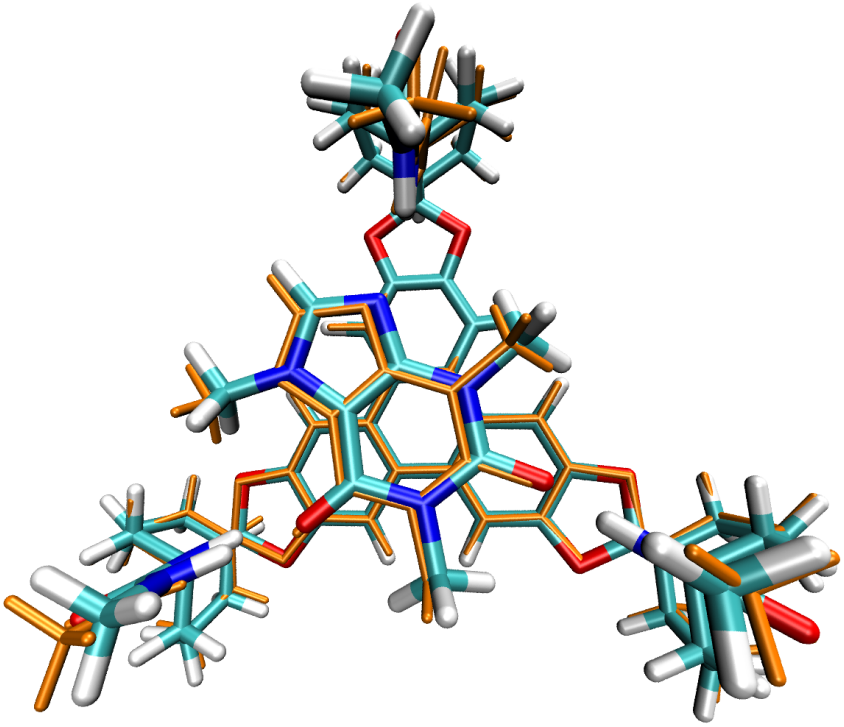


Figure 2.28. Docking result of complex **9** (rank 1, atom coloring) superimposed on the X-ray structure (orange). In the predicted complex structure all three bifurcate hydrogen bonds of the crystal structure were found. The RMSD is 0.70 Å. Despite the large and symmetric synthetic receptor the computation time was rather low with 19 seconds.

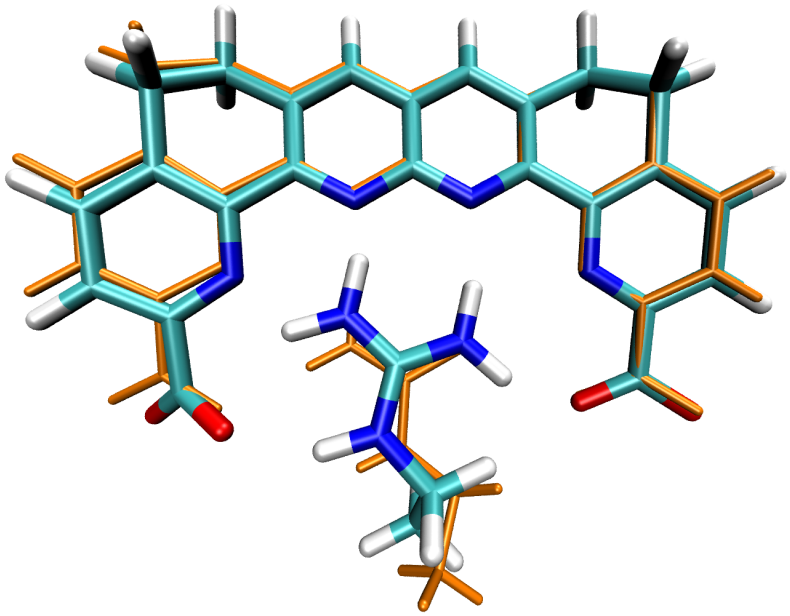


Figure 2.29. Docking result of complex **10** (rank 1, atom coloring) superimposed on the X-ray structure (orange). The predicted complex structure exhibits the bonding pattern of the crystal structure and has a low RMSD of 0.63 Å. The computation time was 3 seconds.

Following the defined RMSD criterion, for all of the complexes within our test dataset our new docking method predicts a reasonable structure at one of the first ten ranks. Furthermore, for complexes **1**,**2**,**4-7**, **9** and **10** near-native solutions were found at rank 1. For the complexes **2**, **5**, **6**, **9** and **10** an almost perfect prediction is obtained (Figures 2.19, 2.23, 2.24, 2.28 and 2.29). Complexes **4** and **7** exhibit slight deviations in the orientation of aromatic rings, whereas in complex **1** mainly the alkyl part of guest molecule differs from the crystal structure. The solution found at rank 1 of complex **3** exhibits an RMSD that exceeds the defined threshold of 2 Å, although the exact hydrogen-bond pattern of the crystal structure is reproduced (Figure 2.20). Nevertheless, the predicted complex structure at rank 7 of complex **3** has an RMSD of 1.0 Å (Figure 2.21) and is scored only slightly worse than the solution at rank 1. The best-scoring solution generated for complex **8** falls within the defined threshold and could thus be considered as near-native. However, regarding the generated hydrogen-bond pattern the solution differs from the crystal structure. In the experimentally determined structure six salt bridges are found of which four are bifurcate. None of the ten best-scoring solutions exhibits this interaction pattern. Considering all generated docking structures of complex **8**, a solution is found at a low rank (minimal RMSD of 0.48 Å observed on rank 263 as shown in Figure 2.27) that offers the same hydrogen-bond pattern as observed in the crystal structure. This supports the ability of our algorithm to generate a near-native structure for this test case, but at the same time reveals some problems of our scoring function in assessing them adequately.

Regarding the computation time, the results can be classified into three groups. Complexes **4-7**, **9** and **10** were generated in a couple of seconds. For complexes **1**, **2** and **8** a couple of minutes were needed. Only the highly flexible complex **3** requires a longer computation time of about 49 minutes. In comparison to the work by Kämper et al. (2006), most importantly, a significant acceleration has been achieved for the two highly flexible complexes **1** and **3**. Here, the computation time could be reduced by about a factor of 80 in case of complex **1** and to about one forth in case of complex **3**. To summarize, for five cases (complexes **1**, **3**, **5-7**) the new docking algorithm is faster than the previous one and for four cases it is marginally slower (complexes **2**, **4**, **8**, **9**).

2.3.3 Discussion

The redocking experiment showed that, in general, our docking strategy produces reliable predictions for complexes between synthetic receptors and guest molecules. Our approach to tackle the flexibility of two molecules simultaneously

successfully predicted all examples of our test set with respect to an RMSD of below 2 Å. In a previous study we already showed the general transferability of the FlexX concepts to the docking of synthetic host-guest systems. Here, additionally we focused on the more efficient handling of systems in which both molecules exhibit a high degree of flexibility. One limitation of our previous approach (Kämper et al., 2006) was observed for docking of complex **1**, where the docking times for forward and inverse docking exceeded several hours of CPU time. The second limitation was observed for complex **3** where no forward docking was possible at all due to the large conformational space of the receptor. The inverse docking strategy, however, worked but was comparably slow. Our new method predicted these structures significantly faster and, at the same time, near-native complexes were obtained. However, considering the complexes **2**, **4**, **8** and **9** our new method was slower than the method of Kämper et al. (2006), although the computation time was still in the range of seconds to minutes.

The following parameters significantly influence the computation time used for a complex:

- the number of possible interactions between the two molecules
- the flexibility of the molecules and
- the symmetry of the system.

The more interactions are possible the larger gets the docking graph. Flexible molecules can cause unspecifically wide distance ranges between centers of directional short-range interactions and thus many interaction pairs are compatible. This results in many edges in the docking graph. Consequently, the search for cliques in the docking graph slows down. It should, however, be noted that in none of the examples presented, the precomputation phase lasted longer than a few seconds. Thus, the precomputation phase is not the rate-limiting step of the algorithm. As the number of interactions and the flexibility of the molecules rises further, this might be a limiting factor for the method.

Our method does not consider symmetry information of the molecules. Due to this fact, many symmetrical cliques are generated if the molecules of the complex exhibit symmetry. Currently, all of them are processed although no additional information is gained and thus, the computation time for one complex rises with dependence to the inherent symmetry of the complex. Consider complex **3** where each of the molecules has threefold symmetry. By manually removing symmetric cliques we can show that the computation can be accelerated by a factor of ten without losing any information. The automatic detection of symmetries in synthetic host-guest complexes thus speeds up the computation of such complexes

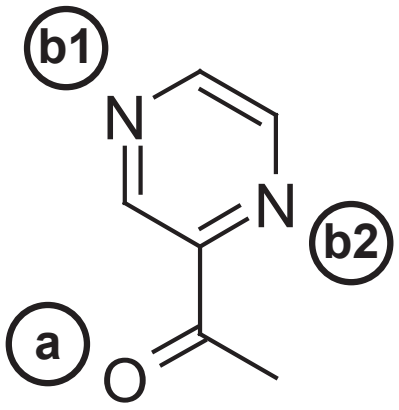


Figure 2.30. Dependence of distance ranges. In this example, the maximal distance between atom *a* and atom *b2* can only be realized, if the distance between atom *a* and atom *b1* is minimal. This shows the dependency of distance ranges which is disregarded in the docking graph generation step. Here, the distance ranges are treated as being independent from each other.

(Chen et al., 2006). In our tests, the number of cliques for the complexes has varied from 2 for complex **4** to 360 for complex **3**. In highly flexible complexes, additionally, many degenerate cliques are generated that do not represent valid complex structures. There are several reasons why this occurs. At first, in the precomputation phase only distances between the centers of the particular directional short-range interaction are considered. Clashes of the remaining atoms are not taken into account at this stage. Furthermore the estimated distance ranges are treated as being independent from each other, although this assumption is not valid in every case (Figure 20).

In our approach both molecules - synthetic receptor and the guest - are incrementally constructed from fragments. As stated above, in the beginning no conformation is known and thus the guiding role of one molecule for docking the other one is missing. In the precomputation phase our method needs the presence of directional short-range interactions such as hydrogen bonds or salt bridges which exhibit spatially much more constrained interaction geometries as lipophilic interactions. The consideration of the latter in this step would lead to infeasibly large docking graphs as generally many lipophilic interaction combinations are possible in common synthetic host-guest complexes. Furthermore, their geometrically ambiguous nature would not allow for applying strict distance filters. However, it is important to note that lipophilic interactions are assessed during the complex construction since the scoring function considers them. The molecules involved in synthetic host-guest systems that are solely based on lipophilic interactions are generally less flexible. Thus, at least one molecule is commonly rigid and an approach as proposed in Chapter 3 could be applied.

The test set used in Kämper et al. (2006) consisted of complexes that have been crystallized from aprotic solvents. Here, a complex has been integrated into the test set that was crystallized from aqueous solution. Although the structural influence of water, which is present in the crystal structure, is not tackled explicitly, near-native structures were generated. So far solvation is considered only implicitly in the scoring function which has been parameterized on experimentally derived protein-ligand complexes that have been crystallized from water.

Besides the forward and inverse docking strategy (Kämper et al., 2006) the new docking algorithm is the second approach that transfers the concepts of the Flex* program suite to the synthetic host-guest system and predicts near-native results for all test cases. This underlines the reliability of the whole concept. Which of the two methods is applied best for a given synthetic host-guest complex depends on the flexibility of the molecules. In the case that only one molecule has to be treated as flexible the approach introduced by Kämper et al. (2006) is the method of choice. If both molecules are flexible our new algorithm should be applied.

2.4 Virtual Screening as a Virtual Test for Selectivity

The redocking experiment in Section 2.3 showed that, in general, our novel method for the structure prediction of synthetic host-guest complexes produced near-native solution for a representative set of test complexes. The ability of a tool to generate near-native structures can be considered as a prerequisite for an application as a virtual screening tool. To furthermore show its potential in a virtual screening scenario, it is essential to test whether the tool is able to reproduce experimentally proven selectivity of a synthetic receptor for a given guest molecule. As an example we chose a synthetic receptor created by Bell et al. (1995), which specifically binds creatinine. In our experiments we virtually test its selectivity by means of two virtual screening scenarios described in the following sections.

2.4.1 Test System

Creatinine is an intermediate metabolite in the muscles and is excreted from the blood by the kidneys. The determination of the creatinine concentration in blood serum and urine is a good indicator for the renal function.

A very simple and commonly used method of experimentally measuring the concentration of creatinine is its reaction with picrate solution (Jaffé reaction), which results in a red-colored product. However, the specificity of this reaction

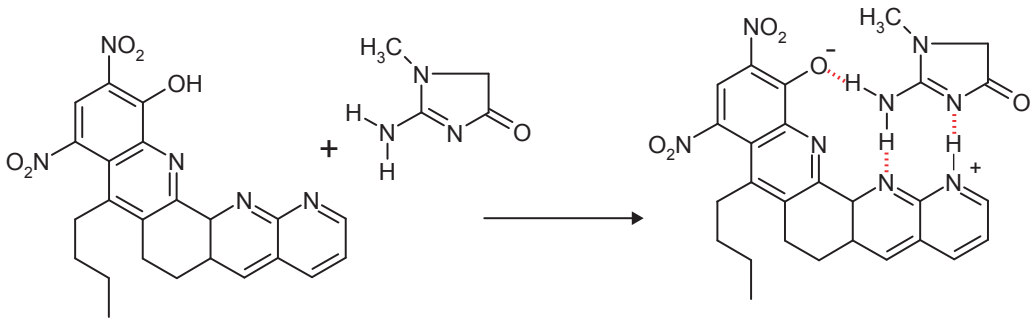


Figure 2.31. The tautomeric form of the synthetic receptor which is shown on the right is stabilized by the complexation of creatinine (Bell et al., 1995). The proton transfer involves a chromogenic response and results in a brownish orange solution that indicates the presence of creatinine.

is poor and hence false-positive detections can occur. Other more advanced approaches use enzymes, which are apparently much more specific, but have the drawback of low stability and higher costs.

Bell et al. (1995) presented the selective receptor for creatinine. This synthetic receptor of creatinine is much more stable than proteins and secures specificity. The dissociation constant of the complex in water saturated chloroform amounts to 0.5 μM. The binding of the guest molecule creatinine goes along with a change in the chromophore of the receptor. This chromogenic response is caused by a proton transfer within the receptor molecule, stabilized by creatinine (see Figure 2.31). The receptor extracts creatinine from water into chlorocarbon solvents and forms an intense brownish-orange-colored complex. This property enables the receptor to be applied as a sensing unit for creatinine.

2.4.2 Design of the Study

In a first step, we test whether FlexR can identify creatinine as one of the best-binding guest molecules of the creatinine receptor among a large set of molecules. Having tested this ability we evaluate the selectivity of the synthetic host for creatinine. To do so, we virtually screen for molecules that occur as metabolites in human bodies and could thus potentially interfere with the detection of creatinine.

2.4.3 Screening Sets

Two screening sets are extracted from public databases. In order to challenge our tool we apply filters, assuring that all selected molecules exhibit similar chemical properties in comparison with the native ligand creatinine. This filtering step helps to focus only on relevant molecules.

The first set consists of molecules taken from the ZINC-database (Irwin & Shoichet, 2005). The ZINC-database provides molecules from a large number of chemical vendor catalogs. All molecules are present as three-dimensional structures with appropriate bond lengths and angles, as well as reasonable protonation states. The molecules have been downloaded in the MOL2 file format. We apply the following filter rules:

- molecular weight 80 - 140 $\mathrm{g\,mol^{-1}}$
- 1-3 hydrogen bond acceptors
- 1-3 hydrogen bond donors
- 0-2 rotatable bonds.

Finally, this set contains 5,371 molecules. These molecules serve as a decoy set. The native ligand creatinine is added to the dataset as a test molecule in order to verify its selective binding.

The second set of molecules is extracted from the KEGG database (Kanehisa et al., 2006). The KEGG database links genomic and molecular information with the aim to provide a reference knowledge base as the basis for understanding higher order biological systems (Kanehisa et al., 2006). One of the integrated modules is called KEGG LIGAND and contains chemical compounds that occur in organisms as metabolites. This set consists of 12,042 molecules. This second set serves for the test in which potentially competing compounds are to be identified. We again only extract molecules with a similar interaction potential as creatinine. The criteria are as follows:

- 1-3 hydrogen bond donors
- 1-3 hydrogen bond acceptors
- 0-2 rotatable bonds.

We generate three-dimensional structures with standard bond length and angles with CORINA(Sadowski & Gasteiger, 1993). A PYTHON script sets the molecules to reasonable protonation states. This means that amines are protonated, and acidic groups are deprotonated. In the end, this set consists of 1,181 molecules.

2.4.4 Results and Discussion

The first test intents to principally determine whether the native ligand creatinine is among the best-scoring molecules of the derived ranking list of the first dataset. In fact, creatinine is found to be among the first 5% of top-ranking molecules. Together with the near-native redocking results presented in the last section, this result can be considered a prerequisite for the selective recognition of creatinine.

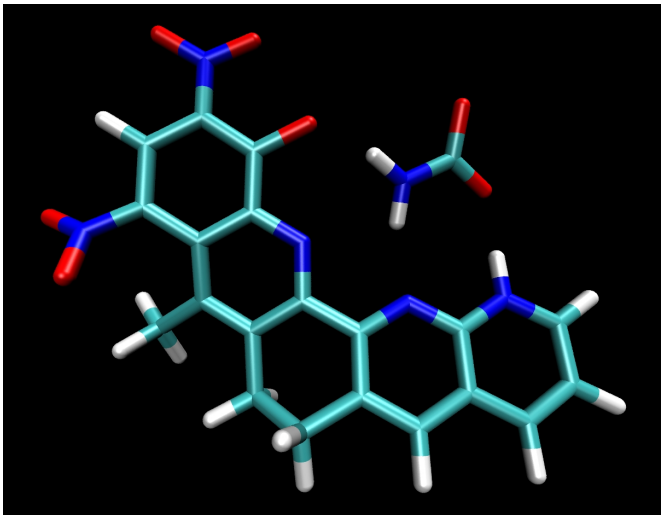


Figure 2.32. Compound C01563 (carbamate) was found on rank 2. The docking time was 1.6 seconds

We can assure that competitive molecules were within the test set, because the decoy structures all were chosen to be similar to creatinine with respect to size and interaction properties.

The second test intends to check whether FlexR can help to identify molecules that could trigger a false-positive detection signal in the *in vitro* testing. All of the tested molecules are metabolites from organisms and can thus potentially occur in the blood sample used for the creatinine test. On the average the complex structure prediction between a ligand and the creatinine receptor took ten seconds. Creatinine was found to be within the first 4% of the derived ranking-list of the virtual screening. Principally, this result confirms the strong affinity of the receptor to creatinine. However, among the best-scoring solutions, several potentially interfering compounds are found. As previously described, the chromogenic response is due to the proton transfer within the creatinine receptor which is stabilized by creatinine. Hence, each molecule that stabilizes this tautomeric form of the synthetic receptor exhibits the potential to interfere with the detection signal. Four high-scoring molecules are shown in Figures 2.32, 2.33, 2.34 and 2.35. Carbamate is a molecule occuring in the nitrogen metabolism, maleamic acid accumulates in the nicotinate and the nicotinamide metabolism, and cytosine is a part of the DNA and thus of high occurence in cells. All of these molecules stabilize the receptor in a manner similar to creatinine. To give the ultimate proof that these molecules will trigger a false-positive detection signal, however, experimental verification is required.

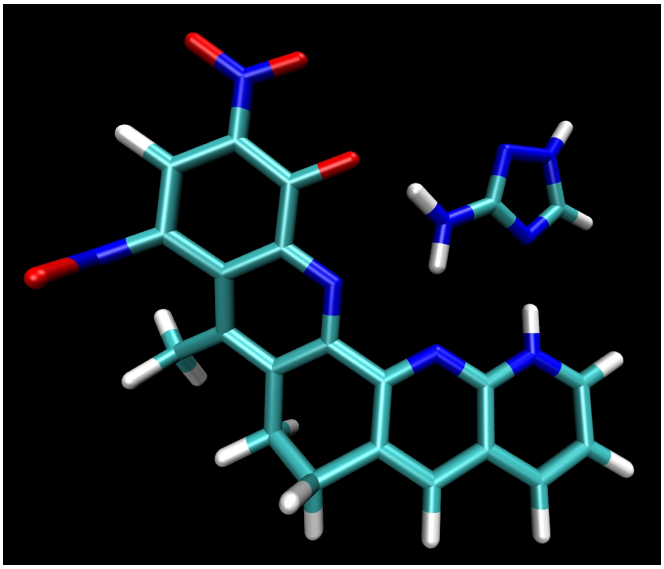


Figure 2.33. Compound C11261 (aminotriazole) was found on rank 35. The docking time was 1.9 seconds.

2.5 Conclusions and Outlook

We have developed a fast and fully automated method for predicting the structure of binary complexes between synthetic receptors and their comprised guest molecules. Our new approach tackles the flexibility of both molecules simultaneously. We created a highly efficient adaptive two-sided incremental build-up approach. We can significantly reduce the search space by building up of each of the two molecules with respect to the counter molecule. In comparison to the work presented by Kämper et al. (2006) we achieved a significant acceleration for complexes that consist of highly flexible molecules. At the same time we maintained the quality of the results. As shown our tool can be applied in a virtual screening

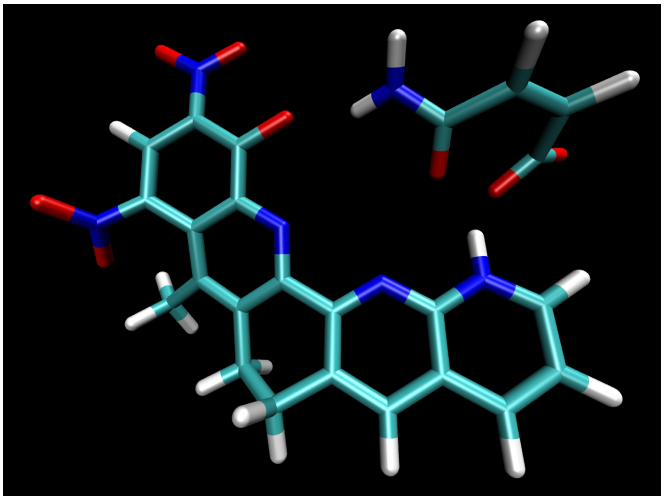


Figure 2.34. Compound C01596 (maleamic acid) was found on rank 48. The docking time was 24.0 seconds.

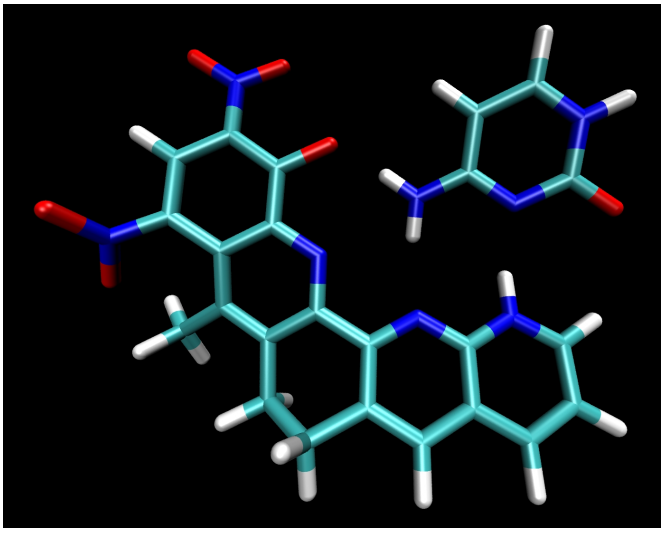


Figure 2.35. Compound C00380 (cytosine) was found on rank 50. The docking time was 1.3 seconds.

scenario, for example to help experimentalists to identify molecules that could possibly compete with the original molecule. This might give insights of how to improve a synthetic receptor in order to increase its selectivity towards the given guest molecule. This approach has the potential of opening up new scenarios for the computer-assisted design of novel synthetic host-guest complexes.

There are a number of aspects which may be further developed in future work concerning FlexR. The structure prediction of host-guest complexes with macrocyclic host molecules is currently not possible. This is due to the missing possibility of building up macrocyclic molecules in an incremental manner, an essential part of our algorithm. Work along this line has been conducted in a master thesis. The results, however, showed that the efficient conformational sampling of macrocyclic molecules is a demanding problem with no straightforward solution. A further aspect involves the implementation of metallic guests in the structure prediction of host-guest complexes. FlexR would have to consider the coordination geometries of the metal ions. Although this is theoretically already possible, almost all host-guest complexes with metal ions as guest molecules comprise a macrocyclic host molecule. One final aspect for the further development includes the docking of entirely hydrophobic guests, since the described version of FlexR focuses on only hydrogen bond based complexes. This class of complexes is mainly formed in non-polar solvents. Host-guest complexes formed in polar solvents are, however, much more demanding: in water, e. g., the main driving for complex formation is usually hydrophobic interactions, because hydrogen bonds in the complex always compete with water molecules. However, our algorithms need directional interactions; otherwise, the geometric filters cannot be applied.

Hydrophobic interactions are non-directional by nature and thus cannot be used for guiding the complex construction. On the other hand, hydrophobic synthetic receptors are usually designed to be rather rigid for entropic reasons, and thus principally the protocols presented in Chapter 3 can be applied.

Part II

Computational studies on β-cyclodextrin

3

Improved Cyclodextrin Based Receptors for Camptothecin by Inverse Virtual Screening

This chapter describes a novel protocol for the computer-aided optimization of a synthetic receptor for a given guest molecule based on the inverse virtual screening of receptor libraries (Steffen et al., 2007b). We chose the anticancer drug camptothecin as the guest molecule and aimed at the identification of β-cyclodextrin based synthetic receptors.

This project was accomplished in collaboration with the group of Dr. Joannis Apostolakis from the Ludwigs-Maximillian University in Munich and the group of Professor Dr. Gerhard Wenz from the Saarland University in Saarbrücken. All experimental work was done by Caroline Thiele and Dr. Christian Strassnig.

3.1 Introduction

Synthetic receptors are molecules that specifically bind guest molecules. As mentioned in Section 1.1, they generally cannot rival proteins in terms of binding affinity and specificity. However, they do exhibit numerous technical advantages over their natural counterparts (see Section 1.1) (Schrader & Hamilton, 2005). β-cyclodextrins have particularly proven to be in high demand within the pharmaceutical industry since their cavity is appropriate for binding druglike molecules (Davis, 2004). Due to the industrial relevance, rational approaches for tailored synthetic receptor design are of great current interest. As detailed in Section 1.2 virtual screening using protein-ligand docking tools is well established in the field of computer-aided drug design. In this field, virtual screening is applied in order to identify novel ligands for a given protein target (see Section 1.2.2) with the potential of being used as drugs. However, the application of virtual screening as a method for the design of novel synthetic host-guest complexes is entirely new. Recently, studies have been published in which protein-ligand docking methods, borrowed from the field of drug design, were applied to synthetic host-guest systems, aiming to identify optimally interacting systems. De Jong et al. (2002) per-

formed a virtual screening for novel guest molecules of a β-cyclodextrin dimer by means of the protein-ligand docking tool DOCK (Ewing & Kuntz, 1997) (de Jong et al., 2002). Docking was applied to place energetically minimized ligand structures into the β-cyclodextrin dimer. In this way, about 110,000 substances were virtually screened. 30 of the manually inspected top-ranking molecules were proposed for further experimental verification. Despite the fact that the docking tool neglected conformational flexibilities of both, the host and the guest molecule, nine out of 30 proposed molecular guests were found to bind to the receptor with high affinity. Corbellini et al. applied a similar approach when searching for guest molecules of a molecular capsule (Corbellini et al., 2004). From about 30,000 virtually screened substances, a restricted number of the computationally predicted binders were selected for experimental testing, which led to five compounds demonstrating strong encapsulation as tested by NMR. These two approaches demonstrate the potential and the possible impact of structure-based virtual screening methods from drug design for the optimization of synthetic host-guest systems. However, the more demanding issue is to look for a receptor that will bind a given guest molecule with high affinity and specificity. While this is clearly more difficult to address, it appears also to be the more relevant for technical applications, such as complexation and controlled delivery of drugs. This approach is referred to as inverse virtual screening as the docking direction is inverted in comparison to common virtual screenings in drug design (Shoichet, 2004).

3.1.1 Camptothecin and Topoisomerase I

For this study, we chose camptothecin as the guest molecule for the design of a tailored receptor by means of inverse virtual screening. Camptothecin (see Figure 3.2) and its derivatives represent a class of antineoplastic agents with a broad spectrum of activity against several types of cancer including colorectal and ovarian cancer (Takimoto et al., 1998). On the molecular level, this class of drugs inhibits topoisomerase I (see Figure 3.1). Topoisomerases are nuclear enzymes that play an important role in DNA replication and in transcription and recombination (Slichenmyer et al., 1993). The enzymes catalyze a three-step process in which they alter the linking of DNA. First, they cut one or both strands of DNA. Second, they allow the passage of a segment of DNA through this break. Finally, they reconnect the DNA break. The topoisomerase of type I only cuts one strand of DNA, whereas the second type can cleave both strands (Berg et al., 2002). The three-dimensional structure of topoisomerase I is shown in Figure 3.1. Topoisomerase I consists of four domains, arranged around a central cavity that

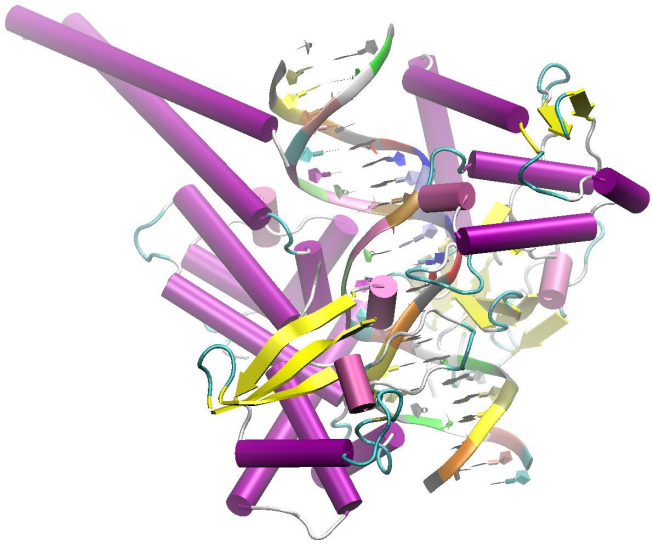


Figure 3.1. The structure of human topoisomerase I in complex to DNA (PDB reference code: 1A36).

binds the double-stranded DNA molecule. Camptothecin and derivatives inhibit topoisomerase I by blocking the reconnection step of the cleavage reaction. This results in an accumulation of a covalent reaction intermediate which is presumed to cause cell death in the S-phase of the cell cycle (Liu et al., 2000). Since the rate of cell replication in tumors is clearly higher than in normal tissues, camptothecin can be applied to halt the tumor growth, whereas the nontumor cells are not strongly affected.

3.1.2 Pharmaceutical Formulations for Camptothecin

Unfortunately the high therapeutic potential of camptothecin is hampered by its low solubility and stability (Lundberg, 1998). Only the closed lactone form is active *in vivo* (see Figure 3.2, left). In relevant literature attempts are described that

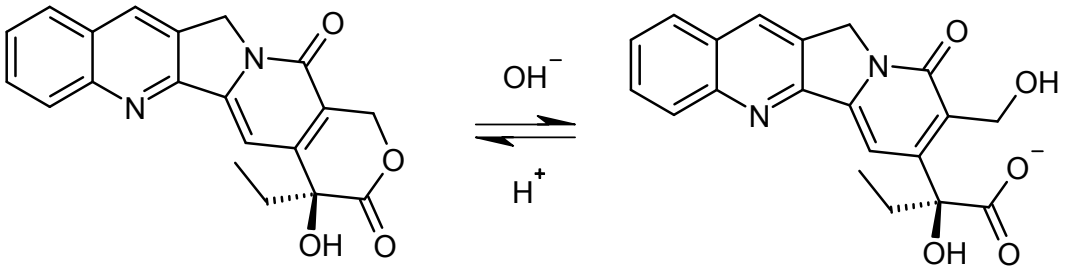


Figure 3.2. Lewis structure of camptothecin in the lactone and carboxylate form.

help circumventing these difficulties by means of pharmaceutical formulations: Lundberg (1998) synthesized oleic acid esters of camptothecin analogs, which can be inserted into liposomes or submicron lipid emulsions. This camptothecin formulation proved to be very stable against the lactone ring opening and, moreover, the cytotoxic activity was retained. Polyethylene glycol-conjugated camptothecin derivatives were synthesized by Conover et al. (1998). These derivatives are water-soluble prodrugs of camptothecin. Cytotoxic activity could be shown in mouse models. It was suggested that this soluble transport form of camptothecin could have a clinical application. Ertl et al. (1999) developed microspheres using poly-(D,L-lactide-co-glycolide) as a building block. These microspheres were loaded with camptothecin. The study showed that the active lactone form of camptothecin was maintained during preparation. Furthermore, a sustained release of camptothecin was achieved, which reduces local toxicity and prolongs efficacy. Kang et al. (2002) introduced the use of β-cyclodextrin and derivatives as solubilisants for camptothecin. Their formulation significantly increased solubility and stability, and motivated the work presented in this chapter.

3.1.3 Cyclodextrins and Inclusion Complexes

Cyclodextrins are among the most relevant synthetic host molecules for aqueous solutions (D'Souza & Lipkowitz, 1998). These molecules are cyclic oligomers of α-D-Glucose. Basically, four types of cyclodextrins can be distinguished, namely α-, β-, γ- and δ-cyclodextrins that correspond to 6, 7, 8 or 9 α-D-glucose units (see Figure 3.3). All are approximately C_n symmetric (Sakurai et al., 1990), with n equal to the number of glucose units. The production of cyclodextrins is conducted by enzymatic degradation of amylose. For this purpose mainly cyclodextrin glycosyl-transferases (CGTases) of bacteria are used (Biwer et al., 2002). This enzymatic conversion, however, is unspecific regarding the produced ring size of the cyclodextrins and therefore, the single homologs are purified with selective media for precipitation (Cramer & Henglein, 1958; Schmid, 1991).

The shape of cyclodextrins has been described as torus- or doughnut-like (see Figure 3.3). The narrower side is called the primary side, since all primary hydroxyl groups[1] of the glucose units point to this side. The wider side is called the secondary side, reflecting the fact that the secondary hydroxyl groups[2] of the glucose units are located at this side (see Figure 3.3). Cyclodextrins have a strong dipole moment because twice as many hydroxyl groups are located on the secondary side than on the primary side (Kitagawa et al., 1987; Sakurai et al., 1990).

[1] Hydroxy groups bound to a carbon atom which is bound to one carbon atom
[2] Hydroxy groups bound to a carbon atom which is bound to two carbon atoms

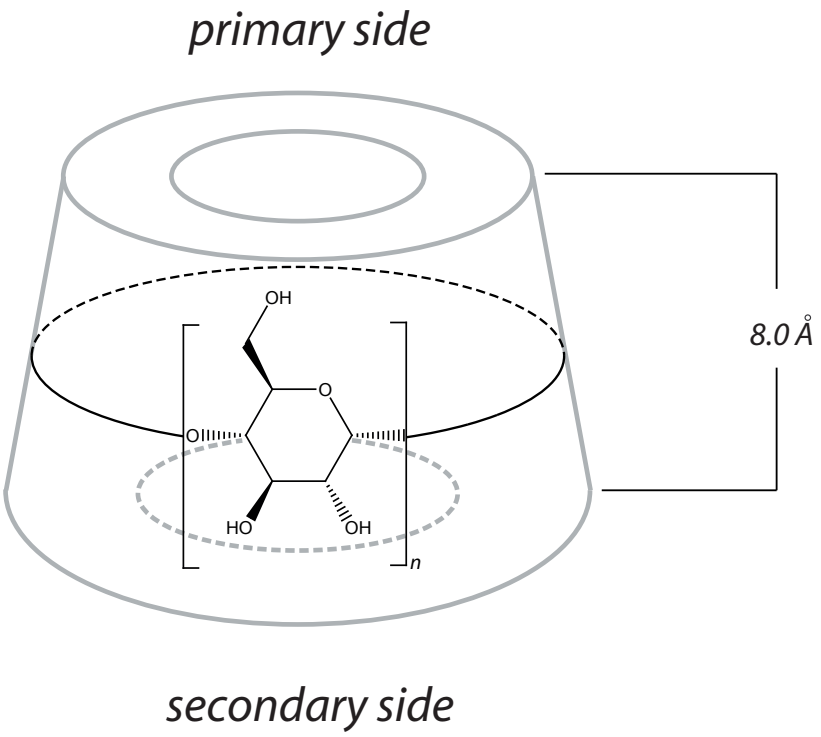


Figure 3.3. Schematic illustration of a cyclodextrin. The number n of α-D-glucose units is equal to 6 for α-, 7 for β-, 8 for γ and 9 for δ-cyclodextrins, respectively.

β-cyclodextrin exhibits so-called flip-flop hydrogen bonds between secondary hydroxyl groups of neighboring glucose units (see Figure 3.4). This restricts the flexibility of the β-cyclodextrin core (Saenger et al., 1983; Betzel et al., 1984). Cyclodextrins possess a cavity within the molecule. The exterior of the cyclodextrins, which is mainly influenced by the hydroxyl groups, is hydrophilic, whereas lipophilic interactions dominate the cavity. This property enables cyclodextrins to form relatively strong host-guest complexes with hydrophobic guests. Due to the hydrophilic nature of the cyclodextrin's exterior, these complexes are soluble in water and thus cyclodextrins act as solubilisants (Wenz, 1994). Furthermore, the inclusion into the cavity can increase the stability of the guest molecule for example against chemicals (Ong et al., 1997), biochemical influences (Brown et al., 1993), or photochemical reactions (Szejtli, 1984). The driving force of complex formation is a combination of hydrophobic interactions, van-der-Waals interactions, dipole-dipole interactions, hydrogen bonds, and changes in the solubilization of the guest molecule and the cyclodextrin cavity (Connors, 1997). The release of water molecules from the cavity is entropically favorable. In general, the higher the shape complementarity of the cyclodextrin to the guest molecule, the lower the binding free energy of the complex. The size of the molecules that can be bound

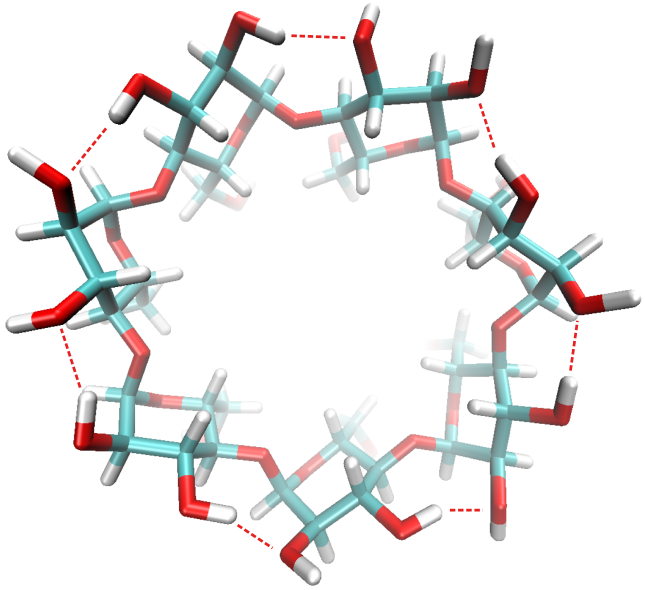


Figure 3.4. View onto the secondary side of a β-cyclodextrin. The structure was taken as is from the CSD (CSD-ID: BUVSEQ03) (Zabel et al., 1986). The red dashed lines denote hydrogen bonds between secondary hydroxyl groups of neighbouring glucose units.

by a particular type of cyclodextrins, however, increases with the size of the cavity. α-cyclodextrins bind alkyl-chains, whereas benzene is already too large for them. The cavity of β-cyclodextrins can complexate more bulky molecules such as adamantane, naphtaline, or various benzene derivatives. γ-cyclodextrins can bind annelated ring systems and even buckyballs up to the size C_{60}. The ability of cyclodextrins to bind molecules of a particular size has been referenced as size recognition (Müller & Wenz, 2007; Wenz et al., 2006a,b). Selectivities and affinities of cyclodextrins can be further increased by means of chemical modifications (Kitae et al., 1998). Particularly β-cyclodextrins are predestined to bind drug-size molecules. Their hydrophobic cavity together with the hydrophilic exterior designates their application as solubilisants for small hydrophobic molecules such as drugs (Connors, 1997). Today, several β-cyclodextrins based drug formulations are on the marketplace (Fenyvesi et al., 1984b; Davis, 2004). These include, for example, formulations for furosemide (Fenyvesi et al., 1984a), prostaglandines (Stuerzebecher et al., 1996), diclofenac (Fugen & Cuijing, 1998), tumor necrosis factors (Stanton & Vincent, 2001), piroxicam (Banerjee et al., 2004), or camptothecin (Kang et al., 2002). Other industrial applications were reported in the food-industry, where cyclodextrins have been used to protect flavors or vitamins from oxidation (Szejtli, 1980).

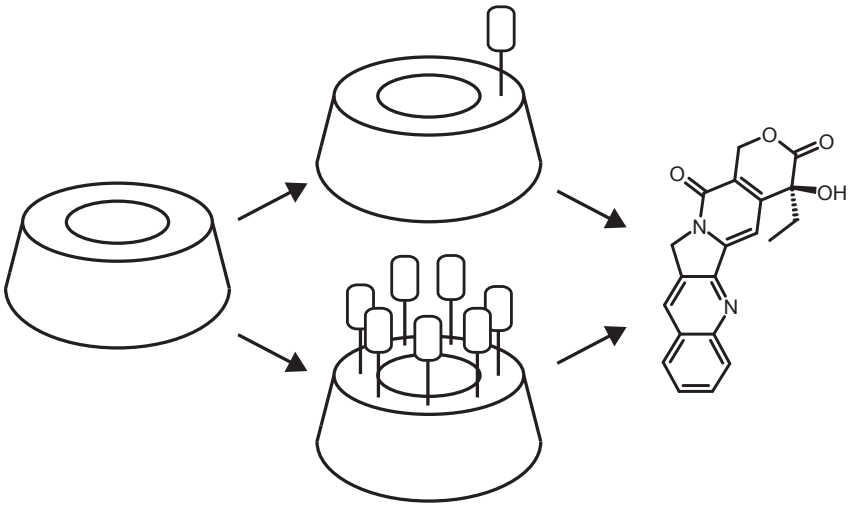


Figure 3.5. Mono and heptakis β-cyclodextrin derivatives are generated. All candidate receptors are sequentially docked onto camptothecin.

3.2 Aim of the Study

In this study we focus on the computer-assisted development of β-cyclodextrin derivatives with a high affinity to the anti-cancer drug camptothecin by means of inverse virtual screening. Top-ranking candidate receptors were synthesized and experimentally tested. The study is considered as a proof principle for the applicability of docking tools for a computer-aided optimization of synthetic receptors.

3.3 Methodology

First, a library of candidate receptors was defined (virtually synthesized). We assured synthesizability by choosing a simple synthesis scheme of a well established reaction for modifying the β-cyclodextrin core(see Section 3.3.4 and Figure 3.5). The synthesis consists of nucleophilic displacement reactions of 6-O-iodo- or 6-O-tosyl-β-cyclodextrin by a set of thiols (Karginov et al., 2006). Each member of this virtual library is then sequentially docked onto camptothecin. This approach is referred to as inverse virtual screening. Similar to the normal virtual screening scenario, we applied scoring functions for ranking the different candidates. Selected top-ranking candidate receptors were synthesized and experimentally tested. See Figure 3.6 for a general overview.

3.3.1 Preparation of Camptothecin

The crystal structure of a iodoacetyl derivative of camptothecin was obtained from the Cambridge Structural Database (Allen, 2002) (ID: CAMPTC10) (McPhail

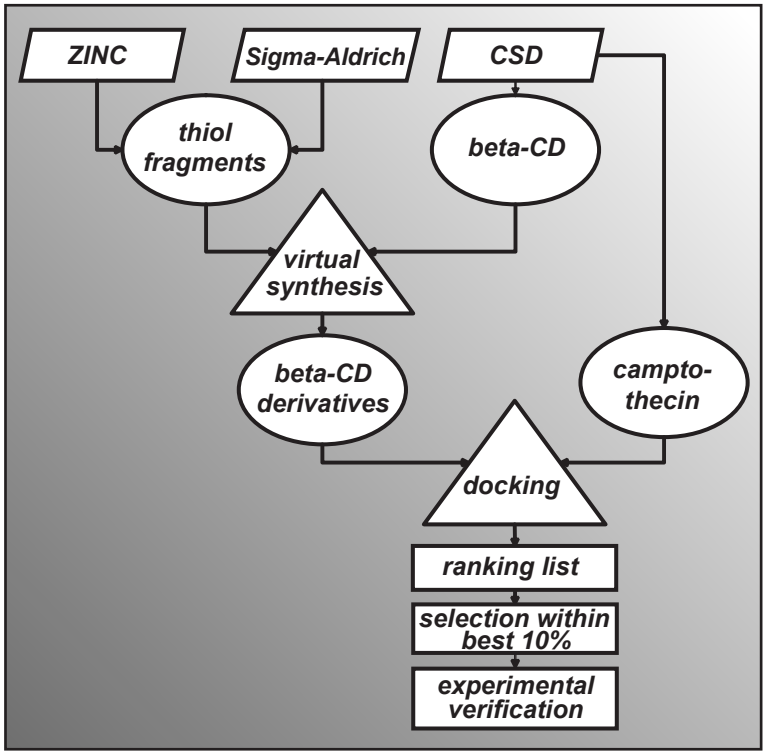


Figure 3.6. Design of the study.

& Sim, 1968) and exported as a MOL2-file. The iodoacetyl-group was replaced by a hydrogen atom to construct the unmodified camptothecin molecule. All missing hydrogen atoms were added with SYBYL 6.7. Subsequently a force field optimization was performed with the MMFF94S force field (Halgren, 1999a) until gradient convergence (0.005 kcal mol^{-1} Å^{-1})). The structure was saved in the MOL2 file format.

3.3.2 Preparation of the β-Cyclodextrin Core Structure

The crystal structure of β-cyclodextrin was obtained from the CSD (ID: BUVSEQ03) (Zabel et al., 1986) and exported as a MOL2 file. The structure is derived from neutron diffraction. All deuterium atom positions were resolved. In the case of a disordering of an atom over two sites of almost equal occupancy only one position has been retained. The atom types of deuterium atoms were changed to hydrogen atoms. All remaining atom types were visually inspected and - if necessary - corrected according to the SYBYL atom type rules. All water molecules present in the crystal structure were manually removed.

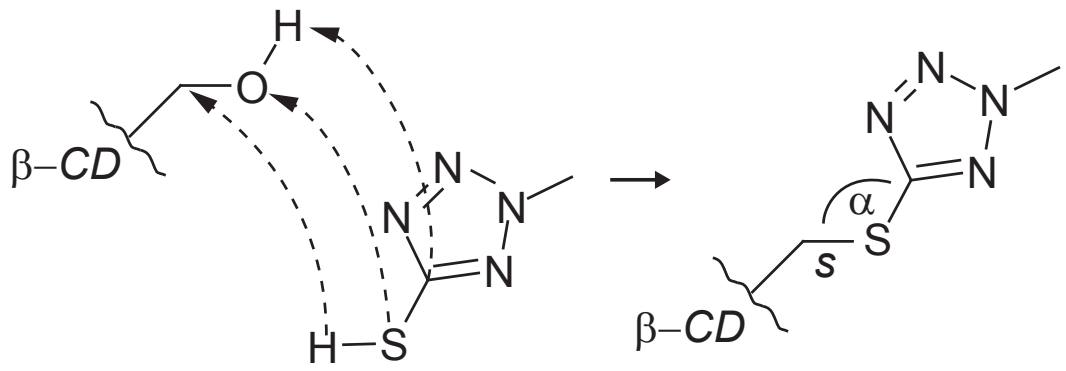


Figure 3.7. Schematic drawing of the virtual synthesis of the β-cyclodextrin-library. The thiol-group (C-S-H) of the fragment is superimposed onto one (mono) or respectively all seven (heptakis) primary hydroxyl groups (C-O-H). The bond length s and the bond angle α are set to standard values ($s = 1.82$Å, $\alpha = 109°$). Apolar hydrogen atoms are omitted for clarity.

3.3.3 Extraction and Preparation of Fragment Libraries

We extracted all compounds from the Sigma-Aldrich catalog and the ZINC database (Irwin & Shoichet, 2005) that contained at least one thiol group by means of the substructure search interfaces provided on the web pages of the suppliers. As an additional filter we set the molecular weight range to 0 - 200 $\mathrm{g\,mol^{-1}}$ in order to limit the size of the fragments. Sigma-Aldrich provides compounds as STRUCTURE DATA FILES (SDF) (Dalby et al., 1992), whereas from ZINC MOL2-files can directly be downloaded. The SD-files from Sigma-Aldrich were converted to MOL2-files and, for each molecule, a low-energy conformation was generated with CORINA (Sadowski & Gasteiger, 1993). Subsequently we removed all compounds with more than one thiol group. This was done in order to secure non-ambiguous synthesis. Altogether we obtained 605 fragments from ZINC and 318 fragments from the Sigma-Aldrich catalog, respectively. In the last step reasonable protonation states were assigned to all fragments, i. e., acidic groups are deprotonated, amines are protonated when they are not in conjugation with an aromatic system.

3.3.4 Virtual Synthesis of β-Cyclodextrin Derivatives

The virtual library of β-cyclodextrin derivatives was defined with the help of a PYTHON script. This script virtually synthesized mono- and heptakis-substituted β-cyclodextrins for each of the thiol group fragments described in the previous section (see Figure 3.7). To do so, the script transforms each fragment in three-dimensional space such that its thiol group is superimposed onto one or respectively all seven primary hydroxyl groups of the β-cyclodextrin structure. Then

the hydroxyl group and the hydrogen atom of the thiol group are removed and a bond of standard length is added between the sulfur atom of the fragment and the carbon atom of the β-cyclodextrin. Finally, the bond angle α is set to 109°. This type of construction guaranteed correct bond lengths and angles, while rotatable torsion angles were optimized during docking. Altogether 1,846 mono- and heptakis-substituted β-cyclodextrin derivatives were generated with this procedure.

3.3.5 Applied Docking Tools

The system of choice is dominated by hydrophobic interactions, which cannot be handled by our tool FlexR (see Chapter 2). Due to this limitation we chose two other docking tools that had proved effective in handling hydrophobic interactions and tailored them for our needs. The applied docking tools AutoDock and GlamDock are detailed below.

AutoDock

AutoDock (Morris et al., 1998) (Version 3.05) is an open-source software package for the automated docking of ligands into macromolecules. It has been successfully applied in some recent virtual screening projects, for example for the discovery of protein phosphatase 2C inhibitors (Rogers et al., 2006), for DNA minor groove binders (Evans & Neidle, 2006), and for anti-SARS drugs (Wei et al., 2006).

The search for the conformation of a ligand with minimal binding energy is regarded as an optimization problem. In AutoDock four optimization algorithms are implemented out of which the Lamarckian genetic algorithm has been shown to be the most effective and reliable (Morris et al., 1998).

Genetic algorithms imitate evolutionary processes, for finding the global optimum of a given optimization problem (Michalewicz, 1996). In protein-ligand docking, the configuration of the ligand to the protein can be described by a set of values, which define the translation, orientation, and conformation of the ligand within the protein binding site. Each of these variables is referenced as a state variable of the ligand and is coded into a virtual gene. The atomic coordinates can be translated from the virtual genome of the ligand and correspond to the phenotype. In the beginning of the evolutionary process, a set of randomly distributed ligands is generated as the starting generation. Two genetic processes allow for evolution: first, crossover recombinations of the genes of two ligands and second, mutations, in which a gene changes its value by a random amount. Based on predicted fitness scores only a restricted number of individuals is selected from

the generated offsprings for the next evolution step. These fitness scores are calculated by means of the energy function described below for a given state of the ligand. For improving the efficiency of the search performance, Morris et al. have extended the classical genetic algorithm by incorporating a local search method (Morris et al., 1998). This variant is called Lamarckian genetic algorithm and allows for the improvement of an individual's fitness by local search optimization of the phenotype, i. e. the atom coordinates corresponding to the genetic state. The optimized position is mapped back into genome.

AutoDock uses a grid-based energy evaluation procedure which speeds up the energy calculation for a given ligand state. For this purpose, a cubic grid with a user adjustable grid size is virtually overlaid onto the binding site of a protein. A set of representative probe atoms are iteratively put onto each grid point. An energy function is used to precalculate an interaction energy for each probe atom on each grid point with the protein's binding site. These precalculated interaction energies are stored in a look-up table which can be rapidly accessed during docking time. The applied energy function is given in equation 3.1.

$$\begin{aligned}\Delta G = &\Delta G_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ &+\Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right) \\ &+\Delta G_{elec} \sum \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \\ &+\Delta G_{tor} N_{tor} \\ &+\Delta G_{sol} \sum_{i_C j} S_i V_j e^{(\frac{-r^2}{2\delta^2})} \end{aligned} \tag{3.1}$$

The five ΔG coefficients[3] are derived from a linear regression analysis on a set of 30 protein-ligand with experimentally determined binding free energies. The terms iterate over all ligand-protein atoms pairs, and furthermore over all pairs of ligand atoms that are more than two bonds apart from each other. The van-der-Waals and the hydrogen bond energies are described by Lennard-Jones potentials. In the case of the hydrogen bond energy a weight $E(t)$ penalizes deviations from ideal bond angles and lengths between the hydrogen bond donor and the acceptor atom. The Coulombic electrostatic potential calculates electrostatic interaction. The number of single rotatable bonds estimates entropic effects. The last term

[3] vdW = van-der-Waals interaction, hbond = hydrogen bond, elec = electrostatic interactions, tor = torsional energy, sol = solvation

implements the calculation of desolvation energies based on the work of Stouten et al. (1993).

GlamDock

The second tool we applied is called GlamDock (Karasz et al., 2004; Tietze & Apostolakis, 2007). GlamDock was validated on benchmark sets from the literature (Kellenberger et al., 2004) and was shown to perform better than state-of-the-art methods on the Kellenberger dataset (Tietze & Apostolakis, 2007).

The current version of GlamDock relies on a Monte Carlo procedure based on the matching of functional groups of the ligand with compatible interaction spots in the binding site. Furthermore, GlamDock uses a local minimization approach (Abagyan et al., 1994; Apostolakis et al., 1998). The search space for the Monte Carlo optimization of the ligand consists of continuous degrees of freedom for the conformation of the ligand and discrete degrees of freedom that link interaction groups of the ligand with interaction spots in the protein binding site. These interactions spots are precalculated by means of interaction probes and the energy function described below. The approach used for this precalculation is in some respects similar to the Protomol procedure in Surflex (Jain, 2003) and works as follows. Compatible interaction probes are uniformly placed around each interaction group of the protein binding site in ideal interaction geometry. The energy function (see below) scores the placed probes. The probes are clustered in order to reduce the number of interaction spots. Only the k best-scoring representatives of each cluster are retained and indexed. Each interaction group of the ligand is assigned to the indices of compatible probes. A given point in the search space is translated into a ligand conformation: first, by the adjustment of the torsion angles, and second, by the rigid placement of the ligand in a manner that optimally fulfills mapped interactions. In this step, the algorithm employs Kabsch rotations (Kabsch, 1976). These Kabsch rotations minimize the distance between interaction groups of the ligand and mapped interaction spots in the binding site. After each rotation GlamDock removes unfulfilled mappings from the list of targeted interactions and the procedure is reiterated. Subsequently GlamDock performs a torsion space minimization with the full energy function in order to relax the conformation in the field of the receptor structure. The remaining (fulfilled) mappings are coded back into the search space point, which led to the particular placement. The best ranking conformation is predicted as most favorable structure of the complex.

The integrated energy function ChillScore for the optimization is a continuous-gradient approximation to the docking version of ChemScore (Baxter

et al., 1998; Eldridge et al., 1997; Verdonk et al., 2003).

$$\begin{aligned}\Delta G_{Chill} = \Delta G_0 + f_{hbond}\Delta G_{hbond} + f_{lipo}\Delta G_{lipo} + f_{metal}\Delta G_{metal} \\ +\Delta G_{rot}N_{rot} + \Delta G_{clash} + \Delta G_{pocket}\end{aligned} \tag{3.2}$$

It consists of seven terms that are summed up to the entire energy. These terms consist of a hydrogen bond term ($f_{hbond}\Delta G_{hbond}$), a lipophilic term ($f_{lipo}\Delta G_{lipo}$), an acceptor-metal interaction term ($f_{metal}\Delta G_{metal}$), an entropic term that accounts for the ligand flexibility ($\Delta G_{rot}N_{rot}$), a clash (atom-atom overlap) term for ligand-protein and intra-ligand atoms (ΔG_{clash}), and a term that penalizes poses in which the center of geometry of the ligand is outside the defined binding pocket (ΔG_{pocket}).

3.3.6 Docking Protocols

As previously described, protein-ligand docking tools explore the conformational space of a ligand within the binding site of a protein. Most state-of-the-art tools tackle the protein as rigid during the simulated binding. In our work, we could not make this assumption, since the conformations of the virtually generated β-cyclodextrin derivatives were unknown and had to be generated during the docking process. Camptothecin, however, is a relatively rigid molecule, whose conformational space can be reasonably described by a single conformation. Due to the enormous combinatorial complexity of the protein's flexibility, today's docking-tools cannot handle the complete conformational space of a protein (see Section 2.1.1). The conformational sampling of a synthetic receptor is, however, computationally feasible during docking, since synthetic receptors are generally only slightly larger than guest molecules. This suggests that we can interchange the roles of ligand and receptor and perform an inverse docking of the receptor onto the rigid guest molecule. In this particular case, we can further reduce the conformational space of the host (β-cyclodextrin): the secondary hydroxyl groups form so-called flip-flop hydrogen bonds to hydroxyl groups of neighboring glucose units and restrict the flexibility of the macrocycle (see Figure 3.4). Only the side-chains on the primary side exhibit a considerable degree of torsional freedom. Hence, the conformational search was performed for the β-cyclodextrin derivatives, whereas camptothecin was kept rigid in both docking tools.[4]

[4] β-cyclodextrin is, however, a particular case. The rigid treatment of the macrocycle is not necessarily reasonable in the case of other macrocyclic host molecules.

AutoDock

Since AutoDock only provides solvation parameters for amino acid atom types we reasonably mapped the atoms of camptothecin onto corresponding amino acid atoms. We used the AutoDock Tools to generate grid maps for camptothecin (Sanner, 1999). We therefore defined a $50{\cdot}50{\cdot}50\,\text{Å}^3$ cube around camptothecin. The grid spacing was set to 0.375 Å. For each docking run the standard AutoDock parameters were used. We only increased the number of energetic evaluations to 5 millions and the number of genetic algorithm runs (GA runs) to 100. The maximal possible number of torsions was set to 30. For the conformational optimization we chose the Lamarckian genetic algorithm. For the energy evaluation in docking and ranking we used the dock score.

GlamDock

The docking protocol consists of five single docking runs each consisting of 650 Monte Carlo minimization (MCM) steps, with 15 steps of Levenberg-Marquardt (Deo & Walker, 1995) minimization in torsion space (Bystroff, 2001) at each MCM step. A maximum of 40 poses are finally post-minimized using 150 steps of Levenberg-Marquardt. The scoring function for docking considers the energies of the receptors, whereas for ranking a size penalizing variant of the scoring function without internal energy was used. In contrast to AutoDock, GlamDock does not constrain the receptor to dock around the ligand. It explicitly allows conformations where the ligand lies on top of the β-cyclodextrin ring (see Figure 3.11). Such conformations are mainly stabilized by the internal energy of the receptor, which on average scales quadratically with its size (the number of atoms). The size penalty has the effect of identifying more specifically interacting complexes and does not necessarily correlate with binding affinity. The reason for the size penalty was that initially both virtual screening results contained mainly large hydrophobic receptors on top ranks.

3.4 Results

We generated a virtual library (1,846 entities) of 6-O-mono- and 6-O-heptakis-substituted β-cyclodextrin derivatives from the β-cyclodextrin core and thiol building blocks. The structure of the complexes between camptothecin and the different derivatives was predicted using the two docking tools and the derivatives were ranked according to the score[5] of the complex.

[5] Scores are used as heuristic estimates of ΔG°. By convention, the lower a score, the more favorable the interaction. Thus, the first rank corresponds to the complex exhibiting the lowest score.

Usually, protein-ligand docking tools explore the conformational space of ligands, e.g., drug molecules, while treating the protein as rigid (in its crystal structure conformation) during the simulated binding. For our work this simplification was not appropriate and the conformations of the virtually generated β-cyclodextrin derivatives had to be generated during the docking process.

For AutoDock docking and ranking was performed based on the overall score for AutoDock (dock score). In the case of GlamDock the scoring function for docking considers the internal energies of the receptors, whereas for ranking a size penalizing variant of the fitness score for GlamDock without internal energy was used.

For the experimental verification we considered only compounds, which were found by at least one docking tool within the top 10% of the respective ranking lists. All potential candidates were visually inspected. We selected promising β-cyclodextrin derivatives for synthesis and further experimental investigation (see Table 3.1). Furthermore, the building blocks for synthesizing the β-cyclodextrin derivative had to be commercially available. Interestingly, AutoDock favored β-cyclodextrin derivatives with aromatic and hydrophobic side-chains, whereas GlamDock mainly suggested derivatives forming hydrogen bonds to camptothecin (see Table 3.1). The predicted affinity scores for the heptakis derivatives were generally more favorable than for the corresponding mono derivatives for both docking programs (see Table 3.2).

Nine heptakis-substituted β-cyclodextrin derivatives were synthesized by nucleophilic displacement reactions in good yields. For a closer investigation of the molecular interactions and for obtaining an estimate of binding affinities of the insoluble heptakis-substituted β-cyclodextrins we also synthesized the nine corresponding mono derivatives, which were all soluble in water. Furthermore the heptakis-substituted thiosulfate β-cyclodextrin (compound **20**) was synthesized. This molecule had an unfavorable predicted binding energy in docking and served as a negative test.

The binding constants K for all synthesized β-cyclodextrin derivatives were determined from the solubility isotherm. The remarkable increase of the solubility of camptothecin as the function of the concentration of the β-cyclodextrin derivatives is demonstrated in Figure 3.8. The binding constants K were derived from the slope (Kang et al., 2002). In order to assure comparability we additionally measured the binding constants of the native β-cyclodextrin, hydroxypropyl-β-cyclodextrin (HP-β-cyclodextrin) and randomly methylated β-cyclodextrin (RDM-β-cyclodextrin), which were already investigated by Kang et al. (2002). It should be noted that the value of K obtained for RDM-β-cyclodextrin (186 M^{-1})

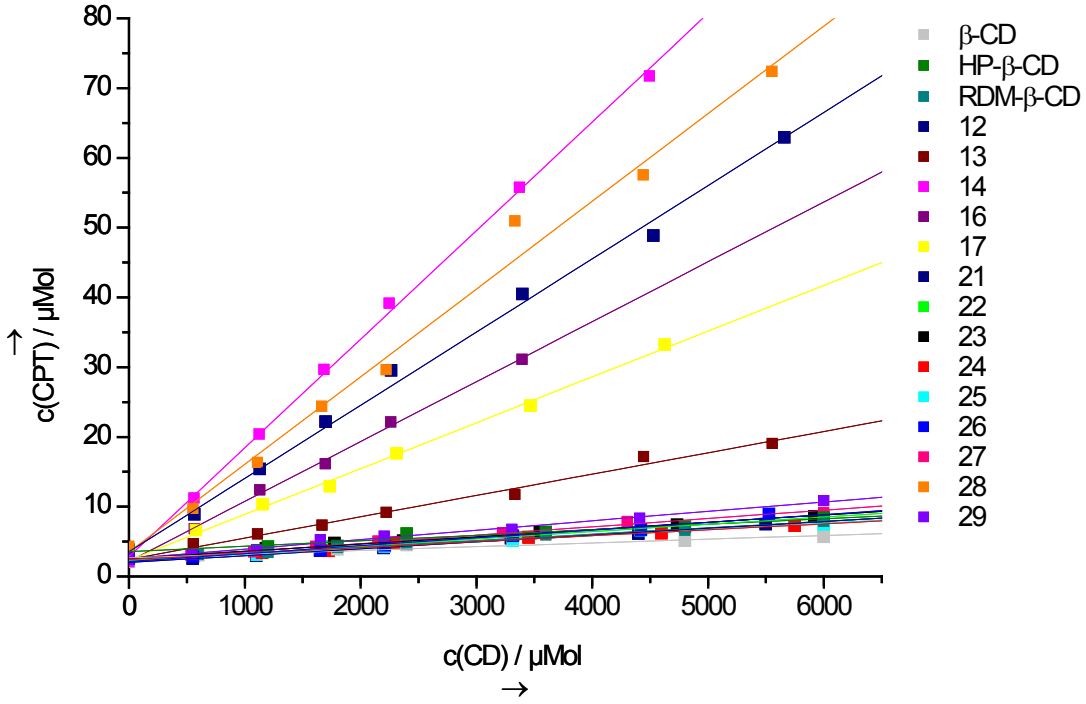


Figure 3.8. Dependence of the solubility of camptothecin on the concentration of the β-cyclodextrin derivatives.

significantly differed from its literature value (909.7 M^{-1}) (Kang et al., 2002). This difference might be caused by different experimental protocols and different substitution patterns of the randomly methylated β-cyclodextrin.

Out of the nine synthesized receptors five exhibit binding constants K clearly superior to the ones of the native β-cyclodextrin and the two other known β-cyclodextrin derivatives from Kang et al. (2002) (see Table 3.2 and Figure 3.8). Heptakis[6-deoxy-6-(2-sulfanyl-ethane-sulfonic acid)]-β-cyclodextrin (compound **14**) showed the highest value of K with 7,496 M^{-1}. Since receptors **11**, **15**, **18** and **19** were insoluble in water, also the corresponding mono derivatives (compounds **21**-**29**) were investigated. Among them, mono-[6-deoxy-6-(6-sulfanyl-9H-purine)]-β-cyclodextrin (compound **28**) showed the strongest binding affinity with 3,629 M^{-1}, which is in the range of the heptakis-substituted β-cyclodextrin derivatives. As predicted, the negative test example (compound **20**) exhibits a comparably low binding affinity with $K = 370\ M^{-1}$. A comparison of the binding free energies ΔG° values for the majority of the mono-substituted β-cyclodextrin derivatives **21** - **27** and **29** with unsubstituted β-cyclodextrin reveals, that one building block causes a stabilization energy of about ΔG°= -2 kJ mol^{-1} . The same comparison for the heptakis-substituted β-cyclodextrins **12** - **17** show a decrease of binding energy of ΔG°= -$(5 \cdots 9)$ kJ mol^{-1} due to seven building blocks. This may suggest

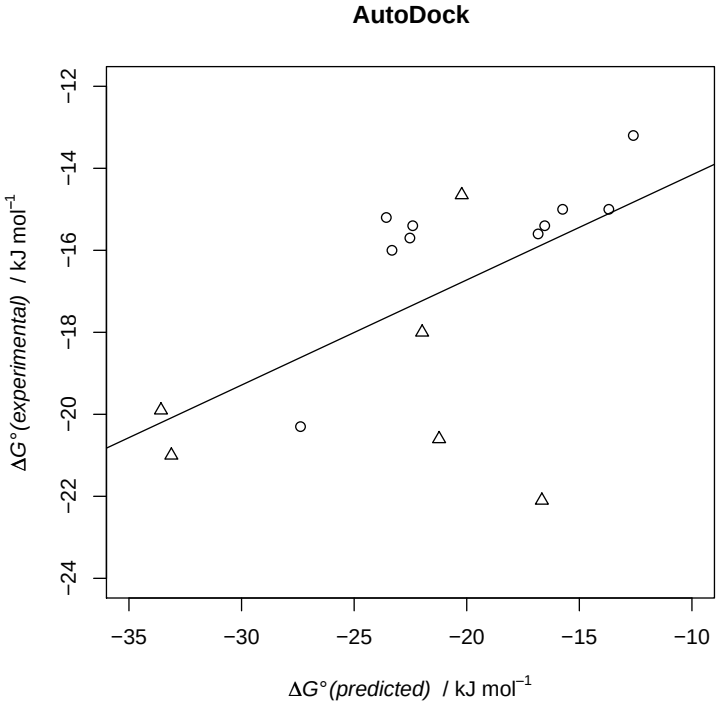
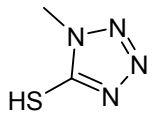
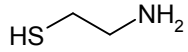
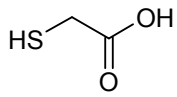
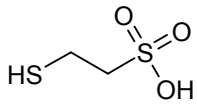
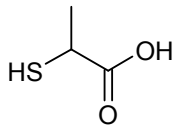
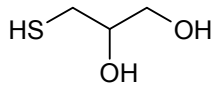
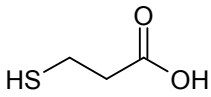
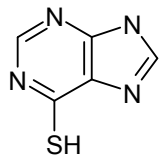
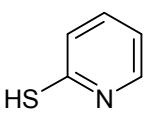


Figure 3.9. The predicted binding energies (AutoDock) are plotted against the experimental binding free energy. The mono derivatives are depicted by circles, the heptakis derivatives are shown as triangles.

an additive effect on binding exerted by only three to four building blocks and may be due to steric barriers. Remarkably, one 6-sulfanyl-9H-purine building block in compound **28** leads to an exceptionally strong stabilization of ΔG°= -7 kJ mol^{-1}.

In Figures 3.9 and 3.10 we plot the binding energies predicted by the AutoDock and the GlamDock scoring functions against the experimentally determined values. The heptakis derivatives are shown as triangles, the mono derivatives as circles. HP-β-cyclodextrin and RDM-β-cyclodextrin were not considered since no docking was performed due to the structural uncertainties (random substitution); for compounds **11**, **15**, **18** and **19** no binding free energy could be experimentally determined due to insolubility in water. The correlation coefficient for AutoDock is equal to r=0.57 (residual standard error of the regression = 2.4 kJ mol^{-1}), for GlamDock equal to r=0.82 (residual standard error of the regression = 1.6 kJ mol^{-1}). Compound **14** is an obvious outlier for both docking tools, but particularly in the case of AutoDock. If this compound is omitted, the correlation coefficient for AutoDock increases to 0.78 (residual standard error of the regression = 1.7 kJ mol^{-1}).

Table 3.1. Building blocks selected by virtual screening of corresponding β-cyclodextrin derivatives.

ID	IUPAC Name	Lewis Structure	CAS-No.	Mono ID	Hepta ID	Tool [1]
1	1-methyltetrazole-5-thiol		13183-79-4	**21**	**11**	AD
2	2-aminoethanethiol		60-23-1	**22**	**12**	GD
3	2-mercaptoacetic acid		68-11-1	**23**	**13**	GD
4	2-mercaptoethanesulfonate		3375-50-6	**24**	**14**	GD
5	2-mercaptopropanoic acid		79-42-5	**25**	**15**	GD
6	3-mercaptopropane-1,2-diol		96-27-5	**26**	**16**	GD
7	3-mercaptopropanoic acid		107-96-0	**27**	**17**	GD
8	9H-purine-6-thiol		50-44-2	**28**	**18**	AD
9	pyridine-2-thiol		2637-34-5	**29**	**19**	AD/GD

[1] AD = AutoDock, GD = GlamDock

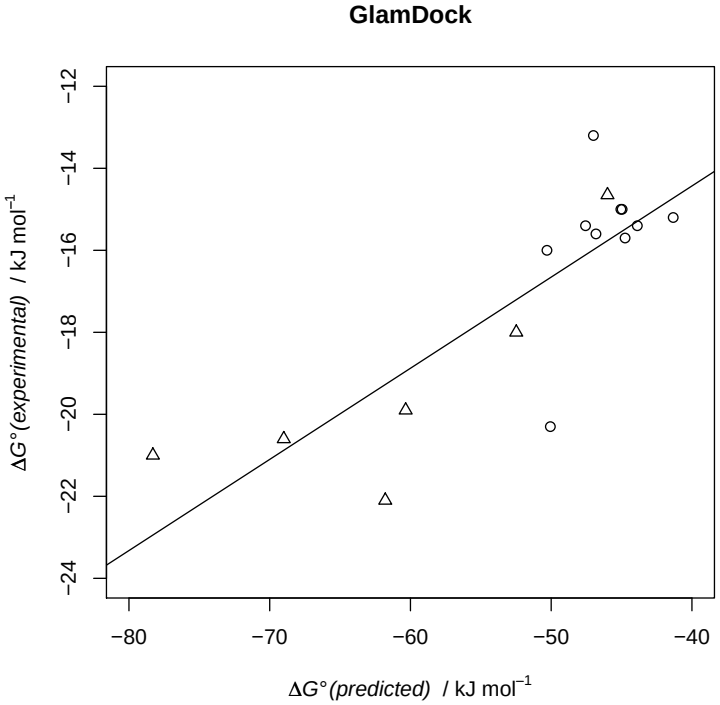


Figure 3.10. The predicted binding energies (GLAMDOCK) are plotted against the experimental binding free energy. The mono derivatives are depicted by circles, the heptakis derivatives are shown as triangles.

3.5 Discussion

Due to the flexibility and the larger size of synthetic receptors in comparison to ligands, virtual screening of receptors (inverse screening) is more complex in general than the conventional virtual screening of ligands (de Jong et al., 2002). For a given complex, the predicted binding free energy ΔG°(predicted) consists of three components, in principle:

$$\Delta G^0(predicted) = \Delta G^0_R + \Delta G^0_L + \Delta G^0_{RL} \tag{3.3}$$

where ΔG^0_R is the change of energy in the receptor molecule, ΔG^0_L is the change of energy in the ligand upon complexation and ΔG^0_{RL} is the interaction energy of the complex. In ligand screening, the comparably large receptor structure is normally treated as rigid and thus ΔG^0_R is assumed to be zero. The estimated binding energy of the system depends only on the interaction energy between the ligand and the receptor (ΔG^0_{RL}) and additionally to a small extent on the change of the internal energy of the flexible ligand (ΔG^0_L).

Table 3.2. Binding constants K and binding free energies ΔG° for camptothecin in 0.02M HCl.

ID	Compound	K [M^{-1}]	ΔG° [kJ mol^{-1}]
	β-cyclodextrin	202±30	-13.2±0.5
	HP-β-cyclodextrin	223±32	-13.4±0.4
	RDM-β-cyclodextrin	186±12	-12.9±0.2
11	Heptakis-[6-deoxy-6-(1-methyl-5-sulfanyl-tetrazole)]-β-cyclodextrin	insoluble	-
12	Heptakis-[6-deoxy-6-(2-aminoethylsulfanyl)]-β-cyclodextrin	4821±572	-21.0±0.3
13	Heptakis-[6-deoxy-6-(2-sulfanyl acetic acid)]-β-cyclodextrin	1450±177	-18.0±0.3
14	Heptakis-[6-deoxy-6-(2-sulfanylethanesulfonic acid)]-β-cyclodextrin	7496±2002	-22.1±0.7
15	Heptakis-[6-deoxy-6-(2-sulfanylpropanoic acid)]-β-cyclodextrin	insoluble	-
16	Heptakis-[6-deoxy-6-(3-sulfanylpropane-1,2-diol)]-β-cyclodextrin	4106±475	-20.6±0.3
17	Heptakis-[6-deoxy-6-(3-sulfanylpropanoic acid)]-β-cyclodextrin	3134±364	-19.9±0.3
18	Heptakis-[6-deoxy-6-(6-sulfanyl-9H-purine)]-β-cyclodextrin	insoluble	-
19	Heptakis-[6-deoxy-6-(2-sulfanyl-pyridine)]-β-cyclodextrin	insoluble	-
20	Heptakis-[6-deoxy-6-sulfanylsulfonyloxysodium)]]-β-cyclodextrin	370±48	-14.65±0.32
21	Mono-[6-deoxy-6-(1-methyl-5-sulfanyl-tetrazole)]-β-cyclodextrin	465±55	-15.2±0.3
22	Mono-[6-deoxy-6-(2-aminoethylsulfanyl)]-β-cyclodextrin	498±69	-15.4±0.3
23	Mono-[6-deoxy-6-(2-sulfanyl acetic acid)]-β-cyclodextrin	493±61	-15.4±0.3
24	Mono-[6-deoxy-6-(2-sulfanylethanesulfonic acid)]-β-cyclodextrin	431±56	-15.0±0.3
25	Mono-[6-deoxy-6-(2-sulfanylpropanoic acid)]-β-cyclodextrin	419±53	-15.0±0.3
26	Mono-[6-deoxy-6-(3-sulfanylpropane-1,2-diol)]-β-cyclodextrin	531±79	-15.6±0.4
27	Mono-[6-deoxy-6-(3-sulfanylpropanoic acid)]-β-cyclodextrin	569±68	-15.7±0.3
28	Mono-[6-deoxy-6-(6-sulfanyl-9H-purine)]-β-cyclodextrin	3629±1567	-20.3±1.1
29	Mono-(2-mercapto-pyridine)-β-cyclodextrin	641±53	-16.0±0.2

In inverse screening the receptors were treated as flexible, whereas the guest molecule was kept rigid (ΔG_L^0=0). Due to the large size of the receptor the change of its internal energy (ΔG_R^0) contributes predominantly to the binding energy. Self-inclusion of the receptor can lead to low-energy conformations of the receptors with only little interaction to the guest molecule. This is shown in Figure 3.11, which depicts a complex with a predicted favorable score (GlamDock). Camptothecin lies on top of the receptor. Instead, one of the hydrophobic side chains is buried in the cyclodextrin cavity and leads to a favorable internal energy

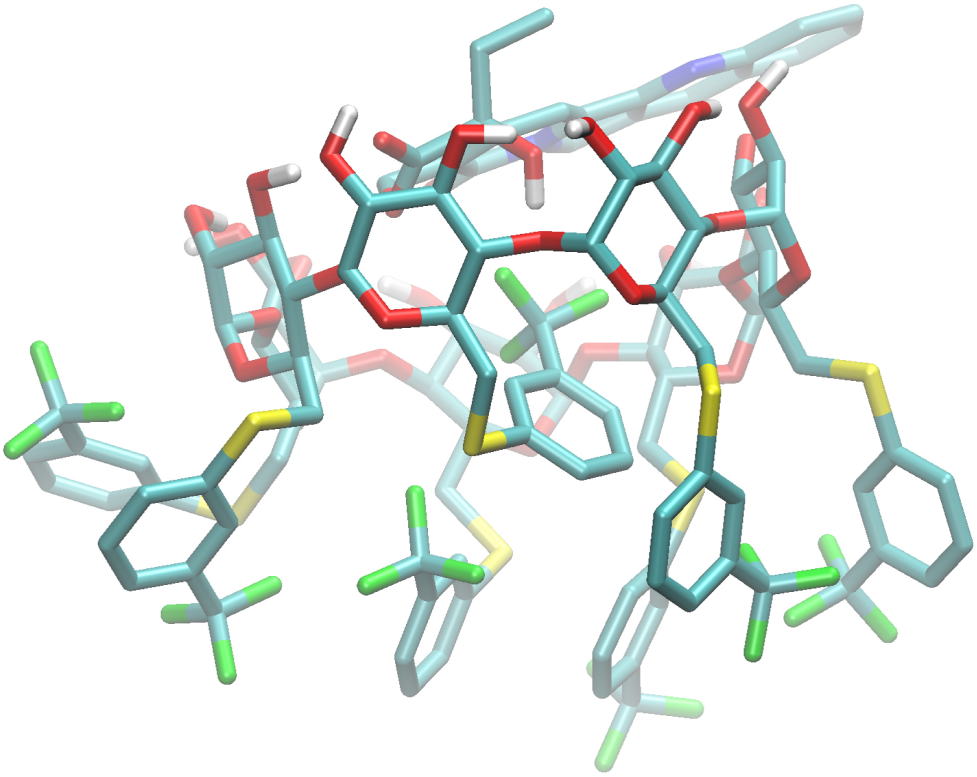


Figure 3.11. GLAMDOCK docking result for a candidate receptor (Heptakis-[6-deoxy-6-(3-(trifluoromethyl)benzenesulfanyl)]-β-cyclodextrin) and camptothecin with a predicted low binding free energy. Hydrogen atoms are omitted for clarity.

(ΔG_R^0), which compensates for the poor intermolecular interactions (ΔG_{RL}^0). This leads to well scoring complexes that show little interaction between ligand and receptor. Furthermore, the average interaction of a system increases quadratically with the number of its atoms, and therefore receptors with large substituents are generally scored more favorably than smaller receptors. There are at least three different approaches to address this type of problem within the paradigm of fast virtual screening:

a) score the complexes only according to the interaction between receptor and ligand
b) add a term to the ranking function, which depends on the number of atoms, to penalize large complexes
c) constrain the docking to allow only conformations with camptothecin in the binding site of the receptor.

In approach (a) it is important to consider the intramolecular receptor energy ΔG_R^0 during the conformational sampling to avoid physically unreasonable conformations of the receptor (e. g. atom-atom overlaps). However, the propor-

tionality of the number of interactions to the size of the receptor remains and leads to better scoring of unspecifically interacting hydrophobic receptors. The second approach (b) reduces the latter problem, but is highly empirical and requires the definition of more or less arbitrary weights for the size-dependent term. Finally, in the last option (c) conformations as shown in Figure 8 are explicitly forbidden, even though they may correspond to the most probable structure of the complex.

In the current work, we chose two different combinations of these approaches. In the screening with GlamDock we used approaches (a) and (b) by explicitly adding a term penalizing the size of the receptor for the ranking, and only used the intermolecular interaction energy ΔG^0_{RL} for scoring. For AutoDock we used approach (c) since the sampling region of the receptor is limited in such a way that camptothecin is always within the binding cleft of the derivatives.

Overall, the results show that these two approaches have their particular advantages and disadvantages. The AutoDock approach led to the selection of receptors which were highly hydrophobic, and could therefore not be measured. On the other hand, it also led to the identification of compound **28**, which is the only mono derivative that can rival the heptakis-substituted derivatives in terms of binding affinity. The GlamDock approach proposed receptors with smaller and more hydrophilic side chains, in general, which show improved binding affinity over β-cyclodextrin. Furthermore the scores correlate reasonably well with the experimental binding affinities. It is interesting to note that one derivative (compound **14**) appears to be an outlier for both scoring functions. Both, AutoDock and GlamDock significantly underpredict its binding affinity.

Nevertheless, in spite of the uncertainties of structure prediction, and the modeling itself, the overall results suggest that at least the tendency of binding affinity is reproduced. For AutoDock a residual standard error of 9.11 kJ mol^{-1} is reported in the literature for a set of 30 protein-ligand complexes (Morris et al., 1998). Furthermore, with regression methods a cross-validated correlation coefficient of 0.89 and a standard deviation of 2.38 were reported for a set of 218 complexes between β-cyclodextrin and different guest molecules (CODESSA-PRO descriptors) (Katritzky et al., 2004). This correlation is clearly better than those achieved in the present work, while the average error is comparable. It should be noted that regression methods are not applicable for the current study since no training data for generating the regression was available before. However, the comparison suggests that our results on this system are probably the best that can be achieved with simple modeling approaches.

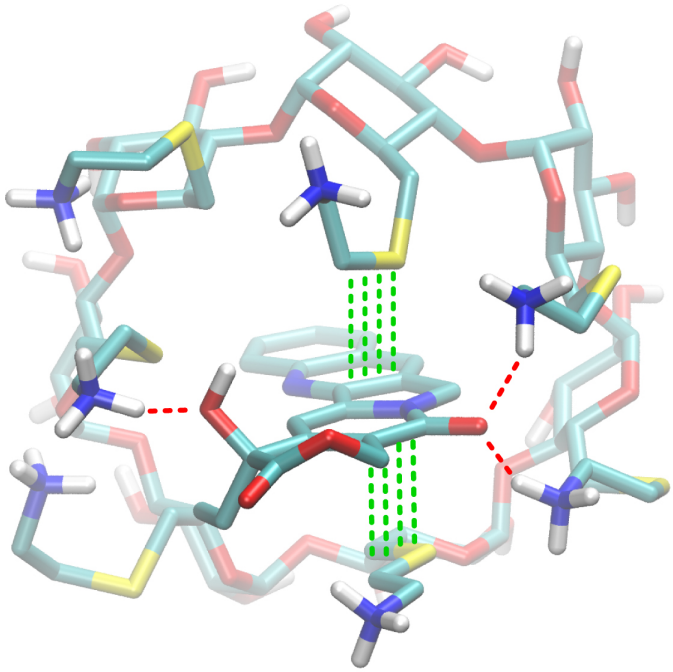


Figure 3.12. The figure shows the generated complex structure of compound **12** to camptothecin (GLAMDOCK). Hydrogen bonds are depicted by dashed red lines, strong hydrophobic interactions are shown as dashed green lines. Apolar hydrogen atoms are omitted for clarity.

To exemplify the interactions involved in the complex formation of camptothecin and the described β-cyclodextrin derivatives we discuss the predicted complex structures of compounds **12** and **18** (see Figures 3.12 and 3.13). The molecular structure of camptothecin offers several possibilities for intermolecular interactions. The large hydrophobic area of camptothecin facilitates dispersive interactions. Consequently, an enlargement of the hydrophobic cyclodextrin cavity by hydrophobic side chains leads to higher binding affinity. This effect is illustrated in Figure 3.12 where the hydrophobic parts of the cysteaminyl side chains of compound **12** show a good shape complementarity and hydrophobic interactions to the camptothecin ring system (dashed green lines). In addition, camptothecin is also able to interact specifically by forming directional hydrogen-bonds. The complex exhibits three intermolecular hydrogen bonds (dashed red lines) of the ammonium groups to hydrogen bond acceptor atoms of camptothecin. On the other hand, polar interacting groups pay a relatively high desolvation penalty in aqueous solution and are most probably not the main driving force behind complex formation for the regarded system.

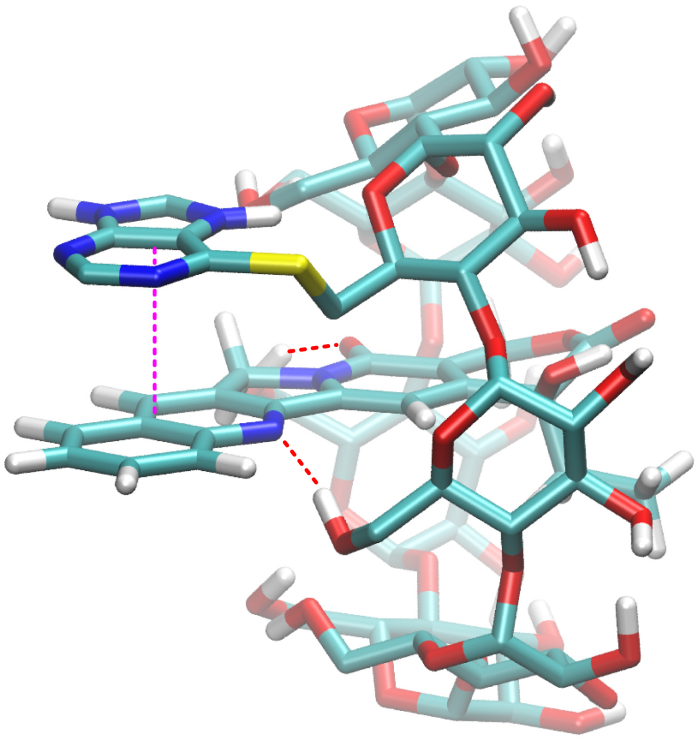


Figure 3.13. The figure shows the generated complex structure of compound **28** to camptothecin (AUTODOCK). The dashed pink line depicts a possible p-stack interaction. Hydrogen bonds are shown as dashed red lines. Apolar hydrogen atoms are omitted for clarity.

Additionally we could show that aromatic building blocks, e.g. purine in compound **28** and, to a smaller extent, pyridine in compound **29** increase complex stability. This result might be best explained by the occurrence of π-stacking (dashed pink line) between camptothecin and the heterocycle (see Figure 3.13). In general hydrophobic interactions are the main driving force behind complex formation in aqueous solution, while polar interactions are more responsible for the specificity of the interaction. While a general size effect can be observed in the data, specific effects are evident, since mono-substituted compounds exist, which bind better than heptakis-substituted compounds and vice versa. Compound **28** binds better than all other mono-substituted derivatives and better than some of the heptakis-substituted compounds. It is recognized by both affinity predictions as the best of the mono-derivatized compounds. Inversely, an extension of the cavity does not necessarily result in an increased binding affinity of the complex. The heptakis-substituted thiosulfate β-cyclodextrin derivative (compound **20**), for example, exhibits ΔG° value of -14.62 kJ mol^{-1} to camptothecin and was also

predicted to have a comparably low binding affinity. These two examples serve to illustrate that rational design of the investigated system towards higher binding affinities is indeed not trivial, yet possible.

3.6 Conclusions

We have investigated a rational optimization approach to synthetic receptor design. Our approach is complementary to the work of de Jong et al. (2002), who described the identification of new ligands for a given host. Our results indicate that inverse virtual screening can support the identification of novel receptors for a given ligand and might open up novel possibilities for the tailored design of drug delivery systems. Finally, it should be noted that this approach is not limited to cyclodextrin derivatives and the idea of receptor design by means of inverse virtual screening can be applied to other host classes. The rules for generating the virtual library of hosts can, in principle, be arbitrarily expanded. Future work might go along this direction.

4

Combined Similarity and QSPR Based Virtual Screening for Guest Molecules of β-cyclodextrin

This chapter reports on the combination of a similarity-based virtual screening technique with a quantitative structure property relationship (QSPR) model for the identification of new guest molecules with high affinity to β-cyclodextrin (Steffen et al., 2007a). Our technique provides a new and successful means for the identification of novel guest molecules for synthetic receptors. The work has been conducted in collaboration with the group of Dr. Joannis Apostolakis from the Ludwigs-Maximilian University in Munich and the group of Professor Dr. Gerhard Wenz from the Saarland University in Saarbrücken. All experimental work was done by Anne Engelke.

4.1 Introduction

The rational design of novel host-guest systems is of particular interest in supramolecular chemistry (Lavigne & Anslyn, 2001). Recently, studies have been published, in which structure-based docking tools were applied to synthetic host-guest systems in order to identify optimally interacting systems (de Jong et al., 2002; Corbellini et al., 2004). As previously described (see Section 3.1), both studies performed a virtual screening for the identification of novel guest molecules. These approaches demonstrated the potential and the possible impact of structure-based virtual screening methods from drug design for the optimization of synthetic host-guest systems. Another virtual screening technique solely uses information of known guest molecules and is known as similarity- or ligand-based virtual screening (see Section 1.2.1). This technique relies on the assumption that structurally similar molecules exhibit similar binding properties with respect to a given target (Patterson et al., 1996; Martin et al., 2002). In general, these types of tools are significantly faster compared to docking (Kämper et al., 2007). Researchers have proposed a broad range of approaches to describe similarity between molecules

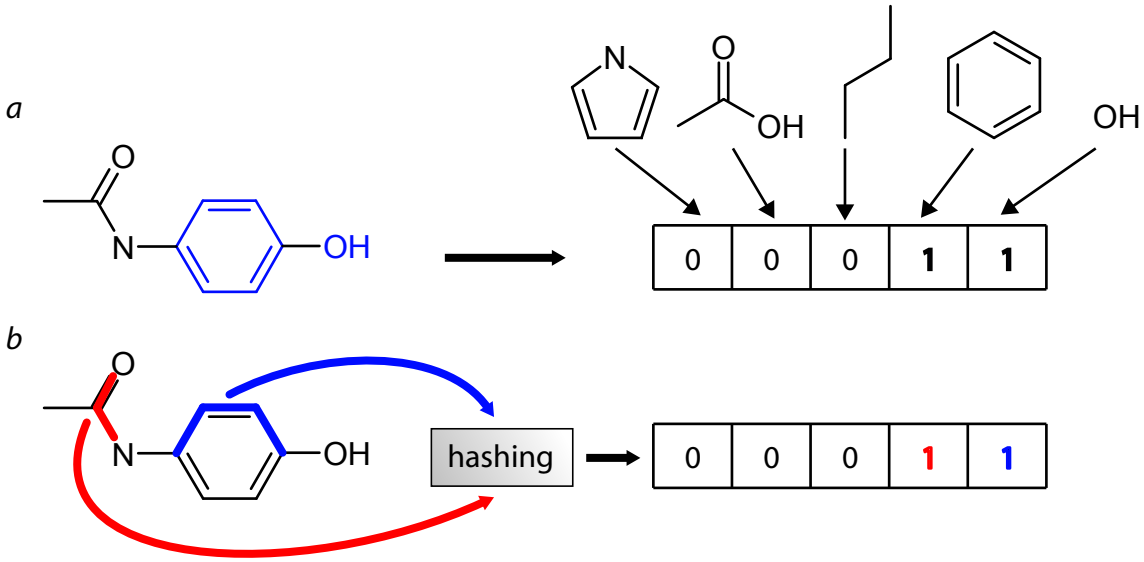


Figure 4.1. The figure shows two different approaches to generate a fingerprint representation. a) Fingerprint based on structural keys use a dictionary of substructures, each of which corresponds to a defined bit in the bit string. b) Hashed fingerprints do not require a predefined dictionary as atom paths are generated on the fly. A hashing function translates them into bits.

(Sheridan & Kerarley, 2002; Lengauer et al., 2004). Some representative examples are detailed below.

4.1.1 Fingerprint-Based Similarity Tools

Fingerprint tools represent the fastest class of similarity tools. Here, the molecules are represented as bit strings (see Figure 4.1). Commonly two different approaches are used to generate a bit string representation of a molecule. The first uses a fixed number of structural keys (substructures) and assigns each of them to a bit within the bit string [MDL Information Systems, Inc. (http://www.mdli.com), Digital Chemistry (http://www.digitalchemistry.co.uk)] (see Figure 4.1, a). If a given structural key is present in a molecule, the corresponding bit is set to one; otherwise the corresponding bit is set to zero. The second approach is called hashed fingerprints [Daylight Chemical Information Systems (http://www.daylight.com)] (see Figure 4.1, b). In contrast to fingerprints based on structural keys, hashed fingerprints do not employ a predefined substructure dictionary. The hashed fingerprints are generated by finding all possible linear paths of connected atoms up to a defined length that occur within a molecule. Then, a hashing function translates each path to a number of bits that are set to one. While the coding step from a given atom path to the corresponding bits is unique, the step backwards from the bits to a path is ambiguous as different path can have the same

bit pattern. Principally, this approach can result in false-positive hits, but does not generate false negatives. In contrast to structural key based fingerprints, this type can be applied to any kind of chemical structure, even if rather uncommon substructures dominate. Besides these two approaches combinations of both are reported. Unity [Tripos Inc. (http://www.tripos.com)], for example, uses a hybrid of hashed fingerprints and structural keys.

The molecular similarity is commonly calculated by means of the Tanimoto coefficient T (Tanimoto, 1957) on the basis of bit strings. T is defined as the following ratio:

$$T = \frac{c}{a + b - c} \tag{4.1}$$

with a equal to the number of bits set to one in molecule A, b equal to the number of bits set to one in molecule B, and c equal to the number of bits that are set to one in the bit strings of both molecules. The Tanimoto coefficient is in the range between 0 and 1. The more similar two molecules are, the closer the corresponding Tanimoto coefficient is to one. Besides the Tanimoto coefficient numerous other similarity coefficients have been proposed (Leach & Gillet, 2003).

The comparison of bit strings is computationally very efficient and therefore fingerprint methods can handle large libraries of molecules. However, bit strings only roughly represent the overall molecular topology. Two molecules might exhibit a high similarity if, for instance, both of them have many functional groups or side chains in common, although they significantly differ in their molecular structure. In comparison to graph- or shape-based similarity techniques the computed similarity might thus not be directly visible.

4.1.2 Graph- and Tree-Based Similarity Tools

In graph-based similarity tools molecules are represented as graphs. Within these graphs, nodes represent structural features, such as an atom or groups of atoms, and edges denote their connectivity. The similarity is calculated on the basis of a generated mapping between two such graphs. This mapping points to corresponding parts of the two molecules. The tool FUZZEE, e. g., which was applied for the study described in this chapter, belongs to this class of tools (see Section 4.3.2.1).

A very advanced tree-based similarity tool is called FTREES (Rarey & Dixon, 1998). Here, molecules are described as trees, comprising nodes that correspond to molecular fragments and edges that describe their connectivity. Each of these nodes contains features, accounting for chemical and steric properties of com-

prised fragments. The similarity between two molecules is calculated by matching their two corresponding feature trees, while preserving their molecular topologies. The method allows for combining connected nodes to a combined node, that represents the features of the contained nodes. In this way, a biphenyl system can for example match a naphthalene ring system.

In general, tools of this class have the advantage that they directly depict corresponding parts of two molecules and the chemical similarity is clearly visible. Furthermore, in contrast to three-dimensional descriptors they are independent from the conformations of the molecules. Reduced graphs, which are very abstract representations of molecules, allow for so-called scaffold hoppings (Böhm et al., 2004), such that molecules of considerably different molecular structure, yet similar physicochemical properties are considered similar (Barker et al., 2006).

4.1.3 Similarity Tools Based on Shape or Structural Superimposition

Molecular similarity can also be deduced from the comparison of the shape or the three-dimensional structure of molecules. Basically, the aim of such methods is to find and to quantify the maximal volume overlap of two molecules, whilst potentially considering physicochemical features of the molecules.

The tool ROCS represents molecules by means of continuous functions that are derived from atom-centered Gaussians (Grant et al., 1996; Rush et al., 2005). This representation allows for the calculation of an alignment of two molecules, in which ROCS maximizes the overlap of the volumes of the molecules. Optionally, a chemical force field maximizes the overlap of parts of the molecules with identical interaction properties.

Lemmen et al. (1998b) presented the tool FLEXS, which superimposes the structures of two molecules, using an incremental construction principle related to FLEXX. FLEXS keeps the reference molecule in a rigid conformation. This conformation can for example be taken from a crystal structure. The test molecule is treated as flexible. Similar to FLEXX, FLEXS cuts the test molecule into fragments, out of which preferably rigid base fragments with many interaction groups are selected. These base fragments are aligned to matching parts of the query molecule. Subsequently, FLEXS builds up the entire test molecule in an incremental manner. During the incremental construction, a scoring function assesses paired intermolecular interactions, as well as the steric overlap of the molecules.

The rather abstract representation of molecules in shape-based tools, which is due the independence from atom types and bonding patterns, is predestinated to allow for scaffold hoppings (Böhm et al., 2004). However, particularly in the case of flexible molecules the major difficulty associated with three-dimensional shape

descriptors is the problem of handling this flexibility. If only one conformation is considered, shape-based similarity tools might not be able to find similarities as the calculation is possibly based on non-corresponding conformations. The consideration of multiple conformations per molecule, however has a direct influence on the performance.

4.2 Aim of the Study

In this work, we have combined a graph-based similarity method with a quantitative structure property relationship (QSPR) model. This model provides the means of estimating the binding free energy ΔG° of the similarity hits and is then used as a second filter. The value of QSPR models for the prediction of ΔG° values of complexes between various guest molecules and β-cyclodextrin was shown in two recent studies (Suzuki et al., 2000; Katritzky et al., 2004). In both cases stable and well predictive models were generated on the basis of computed molecular descriptors (see Section 1.1.4.1). This was motivation for us to use this technique in combination with similarity screening to identify high affinity guest molecules of β-cyclodextrin out of a given database of molecular compounds.

4.3 Methodology

Figure 4.2 illustrates the workflow of our study. First a QSPR model for the prediction of the binding free energy (ΔG°) of β-cyclodextrin inclusion complexes was generated. Second, a similarity-based virtual screening was performed. Then, the QSPR model was used for assessing molecules that were found by similarity-based virtual screening. We selected molecules with a predicted high affinity for β-cyclodextrin in order to experimentally verify our computations.

4.3.1 Generation of a Support Vector Machine Regression Model

We developed a support vector machine regression (SVMR) based QSPR model that was trained to predict the binding free energy ΔG° of β-cyclodextrin inclusion complexes based on molecular descriptors (Tetko et al., 2005) and experimental data from literature. The molecules of our training dataset were taken from Suzuki (2001). All 218 molecules form 1:1 inclusion complexes with β-cyclodextrin. For all molecules the ΔG° values of the complexation to β-cyclodextrin were experimentally determined. The molecules were drawn with

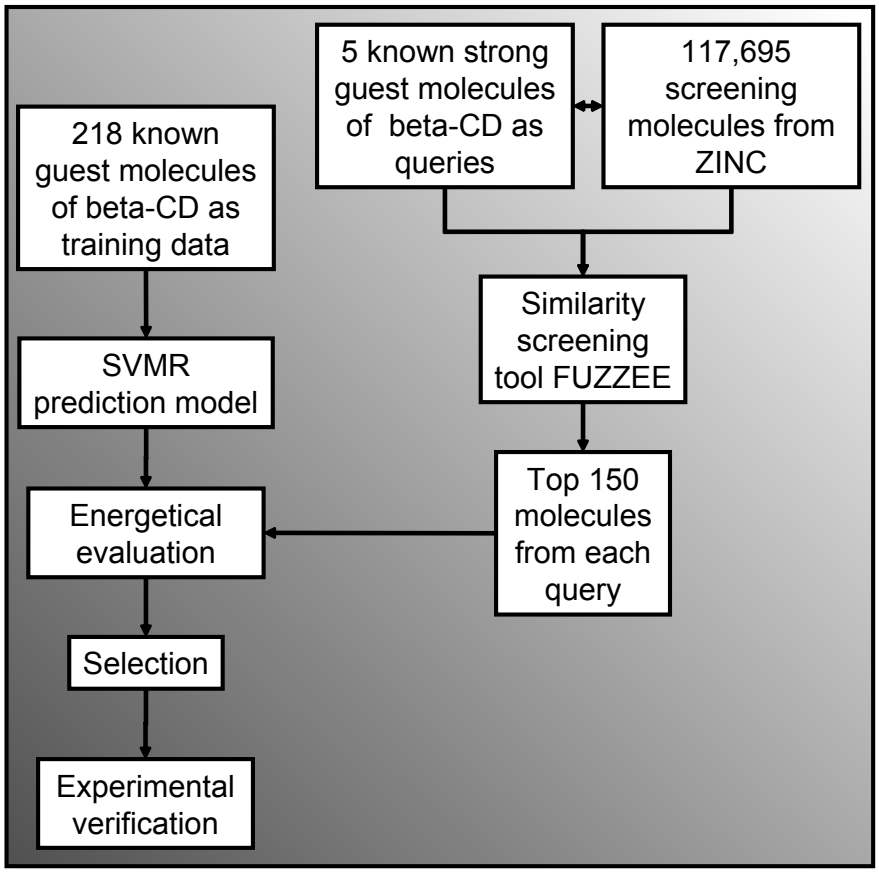


Figure 4.2. Schematic flow of the applied virtual screening method.

reasonable protonation states with ISIS/Draw [www.mdli.com] and exported as MOL files (Dalby et al., 1992). CORINA (Sadowski & Gasteiger, 1993) was used to generate low-energy three-dimensional structures in the SDF format (Dalby et al., 1992). Finally, all molecules were visually inspected and, if necessary, corrected. 1,666 molecular descriptors were calculated for each molecule by means of the E-Dragon web server (Tetko et al., 2005; Todeschini & Consonni, 2000). The descriptors account for simple molecular properties, from molecular weight and topological features up to elaborate quantum chemical characteristics. For the subsequent development of the QSPR model all properties were scaled to the range -1 and 1 in order to avoid numerical problems and prevent a bias in the descriptor space.

The theoretical background of SVMR has been described in detail by Drucker et al. (1997) (see Section 5.2.3.3). In this work we use the LIBSVM implementation (Chang & Lin, 2001) with the linear kernel function and combine it with a forward descriptor selection procedure. The latter helps to limit the number of integrated descriptors, which enhances the interpretability of the regression model. Furthermore the risk of overfitting the model to the underlying data and thereby

decreasing the predictivity of the model for non-training molecules is reduced, if the number of integrated descriptors is limited. The forward descriptor selection procedure is based on a greedy heuristic that works as follows. First, a regression model is generated for each single descriptor with tenfold cross-validation. Second, the descriptor which gives the highest cross-validated squared linear correlation coefficient r^2_{cv} is chosen.[1] Then, this descriptor is combined with each of the remaining descriptors and the pair that leads to the regression model with the highest r^2_{cv} value is selected for the next descriptor extension step. This is repeated until a maximum for r^2_{cv} is reached. All descriptors at this stage are integrated into the final model. This model was used for the prediction of the ΔG° values of the inclusion complexes between β-cyclodextrin and the guest molecules that were identified by virtual screening.

Internal validation of the QSPR model

The squared linear correlation coefficient r^2_{cv} derived from the tenfold cross-validation test is overoptimistic with respect to the prediction accuracy of unseen data, especially since cross-validation was used to select the descriptors for the model. A more realistic estimate of the predictivity of the final model generated in the described manner can be obtained by means of a nested cross-validation protocol (Ruschhaupt et al., 2004) (see Figure 4.3). Therefore the data is split randomly into three equally sized subsets S1, S2 and S3. Out of each possible pairing of the three subsets three combined subsets V1 (S1+S2), V2 (S1+S3) and V3 (S2+S3) are built. Each of the latter serves as a training set for the generation of a QSPR model, which is obtained in the same manner as described above by the tenfold cross-validation based descriptor selection protocol. This is the inner loop of the nested cross-validation. The models generated in the inner loop are then used to predict the respective remaining, unused subset (outer loop - prediction set). The prediction quality of the model on these test sets is taken to mirror the prediction quality for unseen data.

4.3.2 Virtual Screening

Five known β-cyclodextrin guest molecules with ΔG° values less than or equal to -20 kJ mol^{-1} were selected as query compounds (Table 4.1). Three of them,

[1] r^2_{cv} is the Pearson correlation coefficient calculated for the predictions of all ten cross-validation runs together. In principle, also the cross-validation coefficient as used in Chapter 5 could have been used. Since the predictions made in this chapter were experimentally validated and not only theoretically, we decided to use the simple and directly accessible Pearson correlation coefficient, which is a direct output of the applied SVMR-tool.

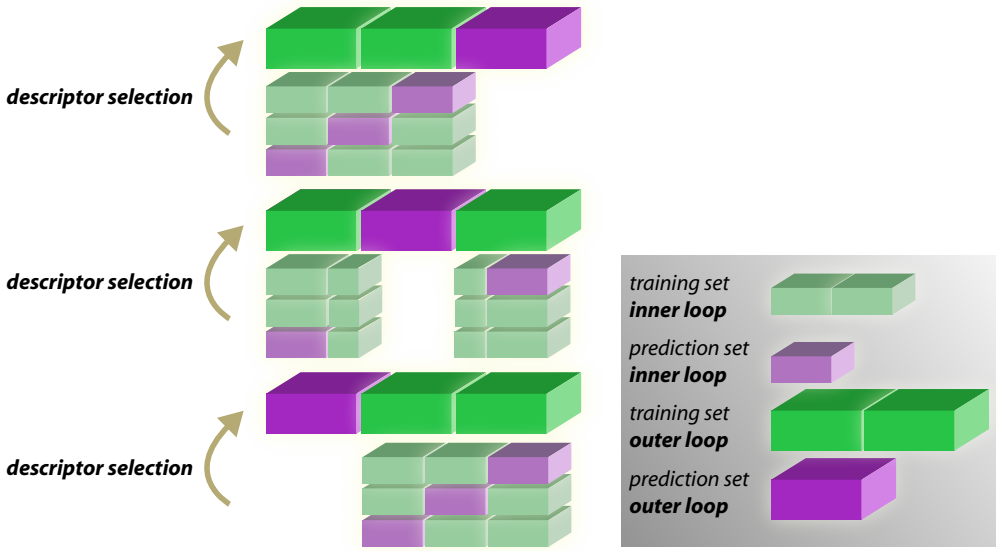


Figure 4.3. Schematic illustration of the nested cross-validation protocol. The data is divided into three equally sized subsets. Then the model generation and the descriptor selection is done for each pair of subsets (inner loop) based on tenfold cross-validation. The remaining subset serves as the prediction set (outer loop).

i. e. chlorpromazine (compound **3**), flurbiprofen (compound **4**) and ibuprofen (compound **5**) are drug molecules. The query compounds were prepared with the same protocol as described for the preparation of the QSPR training set molecules (see Section 4.3.1).

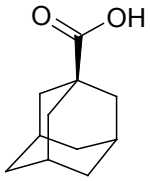
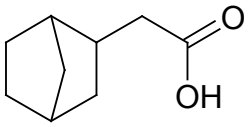
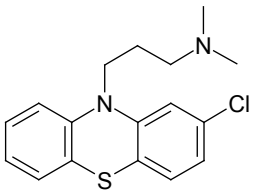
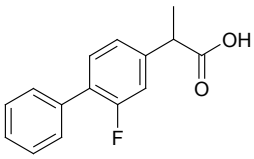
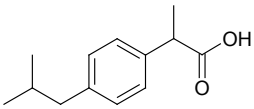
The screening dataset was downloaded from ZINC (Irwin & Shoichet, 2005) as SD-files (Dalby et al., 1992). For reasons of direct and fast commercial availability we chose the Sigma Aldrich subset. Altogether this subset contains 117,695 entries. The structures were taken as provided from ZINC (see Irwin & Shoichet (2005) for closer details of their preparation protocol).

4.3.2.1 FUZZEE

The approach used for similarity screening is based on a variant of the graph matching algorithm in the GMA program (Marialke et al., 2007). Direct graph matching at the atomic level leads to the identification of chemically closely related structures. In order to find molecules of different topology, but yet similar physicochemical features it is preferable to perform the comparison on a more abstract representation of the molecules, for example, at the level of functional groups. The computational representation of molecules used in this work is related to the reduced graphs used by Barker et al. (2006) and is illustrated in Figure 4.4.

Reduced graphs describe molecules as a collection of connected functional groups or fragments. Each node in the graph represents a fragment in the

Table 4.1. Known β-cyclodextrin guest molecules serving as query compounds. Experimental error of ΔG° within $\pm$ 0.3 $\mathrm{kJ\,mol^{-1}}$.

ID	Structure	CAS-No.	ΔG° [$\mathrm{kJ\,mol^{-1}}$]	Lit.
1		828-51-3	-24.9	Harrison & Eftink (1982)
2		1007-01-8	-20.8	Godinez et al. (1995)
3		69-09-0	-22.4	Hardee et al. (1978)
4		5104-49-4	-18.8	Ueda & Perrin (1986)
5		15687-27-1	-22.6	Wenz[1]

[1] The ΔG° value was determined within the laboratory of Professor Wenz, Saarbrücken according to the protocol described in Section 4.3.4.

molecule. Edges between the nodes represent the connectivity of the corresponding fragments. The fragments are obtained as follows: Rings containing up to seven atoms form a fragment. Larger rings are fragmented according to the rules for linear chains. Atoms that belong to more than one ring are assigned to each of the respective fragments. Furthermore atoms with at least two non-hydrogen neighbors form the basis of a fragment. The remaining atoms with only one non-hydrogen neighbor are merged into their neighbor's fragment, unless their neighbor is member of a ring fragment. In this case, the atom forms a single atom fragment. Two nodes are connected, if they share one or more atoms, or if two of the contained atoms are connected to each other by a chemical bond. Each

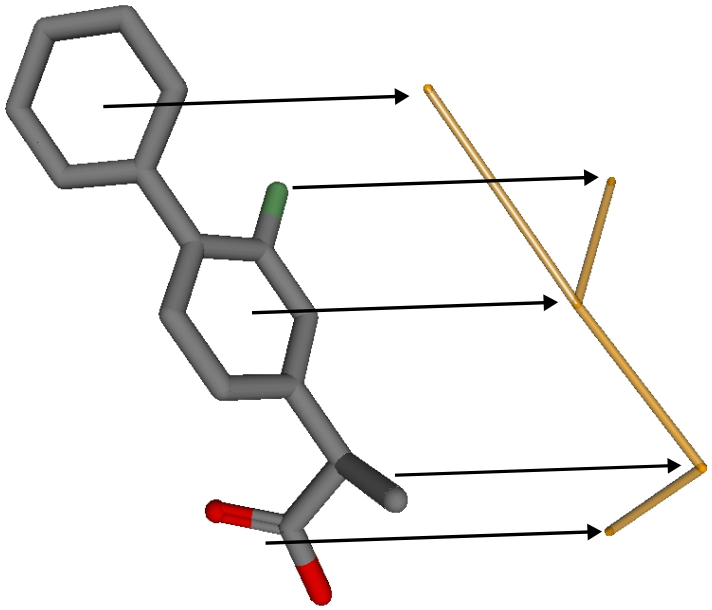


Figure 4.4. Reduced graph representation of flurbiprofen (left: atomic level; right: reduced graph representation.). Hydrogen atoms are omitted for clarity.

node is annotated with a number of features describing the atoms that constitute the original fragment. The features used are shown in Table 4.2. Each feature has

Table 4.2. Features of nodes and weighting scheme.

index	**weight**	**feature**
1	1	carbon sp^3
2	1	carbon $sp^1/sp^2/ar$
3	1	nitrogen sp^3
4	1	nitrogen $sp^1/sp^2/ar$
5	1	Oxygen
6	1	Phosphorus
7	1	Sulphur
8	1	Halogens
9	1	other atom types
10	4	H-Bond donor base
11	4	H-Bond acceptor

a weight and a value which counts the occurrences of the feature in the fragment.

The matching of two nodes in the comparison between two molecules yields a weight given by the following equation:

$$sim(a, M(a)) = \sum_{k=1}^{11} W_k min(a_k, M(a)_k) \tag{4.2}$$

where a is the node in the first molecule, $M(a)$ its match, k iterates over the indices of the features listed in Table 4.2, W_k, a_k, and $M(a)_k$ are the weight, and the value of the corresponding feature in the first and the second node, respectively.

The overall similarity (s) is given as the sum of all similarities between matched nodes (cw), normalized over the maximum of the self-similarities of the compared molecules.

$$cw(K, L, M) = \sum_{a \in K} sim(a, M(a)) \tag{4.3}$$

$$s(K, L, M) = \frac{cw(K, L, M)}{max(cw(K, K, I), cw(L, L, I))} \tag{4.4}$$

where cw is the sum off all matched similarities, s is the overall similarity, K and L are the ligands being compared, M is the mapping found by the algorithm, and I is the identity mapping.

An example of a molecule matching is given in Figure 4.5, where the matching parts of the molecules flurbiprofen and 4-phenoxybenzoic acid are depicted by lines.

4.3.3 The Screening Protocol

For each of the five query compounds a virtual screening run was performed against the screening dataset. Ranking lists were derived from the calculated similarity scores. The top 150 molecules of each of the ranking lists were scored by means of the generated QSPR model. The aim of our study was to search for molecules with low ΔG° value in complex with β-cyclodextrin. From the screening runs only those molecules were selected for which a lower or comparable ΔG° value with respect to the corresponding query structure was predicted. Finally, we were interested in identifying novel molecular scaffolds and thus only molecules with a significant change in the structure compared to the query structure were

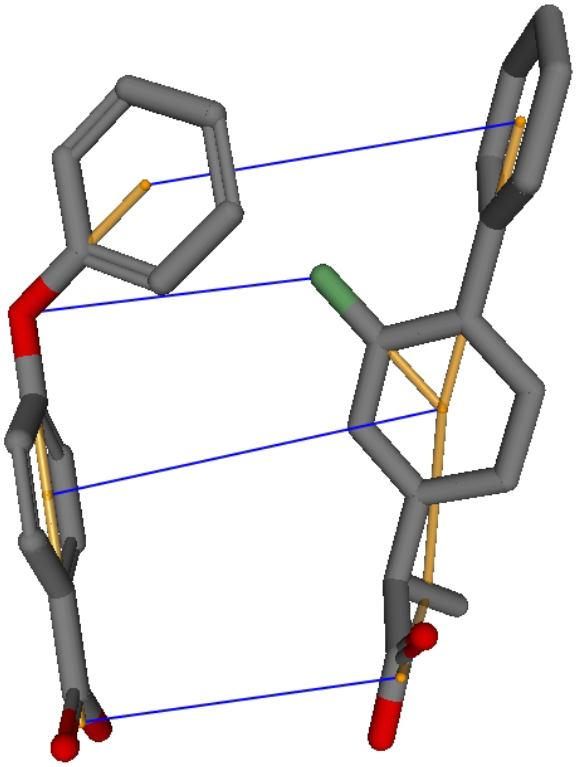


Figure 4.5. The matching between flurbiprofen (right) and 4-phenoxybenzoic acid (left). Hydrogen atoms are omitted for clarity.

considered. Additionally we limited ourselves to molecules with a promising water solubility, allowing for the experimental determination of ΔG° by isothermal microcalorimetry. Furthermore the molecules had to be commercially available.

4.3.4 Binding Studies

Compound **5** was purchased from Avocado, **19** and **10** from Fluka, **6**, **9**, **15**, **16**, **18** and **20** from Sigma, **8**, **12**, **17** and **21** from Aldrich, **11** and **13** from Acros Organics and **14** from ABCR. The ΔG° values of the complexes between β-cyclodextrin and the compounds that were sufficiently water-soluble were measured with isothermal microcalorimetric titrations. All titrations were performed by Anne Engelke from the group of Professor Wenz.

The microcalorimetric titrations were performed at a temperature of 25.0°C with an AutoITC isothermal titration calorimeter (MicroCal Inc., Northampton, USA) using 1.4144 mL sample and reference cells, which were filled with distilled water. The sample cell was filled with a 1.3 mM solution of the respective guest in 25 mM phosphate buffer pH 6.79 and constantly stirred at 450 rpm. A 13 mM solution of β-cyclodextrin was prepared in the same buffer. This solution was au-

tomatically added by a syringe in 20 separate injections of 12.5 μL. The resulting 20 heat signals were integrated to yield the mixing heats, which were corrected by the corresponding dilution enthalpies of β-cyclodextrin. The titration curve was fitted by non-linear regression. Thereby a 1:1 stoichiometry of the inclusion compound and the host molecule was appropriate. The binding constant K_S and the molar binding enthalpy ΔH° were obtained as fitting parameters, from which the binding free energy ΔG° and binding entropy ΔS° were derived.

4.4 Results and Discussion

In the first step we generated a QSPR model for a dataset consisting of 218 molecules. For each of the molecules 1,666 descriptors were calculated with E-DRAGON. The final model integrated 68 descriptors (see Appendix B). For ten-fold cross-validation an r^2_{cv} value of 0.95 and a root mean squared error ($RMSE$) of 1.17 kJ mol^{-1} was obtained (see Figure 4.6). The observed correlation is in good agreement with the one reported by Suzuki (2001) (r^2_{cv}= 0.92) indicating that the chosen regression methodology and the computed descriptors are appropriate for this study. In contrast to Suzuki's descriptors the E-DRAGON descriptors are freely accessible.

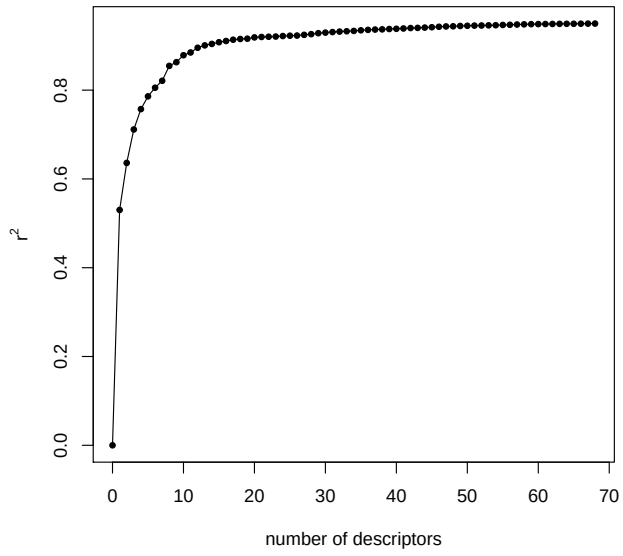
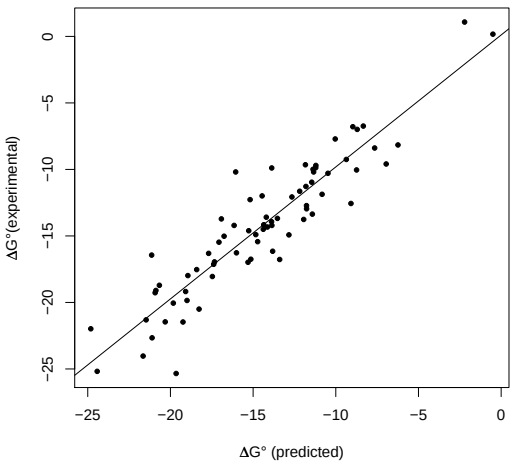
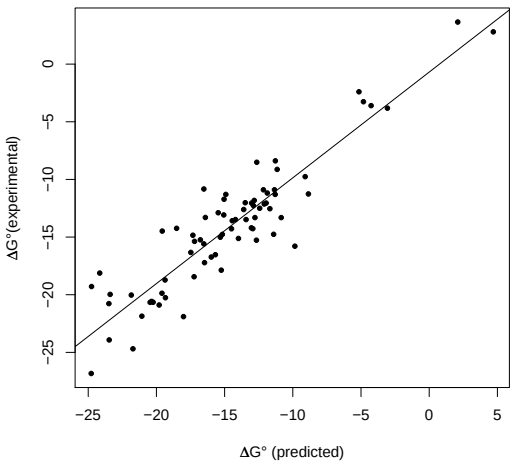


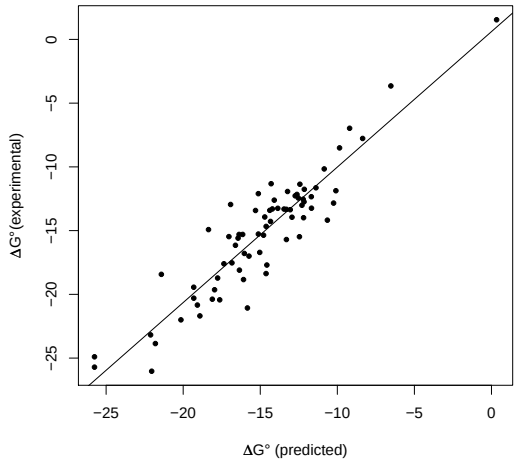
Figure 4.6. Descriptor selection for the training set. For 68 descriptors the maximal r^2_{cv} value of 0.95 was found. The $RMSE$ is 1.17.



(a) Subset S1



(b) Subset S2



(c) Subset S3

Figure 4.7. (a) Prediction of ΔG° for subset S1 by means of the regression model for validation set V3 (r^2 = 0.85, $RMSE$ = 1.98). (b) Prediction of ΔG° for subset S2 by means of the regression model for validation set V2 (r^2 = 0.84, $RMSE$ = 2.32). (c) Prediction of ΔG° for subset S3 by means of the regression model for validation set V1 (r^2 = 0.85, $RMSE$ = 1.89).

For the validation of our approach a nested cross-validation protocol was used. As described in the methodology section, for each of the three validation sets models were generated with the same procedure as applied for the entire final model. Each model was then used to predict the ΔG° values of the corresponding unused molecules. The mean r^2 value of the three sets is $0.84 \pm 0.01\,\mathrm{kJ\,mol^{-1}}$. The mean $RMSE$ for the three sets is $2.06 \pm 0.23\,\mathrm{kJ\,mol^{-1}}$. In Figures 4.7 (a), 4.7 (b)

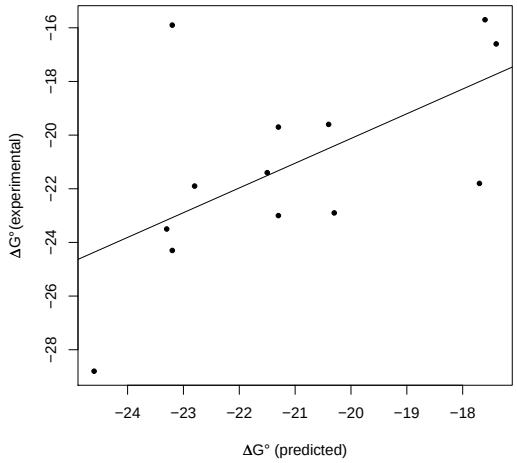


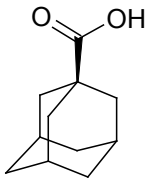
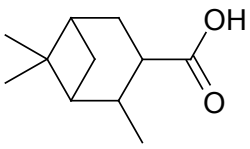
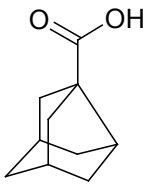
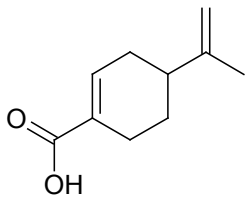
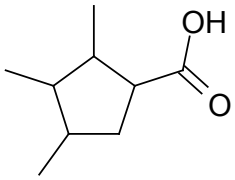
Figure 4.8. Dependence of the predicted and experimental ΔG° values of the screening hits.

and 4.7 (c) the predicted ΔG° values of are plotted against the corresponding experimental values.

We selected five known guest molecules as query structures for the virtual screening. All similarity screening runs together took approximately 1 h on a single Xeon 2.8 GHz CPU. This includes preprocessing of the database. After scoring the top-ranking 150 molecules of each of the five virtual screenings by means of the generated statistical regression model, the 16 most promising molecules regarding their predicted ΔG° value were selected for experimental testing (see Tables 4.3, 4.4, 4.5, 4.6, 4.7).

Two molecules were insoluble and thus no experimental measurement could be performed. Only one molecule displayed no binding affinity at all. Ten molecules exhibited a binding free energy of about -20.0 kJ mol^{-1} or less. Five of them (**10**, **13**, **17**, **19**, **20**) showed a stronger binding affinity than the corresponding query. Thus for three of the five screenings at least one ligand was found with a stronger affinity to β-cyclodextrin than the corresponding query. This is a good result considering that on average only 3.2 new compounds were experimentally tested per query. The $RMSE$ of the predicted values to the experimentally determined values is 2.9 (where we only consider those molecules for which a binding free energy could be measured). The correlation r^2 is equal to 0.35 when all molecules are considered (see Figure 4.8). However, the binding affinity of compound **21** is obviously strongly overrated by the QSPR model. r^2 increases to 0.65 if the data point of compound **21**, an obvious outlier, was omitted. The measured ΔG° for this compound was clearly higher (less favorable) than the predicted value. We attribute this discrepancy to repulsive forces caused by steric interactions

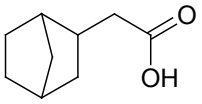
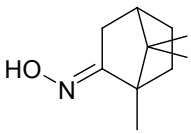
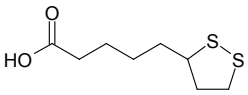
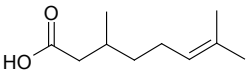
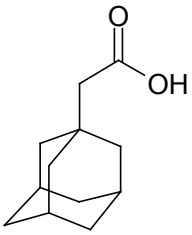
Table 4.3. Selected guest molecules derived from the virtual screening against query **1**.

ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
1		828-51-3	1.00	-	-24.9	-23.0	1.9
6		58096-29-0	0.74	-23.3	-23.5	-21.5	2.1
7		16200-53-6	0.73	-22.8	-21.9	-17.0	5.0
8		23635-14-5	0.73	-21.5	-21.4	-16.5	4.9
9		unknown	0.73	-21.0	no complexation	no complexation	no complexation

due to the branched structure of this guest. The difference between the cross-validated r^2_{cv} and the r^2 for the predicted ligands lies with the fact that we only suggested compounds with a high binding affinity for experimental testing. Thus the variance of the binding free energy of this data is lower compared to data used as the training set, leading to lower r^2 values. However, the accuracy, i. e. the RMSE, is comparable to those obtained in the nested cross-validations.

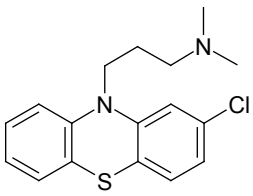
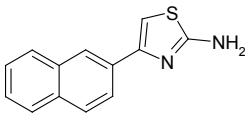
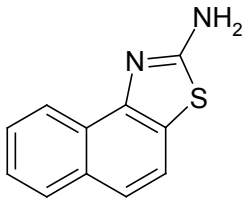
We consider the combination of the similarity-based virtual screening technique and the QSPR model as an effective way to minimize the drawbacks of

Table 4.4. Selected guest molecules derived from the virtual screening against query **2**.

ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
2		1007-01-8	1.00	-	-20.8	-10.7	10.2
10		2792-42-9	0.79	-21.3	-23.0	-25.2	-2.1
11		1077-28-7	0.79	-20.4	-19.6	-15.0	4.7
12		18951-85-4	0.75	-21.3	-19.7	-16.5	3.2
13		4942-47-6	0.74	-24.6	-28.8	-24.6	4.2

each of the two methods when independently used. The sole application of the similarity tool lacks of a concrete estimation of the binding free energy. Although the general principle of similarity is reasonable in many cases, certainly a slight change in structure can have a significant (negative) influence on binding properties. This can be partly tested for by the application of the quantitative filter in the second step. Consider, for example, a structural series of molecules which was investigated in the group of Prof. Wenz some years ago (Table 4.8) (Höfler & Wenz, 1996). All molecules are structurally related to 4-tert-butyl benzoic acid. The overall similarity of each structure to 4-tert-butyl benzoic acid is in the range of 0.96 to 0.83, the corresponding energies however differ significantly. The QSPR model predicts the binding free energy values with an r^2 value of 0.92 to the experimental values. The application of the QSPR model thus helps to filter out molecules with unfavorable binding energies. In fact, about half of the top-

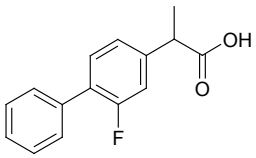
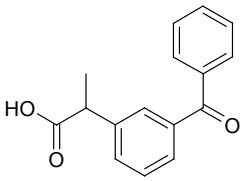
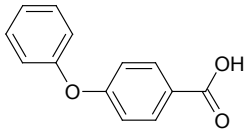
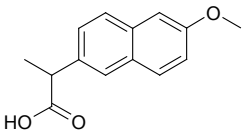
Table 4.5. Selected guest molecules derived from the virtual screening against query **3**.

ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
3		69-09-0	1.00	-	-22.4	-26.8	-4.5
14		21331-43-1	0.72	-18.8	unsoluble	unsoluble	unsoluble
15		40172-65-4	0.68	-17.2	unsoluble	unsoluble	unsoluble

ranking molecules of each of the screenings were filtered out by the application of the QSPR model.

Conversely, virtual screening based only on the output of a regression model is problematic, because the predictions of QSPR models such as the one used are generally relevant only for a limited neighborhood of the chemical space centered around the training set of the model. Thus, the application of the QSPR model alone leads to an unacceptable number of false-positive molecules that do not bind to β-cyclodextrin. To demonstrate this point, we show selected molecules from the screening set in Table 4.9. These molecules were filtered out by FUZZEE. The QSPR model alone, however, predicts a comparably low binding energy. Although not experimentally verified, those molecules do not exhibit the typical structural and functional features of β-cyclodextrin ligands, and are partly simply too large to fit into the cavity. This results from the fact that the QSPR model was only trained on molecules that bind to β-cyclodextrin, while non-binding molecules are not considered. In general, the chemical space of non-binders is too large, that even including negative data into the regression model does not guarantee it is sufficiently covered in quantitative models of affinity. Instead, our strategy in this work has been the prior application of the similarity-based screening technique

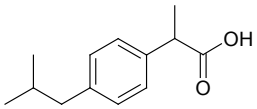
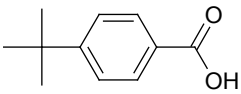
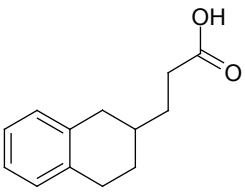
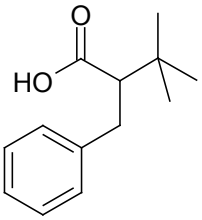
Table 4.6. Selected guest molecules derived from the virtual screening against query **4**.

ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
4		5104-49-4	1.00	-	-18.8	-23.3	-4.5
16		22071-15-4	0.94	-17.6	-15.7	-17.1	-1.3
17		2215-77-2	0.91	-17.7	-21.8	-15.8	6.0
18		22204-53-1	0.91	-17.4	-16.6	-12.6	4.0

to focus on molecules that exhibit the principal features of β-cyclodextrin ligands and lie within the scope of the regression model.

While overall we consider the study successful in the sense that it identified new ligands with high affinity to the targeted host molecule, it is important to note that the combination of the similarity screening with a QSPR model does not solve all problems. Compound **21** exhibits a significantly lower binding affinity in the experimental testing than its predicted value. Even though the QSPR model was trained on a chemically diverse set of molecules, obviously not all features that are important for β-cyclodextrin binding have been taken into account. In the case of compound **21** most probably the sterically demanding tertiary butyl group diminishes the shape complementarity. Compound **9** did not show any binding affinity in the experimental testing. This might be attributed to the cis-standing methyl groups, leading to a too bulky shape that does not fit into the β-cyclodextrin-cavity. To gain closer insights into the mechanisms of molecular recognition certainly further experimental studies would be needed.

Table 4.7. Selected guest molecules derived from the virtual screening against query **5**.

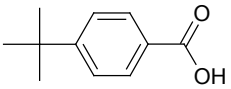
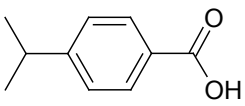
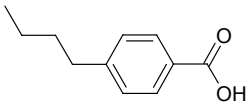
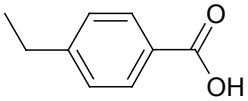
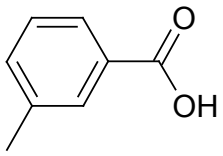
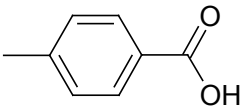
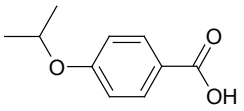
ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
5		15687-27-1	1.00	-	-22.6	-13.5	9.1
19		98-73-7	0.88	-23.2	-24.3	-20.5	3.8
20		8017-39-1	0.88	-20.3	-22.9	-29.3	-6.4
21		53483-12-8	0.88	-23.2	-15.9	-8.7	7.2

The concept of virtual screening should thus not be considered as a replacement of experiments but as an effective way to focus on promising molecules.

4.5 Conclusions

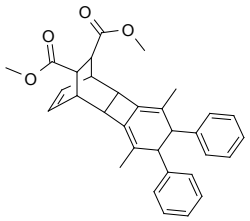
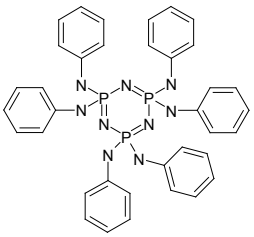
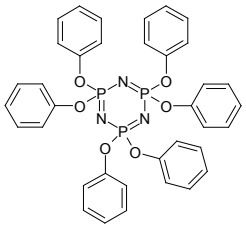
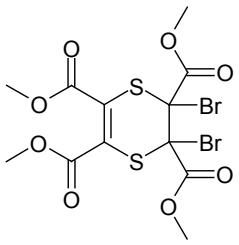
The results validate both the ligand-based screening approach for identifying novel compounds for a given synthetic receptor and the QSPR model for the prediction of binding affinities. Their combination is a promising high throughput alternative to structure-based virtual screenings for the identification of high affinity guests for given receptors. The methodology is faster than docking, allowing the screening of very large chemical libraries in a short time on a single CPU, and does not require knowledge of the receptor structure. While β-cyclodextrin was chosen as a test case because of its technical relevance and the availability of enough ligands with experimentally determined ΔG° the applied methodology can in principle be transferred to other systems. The quality of the results will

Table 4.8. Structural series of benzoic acid derivatives. The similarity was computed against compound **19**.

ID	Structure	CAS-No.	Sim.	ΔG°(pred.) [kJ mol^{-1}]	ΔG°(exp.) [kJ mol^{-1}]	ΔH°(exp.) [kJ mol^{-1}]	$T\Delta S^\circ$(exp.) [kJ mol^{-1}]
19		98-73-7	1.00	-23.2	-24.3	-20.5	3.8
22		536-66-3	0.96	-20.6	-19.7	-13.4	6.3
23		20651-71-2	0.92	-20.1	-21.4	-14.8	6.6
24		619-64-7	0.92	-17.2	-15.0	-9.2	5.7
25		99-04-7	0.88	-13.5	-6.2	-21.0	-14.7
26		99-94-5	0.88	-15.1	-11.0	-8.0	3.0
27		13205-46-4	0.83	-16.3	-16.1	-10.6	5.4

generally depend on the existence of sufficient experimental data for the generation of a reasonably accurate regression model.

Table 4.9. Some structures predicted as favorable binders by virtual screening solely based on the QSPR model.

ID	Structure	ZINC-ID	ΔG°(pred.) kJ mol^{-1}
28		ZINC00622780	-30.0
29		ZINC04552091	-29.6
30		ZINC04552092	-28.3
31		ZINC03263513	-25.7

5

QSPR Study on the Predictability of Thermodynamic Properties of Beta-Cyclodextrin Inclusion Complexes.

This chapter describes the comparison of three different statistical regression methods regarding their ability to establish reliable models for the prediction of thermodynamical parameters on the basis of computed molecular descriptors (Steffen & Apostolakis, 2007). Furthermore, a detailed analysis was performed in order to understand the differences in the predictabilities. As in the previous chapters, we focus on the host-guest systems between β-cyclodextrins as the host molecule and different guest molecules. The work of this chapter was accomplished in collaboration with Dr. Joannis Apostolakis, Ludwigs-Maximillian Universität München.

5.1 Introduction

Several attempts have been made to study and to predict the binding free energies ΔG° of cyclodextrin inclusion complexes by means of computational methods (Connors, 1997; Lipkowitz, 1998). Among them particularly statistical methods based on multiple linear regression (Suzuki, 2001; Katritzky et al., 2004) or neural nets (Liu & Guo, 1999) have proved to lead to robust prediction models. Here, we moreover investigate the predictability of two additional thermodynamical parameters that are of importance during complex formation in β-cyclodextrin based host-guest complexes, i. e. the enthalpy change (ΔH°) and the entropy change (ΔS°). This study is combined with a performance comparison of three different types of statistical regression methods, namely principal components regression (PCR) (see Section 5.2.3.1), partial least squares regression (PLSR) (see Section 5.2.3.2) and support vector regression with forward feature selection (SVMR/FFS) (see Section 5.2.3.3). Whereas the first two methods are well established in the field of chemoinformatics, the latter is a relatively new machine learning technique that has been successfully applied in some recent research projects. Briem & Günther (2005), for example, developed support vector

machine (SVM) models to predict the likeness of a molecular compound to be a kinase inhibitor, Jorissen & Gilson (2005) described the application of SVM models for virtual screenings, and Liu et al. (2006) employed SVMR to materials optimization of sialon ceramics.

5.2 Methodology

For the present study, a new dataset of β-cyclodextrin guest molecules was assembled from the literature.[1] For all molecules the three thermodynamical parameters ΔG°, ΔH° and $T\Delta S^\circ$ were available. This dataset served as a test and validation set for developing statistical prediction models with three different regression methods, PCR, PLSR and SVMR/FFS.

5.2.1 Assembling of the Dataset and Preparation of the Molecules

We assembled a dataset consisting of 176 β-cyclodextringuest molecules (see Appendix C.1). These molecules are a subset of those collected by Rekharsky & Inoue (1998). We applied the following selection criteria:

- The availability of experimental data derived from either calorimetric (cal) or UV-spectroscopic measurements
- The availability of ΔG°, ΔH° and $T\Delta S^\circ$ data
- All ligands with data that deviated from measurements of other groups were excluded.

We drew two-dimensional Lewis structures of the molecules with ISIS-Draw and exported them as MDL MOL files [MDL Information Systems, 1990-2002] The protonation state of each molecule was manually set according to the pH-value in which the measurement was performed. In the case no pH-data was available a reasonable state was set. We used CORINA (Sadowski & Gasteiger, 1993) for generating three-dimensional low-energy structures from the MOL-file. We converted the structures to SD-files (Dalby et al., 1992). Finally, all structures were manually inspected and - if needed - corrected.

5.2.2 Calculation and Processing of Molecular Descriptors

We calculated molecular descriptors for all molecules with the web service E-DRAGON, which is part of the Virtual Computational Chemistry Laboratory

[1] The dataset presented in Chapter 4 could not be used, since for those molecules only the ΔG° values were available.

(Tetko et al., 2005). As described in the previous chapter, E-Dragon calculates 1,666 different molecular descriptors (Todeschini & Consonni, 2000). These descriptors are grouped into different categories ranging from simple atom-type descriptors or fragment counts to more sophisticated topological, geometrical or quantum chemical descriptors. In order to prevent numerical problems and to ensure the avoidance of a bias in the descriptor space we normalized all descriptor values to the range between -1 and 1.

5.2.3 Regression Methods

The statistical methods used in this work exist in numerous implementations. For PCR and PLSR the R-package PLS was used (Wehrens & Mevik, 2007). The support vector machine regression was performed by means of LIBSVM, which was developed by Chang & Lin (2001).

5.2.3.1 Principal Component Regression

In principal component regression (PCR), multiple linear regression is performed on principal components. Principal components are linear combinations of the descriptors in the data matrix and explain their variance. They are derived from the covariance matrix of the calculated descriptors. The number of principal components corresponds to the rank of the data matrix. Its maximal value is the minimum of the number of data points (i. e. the molecules) and the number of descriptors. The first principal component of a data matrix points into the direction that maximizes the variance of the descriptors and corresponds to the eigenvector of the largest eigenvalue of the covariance matrix. The second principal component corresponds to the eigenvector of the second largest eigenvalue and points into the direction that maximizes the variance the data and is orthogonal to the first principal component, and so on for the remaining principal components. The PCR model is generated on a subset of the components. The subset is built by selecting the components in order of their ability to explain the variance in the dependent variable, i. e. in the current study the thermodynamical parameters.

5.2.3.2 Partial Least Squares Regression

Partial least squares regression is very similar to PCR. In contrast to PCR, where the covariance matrix of the data is used to generate the principal components, in PLSR the principal components are derived from the cross-covariance between the data matrix and the dependent variables (i. e. the quantity being predicted). Hence, while in PCR the eigenvectors of the data covariance matrix are used to

span the solution space; in PLSR the directions of maximal covariance between data and the dependent variables are used.

5.2.3.3 Support Vector Machine Regression

Support vector machine regression (SVMR) is a straightforward variant of support vector machines (SVM) classification (Cortes & Vapnik, 1995). In classification problems SVMs find the hyperplane that separates positive from negative examples with a maximum margin. This margin is defined as the distance of the closest data point from the separating hyperplane. In this way a statistical model is generated that only depends on a subset of the training data, namely those data points that are close enough to influence the size of the margin and the orientation of the hyperplane. These are the most difficult examples in the training set. They are called the support vectors, since they define the orientation of the separating plane. In support vector regression the same effect (namely that the final model depends only on a subset of the data) is achieved by the use of a so-called ϵ-insensitive cost function, which during model optimization ignores errors up to a defined threshold. In other words, any training data being predicted by the current model with an accuracy of up to ϵ can be neglected. As in Chapter 4 we added a so-called forward feature selection procedure, which is in some respect similar to the component extension in PCR and PLSR. Forward feature selection increases the learning performance and the interpretability of the regression model as only descriptors are selected that significantly improve the SVMR model. The selection of descriptors incurs combinatorial explosion if all possible subsets of all available descriptors would have to be considered. This, of course, is not feasible if the number of descriptors is too large. To overcome this problem forward feature selection uses the following greedy heuristic. For each single descriptor a support vector regression model is trained with tenfold cross-validation. The descriptor leading to the model with the highest r^2_{cv} of the predicted to the experimental values is selected as the start descriptor. Then this descriptor is combined with each of the remaining descriptors and the best pair is selected. This is repeated until the parameter r^2_{cv} reaches a maximum. This is used as a stopping criterion, at which the final model is obtained.

5.2.4 Internal Validation

In order to validate whether our model generation procedures can lead to a predictive model that provides reliable output, we performed the same nested threefold cross-validation protocol (Ruschhaupt et al., 2004) for each of the regression

methods as in Chapter 4. Therefore, we first split the whole training set into three equally sized subsets by randomly assigning molecules of the training dataset to one of the subsets (S1 and S2 consist of 59 molecules, S3 consists of 58 molecules). Then, we generated three validation sets, each as a combination of two subsets (V1 = S1 and S2, V2 = S1 and S3, V3 = S2 and S3), such that each of the validation sets can be used as a training set for predicting the binding energies of the remaining subset that is not included in the respective training set. We now distinguish between the inner loop and the outer loop of the validation. Within the inner loop models with increasing numbers of components (PCR/PLSR) or descriptors (SVMR/FFS) are built for each of the validation sets by means of the tenfold cross-validation protocol. In the outer loop we validate the component or respectively the descriptor selection procedure by predicting the subsets S1, S2 or S3 with the model of the inner loop, in which the respective subset was not included. This kind of nested validation produces a reliable estimate of the predictive power of our regression model for any molecule that is not included in the training set.

5.2.5 Calculation of Molecular Similarity and Clustering of the Molecules

For clustering the molecules of our datasets and for the nearest neighborhood analysis we calculated all pairwise molecular similarities by means of the graph alignment algorithm of the similarity tool GMA (see Section 4.3.2.1) (Marialke et al., 2007). The molecular similarity was calculated on the basis of a graph-based alignment on the atomic level. The better the molecular graphs, i. e. the topology and the atom types, of two molecules can be matched, the more similar these two molecules are (with 1=identical and 0=dissimilar). On the basis of these similarities we performed a complete-linkage hierarchical clustering. The cluster tree was cut off at a threshold of a similarity of 0.7. Hence, within one cluster only molecules are grouped that exhibit a similarity of 0.7 or higher.

5.3 Results and Discussion

In this work we studied the predictability of experimental thermodynamical data from 176 guest molecules of β-cyclodextrin. For all molecules experimental measurements for three fundamental thermodynamic quantities, i. e. the entropy change ($T\Delta S^\circ$), the enthalpy change (ΔH°) and the binding free energy (ΔG°), were present. Statistical models were developed to predict each of these parameters on the basis of computed molecular descriptors. We applied three different

types of regression methods - principal component regression (PCR), partial least squares regression (PLSR) and support vector machine regression with forward feature selection (SVMR/FFS). For the validation and the closer assessment of our models we performed tenfold cross-validation and a nested cross-validation protocol.

Comparison of the regression methods

In Table 5.1 the results of the cross-validations are detailed. We discuss the cross-validation parameter q^2, which includes the prediction errors.

$$q^2 = 1 - \frac{\sigma^2(\Delta y)}{\sigma^2(y)} \tag{5.1}$$

where q^2 is the cross-validation parameter, $\sigma^2(\dots)$ is the variance of the respective quantity in brackets, Δy is the deviation between predicted and experimental values, and y is the quantity being predicted (the experimental values).[2]

The highest cross-validation values q^2 when applying PCR to predict ΔG°, ΔH° and $T\Delta S^\circ$ are 0.71, 0.54 and 0.35, respectively (see Table 5.1). PLSR leads to models with maximal q^2 values for the three parameters of 0.74, 0.53 and 0.31, respectively. The highest q^2 values are obtained for SVMR/FFS with 0.89, 0.75 and 0.63, respectively.

The shape of the curve when plotting the number of components or descriptors, respectively, against q^2 is characteristic for each of the regression methods (see Table 5.2 - left column for a representative example). PLSR directly steers towards the maximal q^2 value and thus reaches its peak with only a few components. After this maximum, the q^2 value decreases slightly and stays on a plateau until it drastically drops down at one point. The curves for PCR look clearly different. The maximum of q^2 is reached with significantly more components and in-between local minima exist. The differences in the shape of the curves can be explained by the way the components are obtained. While in PLSR the components are derived from the cross-covariance between the descriptors and the predictors, in PCR the components are only derived from descriptor matrix. For SVMR/FFS the q^2 value increases continuously with each added descriptor until it reaches a plateau with the maximal q^2 value. This continuous increase of the q^2 value is due to the selection criterion of the FFS, which is to include the descriptor that shows the highest improvement to cross-validation performance.

[2] In this chapter we discuss the cross-validation parameters instead of the squared linear correlation coefficients and the RMSE values. The cross-validation parameter gives a direct view on the prediction errors and thus provides a concise means to discuss the quality of the prediction. The cross-validation parameter is bounded between $-\infty$ and 1. The closer it approaches to 1 the better.

Table 5.1. Comparison of the regression methods for tenfold cross-validation. The maximal q^2 values are reported for each thermodynamical parameter.

Regression method	ΔG° q^2(max)	ΔH° q^2(max)	$T\Delta S^\circ$ q^2(max)
PCR	0.71	0.54	0.35
PLSR	0.74	0.53	0.31
SVMR/FFS	0.89	0.75	0.63

For the validation of the statistical models we performed the nested cross-validation protocol described by Ruschhaupt et al. (2004). For each regression method this procedure was performed three times resulting in nine different models and prediction assessments.

The PCR model predicts ΔG° values of the molecules in the outer loop with a q^2 of 0.69±0.03 to the experimentally determined ΔG° values, PLSR performs with a q^2 of 0.69±0.03 and SVMR/FFS with a q^2 of 0.71±0.03 (see Tables 5.2 and 5.5). In the case of SVMR/FFS a drastic decrease of the q^2 of the inner loop in comparison to q^2 of the outer loop can be observed. The maximal obtained q^2 value in the inner loop is 0.87 whereas in the outer loop a q^2 of only 0.74 was found. PLSR and PCR show a more stable behavior and the q^2 values of the inner and the outer loops are comparable. It should, however, be noted that the correlations presented here for the prediction of ΔG° are clearly below the one found in Chapter 4 (r^2_{cv}=0.95). Since we applied the same methodology (SVMR/FFS) this finding is due to the different datasets.

The correlations obtained for the prediction of ΔH° and $T\Delta S^\circ$ (Tables 5.3 and , and Tables 5.4 and respectively) are clearly below the ones obtained for the prediction of ΔG° for all regression methods. For both, ΔH° and $T\Delta S^\circ$, none of the regression methods resulted in a q^2 of above 0.5 in the outer loop. This finding particularly shows the risk of overfitting the SVMR/FFS model to the data as in the tenfold cross-validation even for ΔH° and $T\Delta S^\circ$ comparably good correlations were obtained. The overfitting of the SVMR/FFS model is mainly due to the forward feature selection algorithm which uses r^2_{cv} for choosing the next descriptor in the iteration. Thus, the execution of a nested cross-validation is essential for getting a realistic estimate of the method's predictivity.

Table 5.2. Dependence of the cross-validation coefficient q^2 (ΔG°) on the number of components/descriptors integrated into a model for the inner and the outer loop of the nested cross-validation for all three methods.

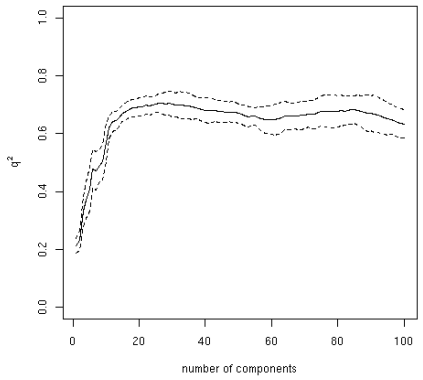
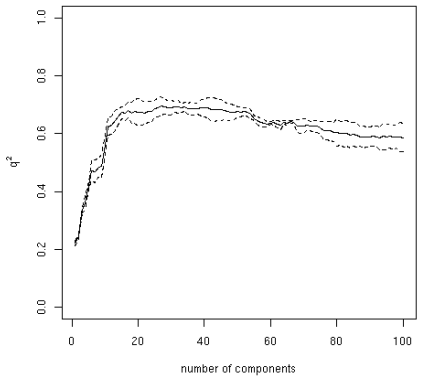
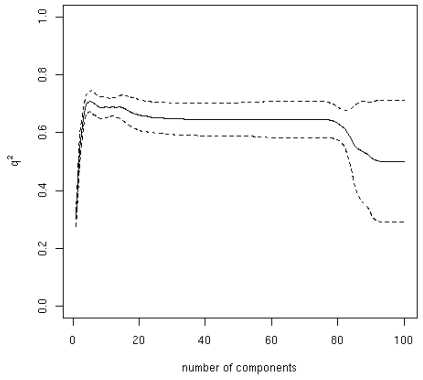
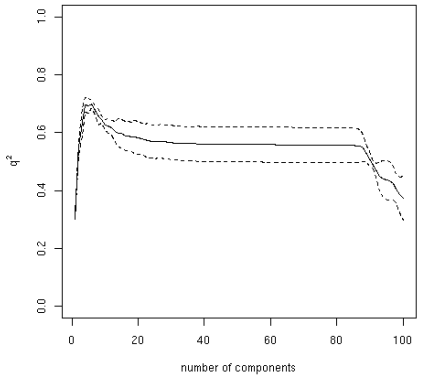
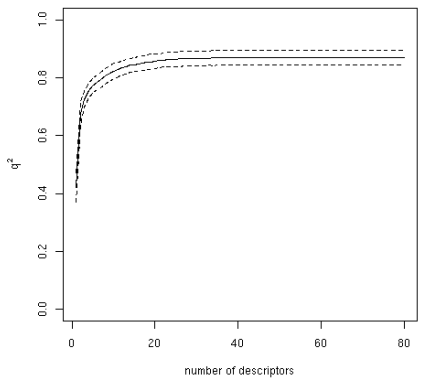
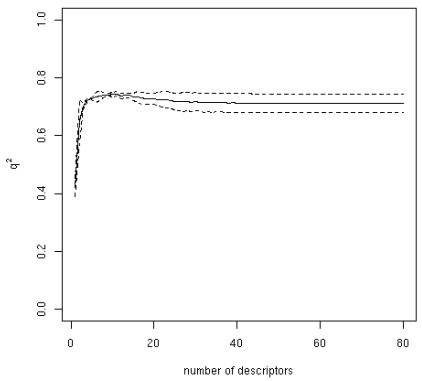
inner loop	outer loop
PCR	
	
PLSR	
	
SVM-R/FFS	
	

Table 5.3. Dependence of the cross-validation coefficient q^2 (ΔH°) on the number of components/descriptors integrated into a model for the inner and the outer loop of the nested cross-validation for all three methods.

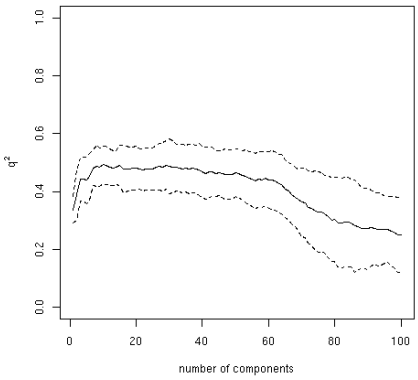
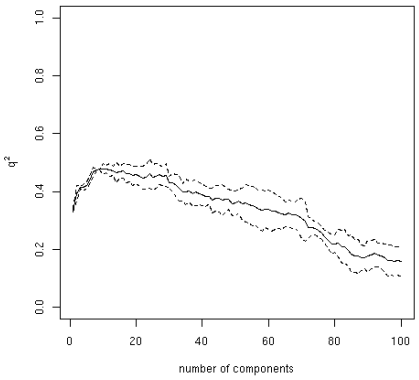
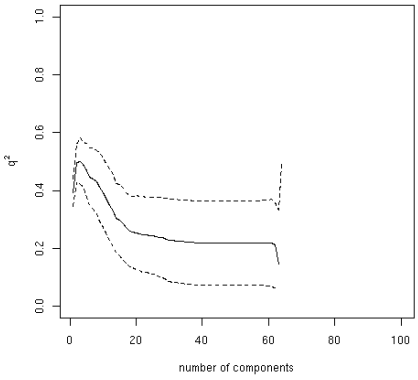
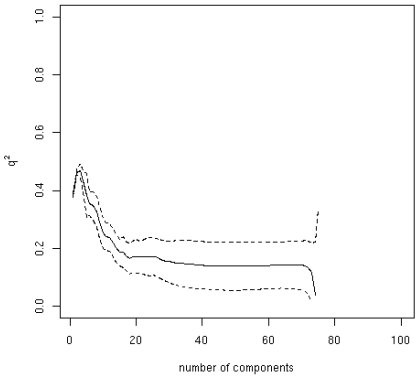
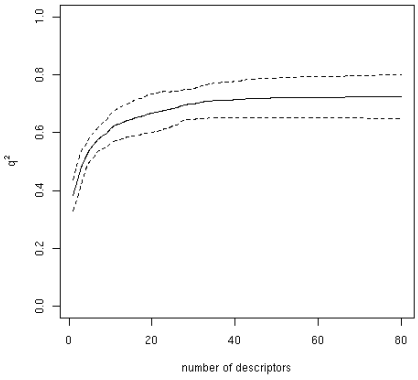
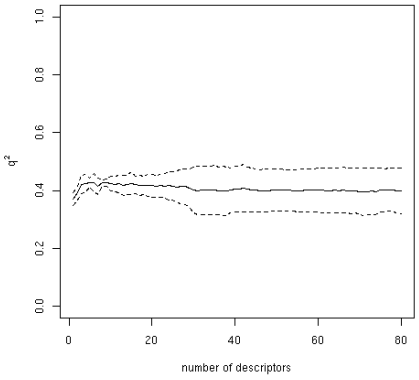
inner loop	outer loop
PCR	
	
PLSR	
	
SVM-R/FFS	
	

Table 5.4. Dependence of the cross-validation coefficient q^2 ($T\Delta S^\circ$) on the number of components/descriptors integrated into a model for the inner and the outer loop of the nested cross-validation for all three methods.

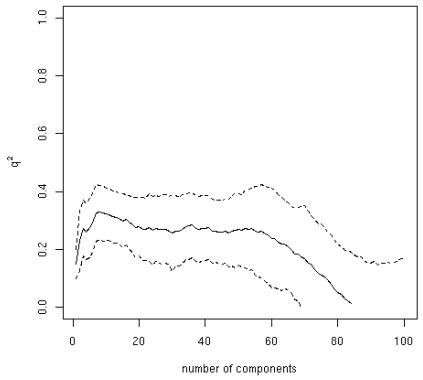
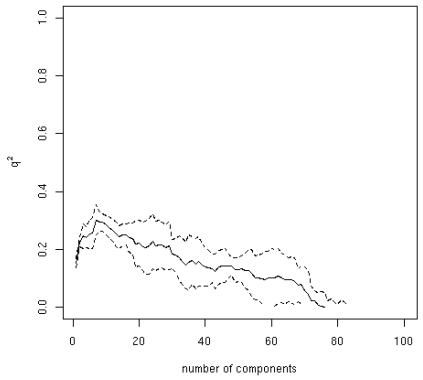
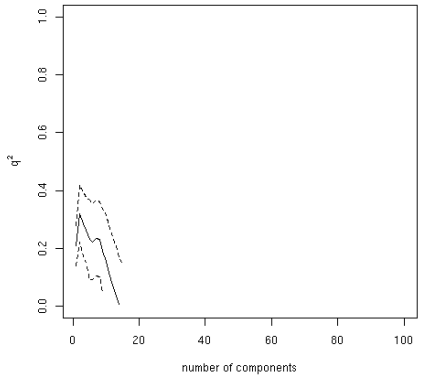
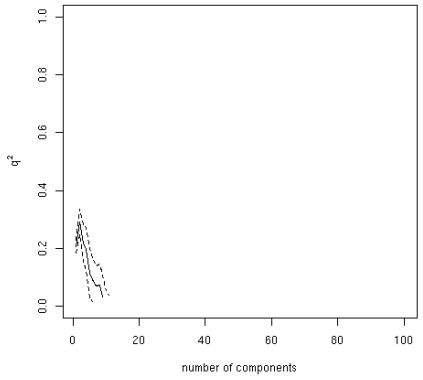
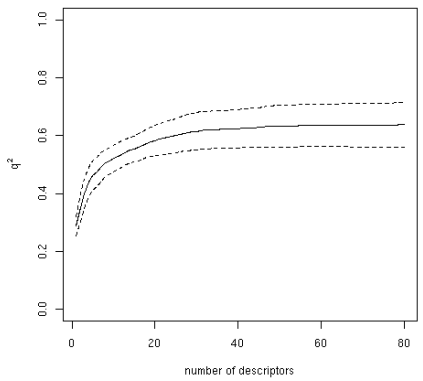
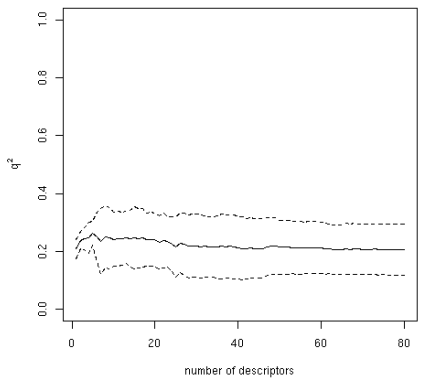
inner loop	outer loop
PCR	
	
PLSR	
	
SVM-R/FFS	
	

Table 5.5. Comparison of the regression methods for nested cross-validation (ΔG°). The maximal q^2 in the inner loop (q^2(max)-inner loop), the maximal q^2 in the outer loop (q^2(max)-outer loop) and the q^2 of the outer loop predicted by the model with the maximal q^2 in the inner loop (q^2(inner loop-max)-outer loop) are shown.

Regression method	q^2(max)-inner loop	q^2(max)-outer loop	q^2(inner loop-max)-outer loop
PCR	0.71±0.03	0.7±0.03	0.69±0.03
PLSR	0.71±0.03	0.7±0.01	0.69±0.03
SVMR/FFS	0.87±0.03	0.74±0.01	0.71±0.03

Table 5.6. Comparison of the regression methods for nested cross-validation (ΔH°). The maximal q^2 in the inner loop (q^2(max)-inner loop), the maximal q^2 in the outer loop (q^2(max)-outer loop) and the q^2 of the outer loop predicted by the model with the maximal q^2 in the inner loop (q^2(inner loop-max)-outer loop) are shown.

Regression method	q^2(max)-inner loop	q^2(max)-outer loop	q^2(inner loop-max)-outer loop
PCR	0.49±0.07	0.48±0.02	0.48±0.02
PLSR	0.5±0.08	0.47±0.02	0.47±0.02
SVMR/FFS	0.73±0.08	0.43±0.02	0.4±0.08

Table 5.7. Comparison of the regression methods for nested cross-validation ($T\Delta S^\circ$). The maximal q^2 in the inner loop (q^2(max)-inner loop), the maximal q^2 in the outer loop (q^2(max)-outer loop) and the q^2 of the outer loop predicted by the model with the maximal q^2 in the inner loop (q^2(inner loop-max)-outer loop) are shown.

Regression method	q^2(max)-inner loop	q^2(max)-outer loop	q^2(inner loop-max)-outer loop
PCR	0.33±0.09	0.3±0.05	0.3±0.03
PLSR	0.32±0.10	0.29±0.04	0.29±0.04
SVMR/FFS	0.64±0.08	0.26±0.04	0.21±-0.09

Predictability of different thermodynamic quantities

The relation between the three quantities is given by statistical thermodynamics on the one hand ($\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$), and the empirical finding of enthalpy-entropy compensation on the other (Sharp, 2001). In Figure 5.1(left) the difference of ΔH° and $T\Delta S^\circ$ is plotted against ΔG°. Except for two outliers (pentylthiobarbituricacid and cyclobarbital) all points are located on the diagonal. This indicates the consistency of the data. Furthermore we can observe the enthalpy-entropy compensation effect (see Figure 5.1 right).

Surprisingly, for all regression methods the best predictions were obtained for ΔG°. Particularly for $T\Delta S^\circ$ no predictive regression models could be generated with any of the methods. One possible reason for the different predictability of the three quantities has been given by Sharp (Sharp, 2001). In his analysis of the thermodynamics of three different protein systems, Sharp suggested that the most probable reason behind entropy-enthalpy compensation is the higher experimental error in the determination of ΔH° and $T\Delta S^\circ$ (Sharp, 2001). If ΔG° can be measured reliably, while there is significant error in the determination of the ΔH° and $T\Delta S^\circ$, then the last two quantities will vary significantly and in a correlated manner due to the thermodynamic equality $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$. This explanation would agree with the apparent difficulties, which we are facing in predicting ΔH° and $T\Delta S^\circ$ in comparison to ΔG°. Furthermore, it has been observed that experimental parameters have a significantly higher influence on $\Delta H^\circ$and $T\Delta S^\circ$ than on ΔG°. Ross et al., for example, measured the thermo-

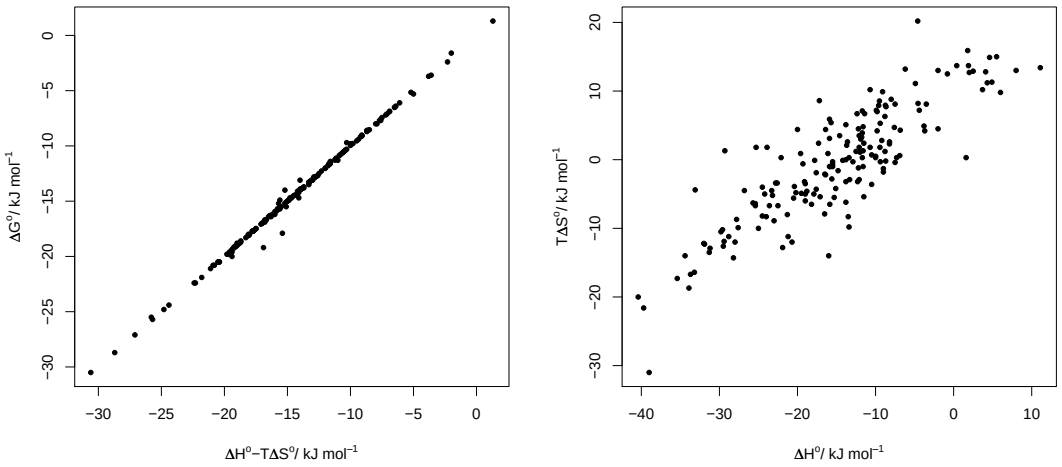


Figure 5.1. Plot of the experimental ΔG° values against the difference between the experimental values for ΔH° and $T\Delta S^\circ$ (left). Plot of the enthalpy-entropy compensation (right).

dynamic parameters of the complex between cyclohexanol and β-cyclodextrin in four different temperatures (288 - 318 K) (Ross & Rekharsky, 1996). While ΔG° is about the same in all measurements (16.3 ± 0.2 kJ mol^{-1}), the ΔH° values vary between -2.8 and -13.0 kJ mol^{-1} and $T\Delta S^\circ$ 13.2 and 3.6 kJ mol^{-1}. The stronger dependence of ΔH° and $T\Delta S^\circ$ on parameters of the experiment leads to higher errors, particularly when data from different laboratories is used. This was the case for the present study. Therefore, the explanation of different experimental accuracies appears plausible.

We analyzed differences in the thermodynamical properties of structurally closely related guest molecules, in order to obtain a more detailed view on the reasons for the poor predictability of ΔH° and $T\Delta S^\circ$. To extend the experimental data basis for this analysis, we integrated additional data if multiple measurements for a guest molecule were listed in the Rekharsky review (Rekharsky & Inoue, 1998). For those compounds for which we had independent data from different publications we calculated the standard deviations for ΔG°, ΔH° and $T\Delta S^\circ$, and averaged these over all compounds. The respective values are 1.8 kJ mol^{-1}, 2.1 kJ mol^{-1}, and 2.7 kJ mol^{-1}. Interestingly enough, the magnitude of these values is entirely consistent with the common practice of determining changes in entropy: the change of entropy is calculated from the difference of the measured change in enthalpy and the measured change in the binding free energy. If we assume independent errors in the two latter quantities, we can calculate the expected error of the change of entropy by means of the laws of error propagation - the root of the sum of the squares of the errors in enthalpy and entropy is 2.8 kJ mol^{-1}. Noteworthy, the magnitudes of the experimental errors found here are higher than what is generally reported in publications of experimental data. This is mainly because they include the systematic error arising from the compilation of data from different laboratories, whose experimental protocols most likely differ. The tendency of the errors certainly agrees with the predictability of the three quantities. However, the error is rather low compared to the overall average variance of the corresponding quantities. The data we used for the analysis varies with standard deviations of 5.3 kJ mol^{-1}, 9.6 kJ mol^{-1}, and 8.5 kJ mol^{-1} for ΔG°, ΔH° and $T\Delta S^\circ$. The average root mean square errors of the predicted to the experimental values obtained with SVMR/FFS are 2.8 kJ mol^{-1} (ΔG°), 7.5 kJ mol^{-1} (ΔH°) and 7.4 kJ mol^{-1} ($T\Delta S^\circ$). While the prediction of ΔG° appears to be limited mainly by the experimental error (2.8 kJ mol^{-1} compared to 1.8 kJ mol^{-1}), ΔH° and $T\Delta S^\circ$ are clearly poorly predicted, and this cannot be only explained by the slightly higher values of the experimental error.

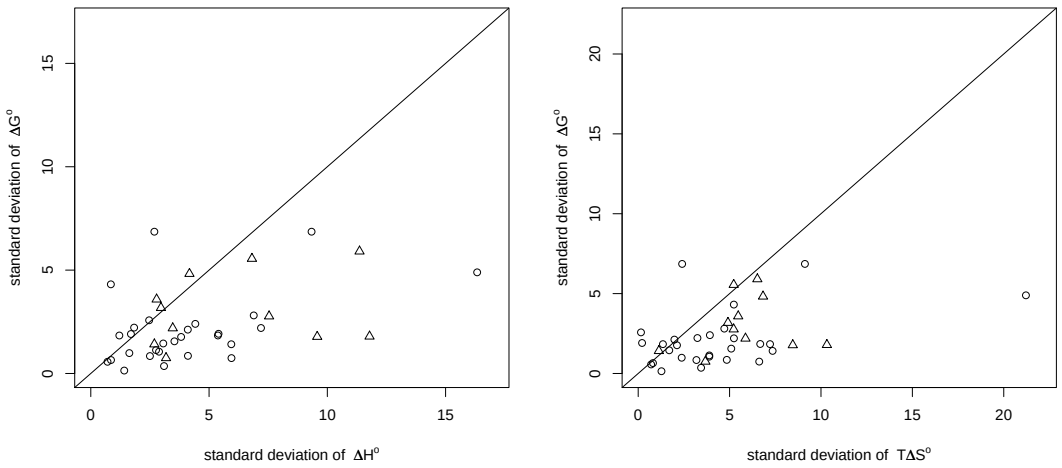


Figure 5.2. Plot of the standard deviations of ΔG° against the standard deviations of ΔH° (left side) and $T\Delta S^\circ$ (right side) for each cluster. Circles stand for clusters in which the experimental measurements all were performed within one laboratory. Triangles denote clusters containing data from different laboratories.

Next, we clustered the compounds of our dataset according to molecular similarity. Clusters were built using a similarity threshold of 0.7 with a complete linkage algorithm. In this way all structures within a cluster have a similarity of 0.7 or higher and thus are structurally closely related compounds (see Appendix C.2). We then calculated the mean values for ΔG°, ΔH° and $T\Delta S^\circ$ together with their standard deviations for all molecules within a cluster. In Figure we plot the standard deviations of ΔG° against the corresponding standard deviations of ΔH° and $T\Delta S^\circ$ within each cluster. In the majority of all cases, the points lie below the diagonal, indicating that the variance in the experimental ΔH° and $T\Delta S^\circ$ values is higher than the variance of the corresponding ΔG° values. This result indicates a higher dependence of the enthalpy and the entropy values on small structural changes in the ligand. This is nicely illustrated, for example, by the calorimetrically derived thermodynamic data for inclusion complexes of a range of sulfonamides (see Appendix Table C.2 - Cluster ID 39), which were all found in one similarity cluster and where studied within one laboratory. The standard deviation of the ΔG° values is relatively small with ± 1.8 kJ mol^{-1}. The corresponding standard deviations of ΔH° and $T\Delta S^\circ$ however, are clearly higher with $\pm$ 5.4 kJ mol^{-1} and $\pm$ 3.84 kJ mol^{-1}, respectively.

Additionally, we attempted a nearest-neighbor prediction of ΔG°, ΔH° and $T\Delta S^\circ$ using the graph-based similarity of the molecules. This method is independent from the E-Dragon descriptors and the regression methods. For each molecule within the dataset the three thermodynamic quantities were predicted

to be equal to those of the most similar compound within the set. We obtain q^2 values equal to 0.50 for ΔG°, 0.47 for ΔH° and 0.29 for $T\Delta S^\circ$. Except for a certain loss of accuracy in the prediction of ΔG°, the results are very similar to the results from the regression based prediction. The principal trend of the predictability of the thermodynamic quantities observed in the regression analysis can also be observed in this analysis and again $T\Delta S^\circ$ is the least predictable thermodynamic parameter. This analysis indicates that the poorer predictability of $T\Delta S^\circ$ (and to a lesser extent of ΔH°) for different ligands is due to a more complex dependence of $T\Delta S^\circ$ on even small structural changes of the ligand. This explanation is also consistent with the empirical observation of enthalpy-entropy compensation. The relative insensitivity of ΔG° to small structural changes compared to the other two quantities, would lead to the compensation effects in enthalpy and entropy due to the equation $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$, and inversely, given entropy enthalpy compensation, changes in entropy will lead to smaller changes in free energy.

5.4 Conclusion

In this work we investigated the predictability of three important thermodynamic quantities the free energy of binding, heat of formation and the entropy change upon binding. To this end, we chose β-cyclodextrin with its guest molecules - a very well studied system with a large amount of high quality binding data. We could show that free energies of binding can be reliably predicted by means of simple, commonly available molecular descriptors with all three linear regression methods studied in a comparable quality. The SVMR/FFS method has the advantage that it leads to a (partly) interpretable model with comparably few descriptors. However, in the application of SVMR/FFS it is important to perform a nested cross-validation in order to obtain a realistic impression of its generalization ability. The predictability of ΔG° obviously cannot be traced to the predictability of ΔH°, since the latter is reproduced with significantly lower accuracy by the models analyzed in this work. $T\Delta S^\circ$ appears almost unpredictable. An analysis of our results in the context of further data from the literature suggests that the poor predictability of $T\Delta S^\circ$ and, to a smaller extent, of ΔH° is due to a stronger dependence of those quantities on structural details of the complex and only to a lesser extent on the larger experimental error. This would also explain the well documented empirical finding of entropy-enthalpy compensation.

6

Summary and outlook

Within this thesis novel computational tools were developed, validated and applied that transfer the concepts of efficient virtual screening approaches from the field of medicinal chemistry to supramolecular chemistry. In the first part I described the development of a fast and reliable structure prediction tool for synthetic host-guest complexes. The method is based on the protein-ligand docking program FlexX. In contrast to protein-ligand docking both molecules, the synthetic receptor and the guest molecule, had to be tackled flexible. In order to handle this flexibility, I applied a novel docking strategy that uses an adaptive two-sided incremental construction algorithm which incorporates the structural flexibility of both, the guest molecule and the synthetic receptor. The algorithm follows an adaptive strategy, in which one molecule is expanded by attaching its next fragment in all possible torsion angles whereas the other (partially assembled) molecule serves as a rigid binding partner. Then the roles of the molecules are exchanged. Geometric filters are used to discard partial conformations that cannot realize a targeted interaction pattern derived in a graph-based precomputation phase. The process is repeated until the entire complex is built up. The algorithm was validated on a test dataset comprising ten complexes of synthetic receptors and ligands. The method generated near-native solutions compared to crystal structures. It is able to generate solutions generally within less than a minute and can be used as a virtual screening tool, e. g. for searching for suitable guest molecules for a given synthetic receptor in large databases of guests and vice versa.

In the second part of the thesis efficient computational techniques were applied for designing optimally interacting host-guest systems based on β-cyclodextrins. I reported on the computer-aided optimization of a synthetic receptor for a given guest molecule, based on inverse virtual screening of receptor libraries. As an example, a virtual set of β-cyclodextrin derivatives was generated as receptor candidates for the anticancer drug camptothecin. I applied docking tools to generate

camptothecin complexes of every candidate receptor. Scoring functions were used to rank all generated complexes. From the candidates within the top 10% of the derived ranking list candidates nine were selected for experimental verification. The stabilities of the camptothecin complexes obtained from solubility measurements of five of the nine β-cyclodextrin derivatives were significantly higher than for any other β-cyclodextrin derivative known from literature. The remaining four β-cyclodextrin derivatives were insoluble in water. In addition, corresponding mono-substituted β-cyclodextrin derivatives were synthesized, which also showed improved binding constants. Among them the 9-H-purine derivative was the best, being comparable to the investigated hepta-substituted β-cyclodextrins.

The third project focused on the identification of novel guest molecules for β-cyclodextrin. Here, I applied a combination of a similarity-based virtual screening technique with a quantitative structure property relationship model to retrieve new guest molecules with high affinity to β-cyclodextrin. Five known β-cyclodextrin guest molecules were chosen as query molecules. A subset of the ZINC database with 117,695 molecular entries served as the screening set. For all five query compounds a virtual screening was performed by means of FUZZEE - a graph-based molecular similarity algorithm. Ranking lists were derived from the similarity scores. The 150 best-ranking molecules of each of the ranking lists were then scored by means of a QSPR model. This model was built on the basis of 218 β-cyclodextrin guest molecules with experimentally determined binding data and 1,666 computed molecular descriptors with support vector machine regression. The best-scoring and most-promising molecules of the five screening runs that were commercially available were selected for experimental verification. Altogether 16 compounds were purchased and their binding free energy to β-cyclodextrin was determined by isothermal microcalorimetry. Ten molecules exhibited a binding free energy of about or lower than -20 kJ mol^{-1}. Five of these molecules even had a higher binding affinity than their corresponding query structures. Two compounds were insoluble; only one molecule did not show any complexation to β-cyclodextrin. This technique provides a new and very fast means for the design of synthetic host-guest complexes.

In the last chapter of this thesis, I investigated the predictability of three thermodynamic quantities related to complex formation. As a model system I chose the host-guest complexes of β-cyclodextrin with different guest molecules. A training dataset comprising 176 β-cyclodextrin guest molecules with experimentally determined thermodynamic quantities was taken from the literature. I compared the performance of three different statistical regression methods (principal component regression PCR, partial least squares regression PLSR and support

vector machine regression combined with forward feature selection SVMR/FSS) with respect to their ability to generate predictive quantitative structure property relationship models for ΔG°, ΔH°, $T\Delta S^\circ$ on the basis of computed molecular descriptors. SVMR/FFS marginally outperformed PLSR and PCR in the prediction of ΔG°. PLSR performed slightly better than PCR. PLSR and PCR proved to be more stable methods in a nested cross-validation protocol. Whereas ΔG° can be predicted in good agreement with experimental values, none of the methods led to comparably good predictive models for ΔH°. $T\Delta S^\circ$ appears almost unpredictable with the methods described here. I performed a detailed analysis in order to understand the differences in the predictabilities. As a result I could show that free energies are less sensitive to small structural variations of the guest molecules than enthalpy or entropy. This property and the lower sensitivity of ΔG° to different experimental conditions are possible reasons for its better predictability.

Within this thesis a focus was put onto supramolecular complexes based on hydrogen bonds on the one hand, and onto complexes involving β-cyclodextrins on the other. However, there are multiple other classes of supramolecular systems that can be studied by means of computational tools in future work. For example, a very exciting field of supramolecular chemistry is the field of molecularly imprinted polymers. Therein, a polymer is synthesized in the presence of a guest molecule. In this way, polymers posses cavities that are particularly tailored towards the guest molecule of interest. Until now, relatively little attention has been payed on the computer-assisted design of such polymers. A promising idea would be to computationally identify appropriate monomers, which show good interaction with the guest molecule. This could be achieved on the basis of a virtual screening, in which a dataset of monomers is screened against the guest molecule of interest. For this purpose I could apply the technology of FlexR, in principle. A further direction for future work is the computational design of template molecules for chemical reactions. These template molecules stabilize two substrates in a manner that facilitates the reaction between the substrates by stabilizing the transition state of the reaction. Here, a huge problem is the release of the product. If it is too tightly bound it inhibits any further reaction. A computational tool for the design of such template molecules should therefore balance between a high binding affinity of the template molecule to the reaction educts and a significantly lower one for the product.

In this thesis, I showed the usefulness of computational concepts taken from the field of medicinal chemistry for supramolecular chemistry. Conversely, computational chemistry tools for medicinal chemistry can benefit from supramolecular

chemistry. In general, contrasting the structural simplicity of synthetic receptors against the complex nature of proteins can afford fruitful insights. For example, the parametrization of scoring functions on experimental data of synthetic host-guest complexes instead of on experimentally determined protein-ligand complexes appears promising. In principle synthetic complexes could be designed that avoid complex interaction interferences, but rather focus on particular interactions.

Own Publications

Journal Papers

- A. Steffen, A. Kämper, and T. Lengauer (2006). Flexible docking of ligands into synthetic receptors using a two-sided incremental construction algorithm, *J. Chem. Inf. Model.*, 46, 1695.
- A. Steffen, C. Thiele, S. Tietze, C. Strassnig, A. Kämper, T. Lengauer, G. Wenz, and J. Apostolakis (2007). Improved cyclodextrin based receptors for camptothecin by inverse virtual screening, *Chem. Eur. J.*, 13, 6801.
- A. Steffen, M. Karasz, C. Thiele, A. Kämper, T. Lengauer, G. Wenz, and J. Apostolakis (2007). Similarity and QSPR based virtual screening technique for the identification of novel beta-cyclodextrin guest molecules, *New J. Chem.*, 31, 1941.
- A. Steffen and J. Apostolakis (2007). On the ease of predicting the thermodynamic properties of β-cyclodextrin inclusion complexes, *Chem. Cent. J.*, 1, 29.
- S. Raub, A. Steffen, A. Kämper, and C.M. Marian, AIScore - A chemically diverse empirical scoring function derived from quantum chemical binding energies of hydrogen-bonded complexes and experimental protein-ligand binding affinities, *J. Chem. Inf. Model.*, submitted.

Conference Proceedings

- A. Steffen, A. Kämper, and T. Lengauer, Virtual screening for guest molecules of a biosensor. In: Biosensors 2006: The ninth world congress on biosensors in Toronto, (Ed.) A.P.F. Turner. Elsevier, Oxford 2006, O66.
- A. Steffen, A. Kämper, and T. Lengauer, FlexR - A new tool for predicting the structure of synthetic host-guest complexes. In: Synthetic Receptors 2005: Second World Congress on Synthetic Receptors in Salzburg, (Ed.) A.P.F. Turner. Elsevier, Oxford 2005, O7.

Poster

- A. Steffen, A. Kämper, and T. Lengauer, Development of a new tool for predicting the structure of host-guest complexes In: Proceedings of the 2nd Summer School Medicinal Chemistry Regensburg 2004, Regensburg, Germany, 2004, 83.
- A. Steffen, A. Kämper, and T. Lengauer, FlexR - Predicting the structure of host-guest complexes In: Proceedings of the Annual Meeting of the German Pharmaceutical Society - Joint Meeting 2004 (DPhG 2004), Regensburg, Germany, 2004, 130.

References

Abagyan, R., Totrov, M., & Kuznetsov, D. (1994). ICM - a new method for protein modeling and design - applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15, 488.

Abraham, M. H. (1982). Free energies, enthalpies, and entropies of solution of gaseous nonpolar nonelectrolytes in water and nonaqueous solvents. The hydrophobic effect. *J. Am. Chem. Soc.*, 104, 2085.

Allen, F. H. (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr B*, 58, 380.

Apostolakis, J., Pluckthun, A., & Caflisch, A. (1998). Docking small ligands in flexible binding sites. *J. Comput. Chem.*, 19, 21.

Bajorath, J. (2002). Virtual screening in drug discovery: Methods, expectations and reality. *Curr. Drug. Discov.*, 2, 24.

Bajorath, J., Ed. (2004). *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, volume 275 of *Methods in Molecular Biology*. Totowa, NJ: Humana Press.

Ballester, P., Capo, M., Costa, A., Deya , P. M., Gomila, R., Decken, A., & Deslongchamps, G. (2001). Selective binding of cis-1,3,5-cyclohexane tricarboxylic acid vs its epimeric trans isomer by a tripodal amidopyridine receptor; Crystal structures of the 1:1 complexes. *Org. Lett.*, 3, 267.

Banerjee, R., Chakraborty, H., & Sarkar, M. (2004). Host-guest complexation of oxicam nsaids with beta-cyclodextrin. *Biopolymers*, 75(4), 355.

Barker, E. J., Buttar, D., Cosgrove, D. A., Gardiner, E. J., Kitts, P., Willett, P., & Gillet., V. J. (2006). Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.*, 46, 503.

Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., & Eldridge, M. D. (1998). Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Bioinf.*, 33, 367.

Bell, T. W., Hou, Z., Luo, Y., Drew, M. G., Chapoteau, E., Czech, B. P., & Kumar, A. (1995). Detection of creatinine by a designed receptor. *Science*, 269, 671.

Bell, T. W., Khasanov, A. B., & Drew, M. G. (2002). Role of pyridine hydrogen-bonding sites in recognition of basic amino acid side chains. *J. Am. Chem. Soc.*, 124, 14092.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Biochemistry*. W. H. Freeman.

Berl, V., Huc, I., Lehn, J.-M., DeCian, A., & Fischer, J. (1999). Induced fit selection of a barbiturate receptor from a dynamic structural and conformational/configurational library. *Eur. J. Org. Chem.*, 1999, 3089.

Betzel, C., Saenger, W., Hingerty, B. E., & Brown, G. M. (1984). Topography of cyclodextrin inclusion complexes. circular and flip-flop hydrogen-bonding in beta-cyclodextrin undecahydrate - a neutron-diffraction study. *J. Am. Chem. Soc.*, 106, 7545.

Böhm, H.-J. (1992). LUDI: Rule-based automatic design of new substituents for enzyme-inhibitor leads. *J. Comput.-Aided Mol. Des.*, 6, 593.

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

Böhm, H. J., Flohr, A., & Stahl, M. (2004). Scaffold hopping. *Drug Discov. Today: Technol.*, 1, 217.

Biwer, A., Antranikian, G., & Heinzle, E. (2002). Enzymatic production of cyclodextrins. *Appl. Microbiol. Biotechnol.*, 59, 609.

Blum, L. P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97, 245.

Bolten, B. M. & DeGregorio, T. (2002). From the analyst's couch. Trends in development cycles. *Nat. Rev. Drug Discov.*, 1, 335.

Booth, B. & Zemmel, R. (2004). Prospects for productivity. *Nat. Rev. Drug Discov.*, 3, 451.

Briem, H. & Günther, J. (2005). Classifying 'kinase inhibitor-likeness' by using machine-learning methods. *ChemBioChem*, 6, 558.

Bron, C. & Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16, 575.

Brooijmans, N. & Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, 32, 335.

Brown, N. D., Butler, D. L., & Chiang, P. K. (1993). Stabilization of thymopentin and preservation of its pharmacological properties by 2-hydroxypropyl-beta-cyclodextrin. *J. Pharm. Pharmacol.*, 45, 666.

Buckingham, A. D. (2000). *Recent Theoretical and Experimental Advances in Hydrogen Bonded Clusters*, chapter The hydrogen bond, (pp.1̃). Kluwer, Boston.

Bystroff, C. (2001). An alternative derivation of the equations of motion in torsion space for a branched linear chain. *Protein Eng.*, 14, 825.

Carlson, H. A. (2002). Protein flexibility is an important component of structure-based drug discovery. *Curr. Pharm. Des.*, 8, 1571.

Chang, C.-C. & Lin, C.-J. (2001). LibSVM: A library for support vector machines.

Chang, C.-E. & Gilson, M. K. (2003). Tork: Conformational analysis method for molecules and complexes. *J. Comput. Chem.*, 24, 1987.

Chen, H., Lyne, P. D., Giordanetto, F., Lovell, T., & Li, J. (2006). On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.*, 46, 401.

Chen, W. & Gilson, M. K. (2007). ConCept: De novo design of synthetic receptors for targeted ligands. *J. Chem. Inf. Model.*, 47, 425.

Chin, J., Walsdorff, C., Stranix, B., Oh, J., Chung, H. J., Park, S.-M., & Kim, K. A. (1999). Rational approach to selective recognition of NH_4^+ over K^+. *Angew. Chem. Int. Ed.*, 38, 2756.

Clark, M., Cramer, R. D., & Van Opdenbosch, N. (1989). Validation of the general-purpose TRIPOS 5.2 force-field. *J. Comp. Chem.*, 10, 982.

Claussen, H., Buning, C., Rarey, M., & Lengauer, T. (2001). FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308, 377.

Connors, K. (1997). The stability of cyclodextrin complexes in solution. *Chem. Rev.*, 97, 1325.

Conover, C. D., Greenwald, R. B., Pendri, A., Gilbert, C. W., & Shum, K. L. (1998). Camptothecin delivery systems: Enhanced efficacy and tumor accumulation of camptothecin following its conjugation to polyethylene glycol via a glycine linker. *Cancer Chemother. Pharmacol.*, 42, 407.

Cooke, G. & Rotellob, V. (2002). Methods of modulating hydrogen bonded interactions in synthetic host-guest systems. *Chem. Soc. Rev.*, 31, 275.

Corbellini, F., Knegtel, R. M. A., Grootenhuis, P. D. J., Crego-Calama, M., & Reinhoudt, D. N. (2004). Water-soluble molecular capsules: Self-assembly and binding properties. *Chem. Eur. J.*, 11, 298.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273.

Cram, D. (1988). The design of molecular hosts, guests, and their complexes (nobel lecture). *Angew. Chem. Int. Ed.*, 27, 1009.

Cramer, F. & Henglein, F. M. (1958). Determination of binding energies between cyclodextrins and aromatic guest molecules by microcalorimetry. *Chem. Ber.*, 91, 308.

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A., & Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.*, 32, 244.

Davis, ME; Brewster, M. (2004). Cyclodextrin-based pharmaceutics: Past, present and future. *Nat. Rev. Drug Discov.*, 3, 1023.

de Jong, M. R., Knegtel, R. M. A., Grootenhuis, P. D. J., Huskens, J., & Reinhoudt, D. N. (2002). A method to identify and screen libraries of guests that complex to a synthetic host. *Angew. Chem. Int. Ed.*, 41, 1004.

Degen, J. & Rarey, M. (2006). FlexNovo: Structure-based searching in large fragment spaces. *ChemMedChem*, 1, 854.

Deo, A. S. & Walker, I. D. (1995). Overview of damped least-squares methods for inverse kinematics of robot manipulators. *J. Intell. Robot. Syst.*, 14, 43.

Doman, T. N., McGovern, S. L., Witherbee, B. J., Kasten, T. P., Kurumbail, R., Stallings, W. C., Connolly, D. T., & Shoichet, B. K. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1b. *J. Med. Chem.*, 45, 2213.

Dougherty, D. (1996). Cation-π interactions in chemistry and biology: A new view of benzene, PHE, TYR, and TRP. *Science*, 271, 163.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, volume 9 (pp. 155).: The MIT Press.

D'Souza, V. & Lipkowitz, K. (1998). Introduction. *Chem. Rev.*, 98, 1741.

Eisen, M. B., Wiley, D. C., Karplus, M., & Hubbard, R. E. (1994). HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Struct., Funct., Genet.*, 19, 199.

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

Ertl, B., Platzer, P., Wirth, M., & Gabor, F. (1999). Poly-(D,L-lactic-co-glycolic acid) microspheres for sustained delivery and stabilization of camptothecin. *J. Controlled Release*, 61, 305.

Evans, D. A. & Neidle, S. (2006). Virtual screening of DNA minor groove binders. *J. Med. Chem.*, 49, 4232.

Ewing, T. J. A. & Kuntz, I. D. (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.*, 18, 1175.

Fenyvesi, E., Shirakura, O., Szejtli, J., & Nagai, T. (1984a). Properties of cyclodextrin polymer as a tabletting aid. *Chem. Pharm. Bull. (Tokyo)*, 32, 665.

Fenyvesi, E., Takayama, K., Szejtli, J., & Nagai, T. (1984b). Evaluation of cyclodextrin polymer as an additive for furosemide tablet. *Chem. Pharm. Bull. (Tokyo)*, 32, 670.

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

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, 27, 2985.

Fugen, G. & Cuijing, L. (1998). The preparation of inclusion compound of diclofenac sodium-β-cyclodextrin. *Chin. Pharm. J.*, 33, 153.

Garcia-Tellado, F., Goswami, S., Chang, S.-K., Geib, S. J., & Hamilton, A. D. (1990). Molecular recognition: A remarkably simple receptor for the selective complexation of dicarboxylic acids. *J. Am. Chem. Soc.*, 112, 7393.

Gemmel, E., Beck, H., & Bolte, M. und Eger, E. (1999). MOMO - molecular modelling program, version 2.00. Universität Frankfurt.

Gerstein, M. & Krebs, W. (1998). A database of macromolecular motions. *Nucl. Acids Res.*, 26, 4280.

Gilson, M. K., Given, J. A., Bush, B. L., & McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.*, 72, 1047.

Godinez, L. A., Patel, S., Criss, C. M., & Kaifer, A. E. (1995). Calorimetric studies on the complexation of several ferrocene derivatives by alpha-cyclodextrin and

beta-cyclodextrin - effects of urea on the thermodynamic parameters. *J. Phys. Chem.*, 99, 17449.

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

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

Graf, E. & Lehn, J.-M. (1975). Synthesis and cryptate complexes of a spheroidal macrotricyclic ligand with octahedrotetrahedral coordination. *J. Am. Chem. Soc.*, 97, 5022.

Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, 17, 1653.

Halgren, T. A. (1998). Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *Abstr. Pap. Am. Chem. Soc.*, 216, U702.

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

Halgren, T. A. (1999b). MMFF vi. MMFF94s option for energy minimization studies. *J. Comput. Chem.*, 20, 720.

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

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

Hamilton, A. D. & van Engen, D. (1987). Induced fit in synthetic receptors: nucleotide base recognition by a molecular hinge. *J. Am. Chem. Soc.*, 109, 5035.

Hansch, C. & Fujita, T. (1964). $\rho - \sigma - \pi$ analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, 86, 1616.

Harada, A. & Kataoka, K. (1999). Chain length recognition: core-shell supramolecular assembly from oppositely charged block copolymers. *Science*, 283, 65.

Hardee, G. E., Otagiri, M., & Perrin, J. H. (1978). Micro-calorimetric investigations of pharmaceutical complexes. 1. Drugs and beta-cyclodextrin. *Acta Pharm. Suec.*, 15, 188.

Harrison, J. C. & Eftink, M. R. (1982). Cyclodextrin adamantanecarboxylate inclusion complexes - a model system for the hydrophobic effect. *Biopolymers*, 21, 1153.

Herm, M., Molt, O., & Schrade, T. (2001). Towards synthetic adrenaline receptors-shape-selective adrenaline recognition in water. *Angew. Chem. Int. Ed.*, 40, 3148.

Herm, M. & Schrader, T. (2000). Towards synthetic adrenaline receptors. *Chemistry*, 6, 47.

Höfler, T. & Wenz, G. (1996). Determination of binding energies between cyclodextrins and aromatic guest molecules by microcalorimetry. *J. Inclusion Phenom. Mol. Recognit. Chem.*, 25, 81.

Hodges, M. P., Stone, A. J., & Xantheas, S. S. (1997). Contribution of many-body terms to the energy for small water clusters: A comparison of ab initio calculations and accurate model potentials. *J. Phys. Chem. A*, 101, 9163.

Hof, F., Trembleau, L., Ullrich, E. C., & Rebek, J. (2003). Acetylcholine recognition by a deep, biomimetic pocket. *Angew. Chem. Int. Ed.*, 42, 3150.

Hohenberg, P. & Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev. B*, 136, 864.

http://autodock.scripps.edu/ (2007). Autodock 4. website.

Hunter, C. & Sanders, J. (1990). The nature of pi-pi interactions. *J. Am. Chem. Soc.*, 112, 5525.

Irwin, J. J. & Shoichet, B. K. (2005). ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45, 177.

Jain, A. N. (2003). SurFlex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, 46, 499.

Jenkins, J. L., Kao, R. Y. T., & Shapiro, R. (2003). Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. *Proteins: Struct., Funct., Bioinf.*, 50, 81.

Jiang, F. & Kim, S. H. (1991). "Soft docking": Matching of molecular surface cubes. *J. Mol. Biol.*, 219, 79.

Jones, G., Willett, P., & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245, 43.

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

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

Jorissen, R. N. & Gilson, M. K. (2005). Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.*, 45(3), 549.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32, 922.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: New developments in KEGG. *Nucl. Acids Res.*, 34, D354.

Kang, J., Kumar, V., Yang, D., Chowdhury, P. R., & Hohl, R. J. (2002). Cyclodextrin complexation: influence on the solubility, stability, and cytotoxicity of camptothecin, an antineoplastic agent. *Eur. J. Pharm. Sci.*, 15, 163.

Karasz, M., Koerner, R., Marialke, M., Tietze, S., & Apostolakis, A. (2004). In E. A. Sener & I. Yalcin (Eds.), *Proceedings of the 15th European Symposium*

on Quantitative Structure Activity Relationships & Molecular Modelling (pp. 493). Turkey: CADD&Society Turkey.

Karginov, V. A., Hecht, S. M., Fahmi, N., & Aliben, K. (2006). Beta-cyclodextrin derivatives and their use against anthrax lethal toxin. WO 2006/001844 A2.

Katritzky, A. R., Fara, D. C., Yang, H., Karelson, M., Suzuki, T., Solov'ev, V. P., & Varnek, A. (2004). Quantitative structure-property relationship modeling of beta-cyclodextrin complexation free energies. *J. Chem. Comput. Sci.*, 44, 529.

Kellenberger, E., Rodrigo, J., Muller, P., & Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins Struct. Func. Bioinf.*, 57, 225.

Kitae, T., Nakayama, T., & Kano, K. (1998). Chiral recognition of alpha-amino acids by charged cyclodextrins through cooperative effects of coulomb interaction and inclusion. *Perkin Trans. 2*, 2, 207.

Kitagawa, M., Hoshi, H., Sakurai, M., Inoue, Y., & Chujo, R. (1987). The large dipole-moment of cyclomaltohexaose and its role in determining the guest orientation in inclusion complexes. *Carbohydr. Res.*, 163, C1–C3.

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

Klebe, G. (2006). Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discov. Today*, 11, 580.

Klebe, G. & Mietzner, T. (1994). A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.*, 8, 583.

Kämper, A., Apostolakis, J., M.Rarey, Marian, C. M., & Lengauer, T. (2006). Fully automated flexible docking of ligands into flexible synthetic receptors using forward and inverse docking strategies. *J. Chem. Inf. Model.*, 46, 903.

Kämper, A., Rognan, D., & Lengauer, T. (2007). Lead identification by virtual screening. In T. Lengauer (Ed.), *Bioinformatics - From Genomes to Therapies*. Weinheim: Wiley-VCH.

Kneeland, D. M., Ariga, K., Lunch, V. M., Huang, C.-Y., & Anslyn, E. V. (1993). Bis(alkylguanidinium) receptors for phosphodiesters: Effect of counterions, solvent mixtures, and cavity flexibility on complexation. *J. Am. Chem. Soc.*, 115, 10042.

Koehler, J. E., Saenger, W., & van Gunsteren, W. F. (1988). Conformational differences between alpha-cyclodextrin in aqueous solution and in crystalline form. a molecular dynamics study. *J. Mol. Biol.*, 203, 241.

Kohn, W. & Sham, L. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev. A*, 140, 1133.

Koshland, D. E. (1994). Das Schlüssel-Schloss-Prinzip und die Induced-Fit-Theorie. *Angew. Chem.*, 106, 2468.

Kramer, B., Rarey, M., & Lengauer, T. (1999). Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Bioinf.*, 37, 228.

Krebs, W. G., Tsai, J., Alexandrov, V., Junker, J., Jansen, R., & Gerstein, M. (2003). Tools and databases to analyze protein flexibility; approaches to mapping implied features onto sequences. *Methods Enzymol.*, 374, 544.

Kubinyi, H. (1999). Chance favors the prepared mind–from serendipity to rational drug design. *J. Recept. Signal. Transduct. Res.*, 19, 15.

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

Lahana (1999). How many leads from HTS? *Drug Discov. Today*, 4, 447.

Lamb, M. L. & Jorgensen, W. (1997). Computational approaches to molecular recognition. *Curr. Opin. Chem. Biol.*, 1, 449.

Lavigne, J. J. & Anslyn, E. V. (2001). Sensing a paradigm shift in the field of molecular recognition: From selective to differential receptors. *Angew. Chem. Int. Ed.*, 40, 3118.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235, 345.

Leach, A. R. & Gillet, V. J. (2003). *An Introduction to Chemoinformatics*. Springer, Netherlands.

Lehn, J. (1988). Supramolecular chemistry - scope and perspectives molecules, supermolecules, and molecular devices. *Angew. Chem. Int. Ed.*, 27, 89.

Lehn, J. (1995). *Supramolecular Chemistry: Concepts and Perspectives*. VCH.

Lemmen, C., Hiller, C., & Lengauer, T. (1998a). RigFit: a new approach to superimposing ligand molecules. *J. Comput.-Aided Mol. Des.*, 12, 491.

Lemmen, C., Lengauer, T., & Klebe, G. (1998b). FlexS: a method for fast flexible ligand superposition. *J. Med. Chem.*, 41, 4502.

Lengauer, T., Lemmen, C., Rarey, M., & Zimmermann, M. (2004). Novel technologies for virtual screening. *Drug Discov. Today*, 9, 27.

Lipkowitz, K. (1998). Applications of computational chemistry to the study of cyclodextrins. *Chem. Rev.*, 98, 1829.

Liu, L. & Guo, Q. X. (1999). Wavelet neural network and its application to the inclusion of beta-cyclodextrin with benzene derivatives. *J. Chem. Inf. Comput. Sci.*, 39, 133.

Liu, L. F., Desai, S. D., Li, T.-K., Mao, Y., Sun, M., & Sim, S.-P. (2000). Mechanism of action of camptothecin. *Annals of the New York Academy of Sciences*, 922, 1.

Liu, X., Lu, W. C., Jin, S. L., Li, Y. W., & Chen, N. Y. (2006). Support vector regression applied to materials optimization of sialon ceramics. *Chemometr. Intell. Lab. Sys.*, 82, 8.

Lorber, D. M. & Shoichet, B. K. (1998). Flexible ligand docking using conformational ensembles. *Protein Sci.*, 7, 938.

Lundberg, B. B. (1998). Biologically active camptothecin derivatives for incorporation into liposome bilayers and lipid emulsions. *Anti-Cancer Drug Des.*, 13, 453.

MacKerell, A. D., J., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy,

D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., I., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102, 3586.

Mangoni, M., Roccatano, D., & Nola, A. D. (1999). Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins: Struct., Funct., Bioinf.*, 35(2), 153.

Marialke, J., Körner, R., Tietze, S., & Apostolakis, J. (2007). Graph based molecular alignment. *J. Chem. Inf. Model.*

Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activities? *J. Med. Chem.*, 45, 4350.

McGann, M., Almond, H., Nicholls, A., Grant, J., & Brown, F. (2003). Gaussian docking functions. *Biopolymers*, 68, 76.

McPhail, A. & Sim, G. (1968). The structure of camptothecin: X-ray analysis of camptothecin iodoacetate. *J. Chem. Soc. B*, (pp. 923).

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, New York.

Müller, A. & Wenz, G. (2007). Thickness recognition of bolaamphiphiles by alpha-cyclodextrin. *Chem. Eur. J.*, 13, 2218.

Mo, Y. (2006). Probing the nature of hydrogen bonds in DNA base pairs. *J. Mol. Model.*, 12, 665.

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

Muderawan, I. W., Ong, T.-T., & Ng, S.-C. (2006). Urea bonded cyclodextrin derivatives onto silica for chiral HPLC. *J. Sep. Sci.*, 29, 1849.

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

Murray, C. W., Baxter, C. A., & Frenkel, A. D. (1999). The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.*, 13, 547.

Ong, J. K., Sunderland, V. B., & McDonald, C. (1997). Influence of hydroxypropyl beta-cyclodextrin on the stability of benzylpenicillin in chloroacetate buffer. *J. Pharm. Pharmacol.*, 49, 617.

Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J., & Goodsell, D. S. (2002). Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Struct., Func., Genet.*, 46, 34.

Otto, S. (2006). Reinforced molecular recognition as an alternative to rigid receptors. *Dalton Trans.*, (pp. 2861).

Paiva, A. M., Vanderwall, D. E., Blanchard, J. S., Kozarich, J. W., Williamson, J. M., & Kelly, T. M. (2001). Inhibitors of dihydrodipicolinate reductase, a

key enzyme of the diaminopimelate pathway of mycobacterium tuberculosis. *Biochim. Biophys. Acta*, 1545, 67.

Pascal, R. A. & Ho, D. M. (1994). Molecular structures of host-guest complexes with rebek's di-acid. *Tetrahedron*, 50.

Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., & Weinberger, L. E. (1996). Neighborhood behavior: A useful concept for validation of 'molecular diversity' descriptors. *J. Med. Chem.*, 39, 3049.

Pedersen, C. (1988). The discovery of crown ethers (noble lecture). *Angew. Chem. Int. Ed.*, 27, 1021.

Polgár, T., Baki, A., Szendrei, G. I., & Keseru, G. M. (2005). Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. *J. Med. Chem.*, 48, 7946.

Polgár, T. & Keserü, G. M. (2006). Ensemble docking into flexible active sites. critical evaluation of flexe against JNK-3 and beta-secretase. *J. Chem. Inf. Model.*, 46, 1795.

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

Rarey, M., Degen, J., & Reulecke, I. (2007). Docking and scoring for structure-based drug design. In T. Lengauer (Ed.), *Bioinformatics - From Genomes to Therapies* chapter Docking and scoring for structure-based drug design. Weinheim: Wiley-VCH.

Rarey, M. & Dixon, J. S. (1998). Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.*, 12, 471.

Rarey, M., Kramer, B., & Lengauer, T. (1997). Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.*, 11, 369.

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

Rarey, M. & Stahl, M. (2001). Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.*, 15, 497.

Rarey, M., Wefing, S., & Lengauer, T. (1996b). Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.*, 10, 41.

Raub, S. & Marian, C. M. (2007). Quantum chemical investigation of hydrogen-bond strengths and partition into donor and acceptor contributions. *J. Comput. Chem.*, (pp. accepted).

Raub, S., Steffen, A., Kämper, A., & Marian, C. (2007). AIScore - A chemically diverse empirical scoring function derived from quantum chemical binding energies of hydrogen-bonded complexes and experimental protein-ligand binding affinities. *J. Med. Chem.*, to be submitted.

Rekharsky, M. V. & Inoue, Y. (1998). Complexation thermodynamics of cyclodextrins. *Chem. Rev.*, 98, 1875.

Roche, O., Schneider, P., Zuegge, J., Guba, W., Kasny, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E.-M., Rogers-Evans, M., Neidhart, W., Stalder, H.,

Dillon, Michael andn Sjögren, E., Fotouhi, Nader andn Gillespie, P., Goodnow, R., Harris, William andn Jones, P., Taniguchi, M., Tsujii, S., von der Saal, W., Zimmermann, G., & Schneider, G. (2002). Development of a virtual screening method for identification of 'frequent hitters' in compound libraries. *J. Med. Chem.*, 45, 137.

Rogers, J. P., Beuscher, A. E., Flajolet, M., McAvoy, T., Nairn, A. C., Olson, A. J., & Greengard, P. (2006). Discovery of protein phosphatase 2C inhibitors by virtual screening. *J. Med. Chem.*, 49, 1658.

Ross, P. D. & Rekharsky, M. V. (1996). Thermodynamics of hydrogen bond and hydrophobic interactions in cyclodextrin complexes. *Biophys. J.*, 4, 2144.

Ruschhaupt, M., Huber, W., Poustka, A., & Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*.

Rush, T. S., Grant, J. A., Mosyak, L., & Nicholls, A. (2005). A shape-based 3d scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.*, 48, 1489.

Sadowski, J. & Gasteiger, J. (1993). From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.*, 93, 2567.

Sadowski, J., Rudolph, C., & Gasteiger, J. (1992). The generation of 3D models of host-guest complexes. *Anal. Chim. Acta*, 265, 233.

Saenger, W., Betzel, C., Hingerty, B., & Brown, G. M. (1983). Flip-flop hydrogen bonds in β-cyclodextrin – a generally valid principle in polysaccharides? *Angew. Chem. Int. Ed.*, 22, 883.

Sakurai, M., Kitagawa, M., Hoshi, H., Inoue, Y., & Chujo, R. (1990). A molecular-orbital study of cyclodextrin (cyclomalto-oligosaccharide) inclusion complexes .3. dipole-moments of cyclodextrins in various types of inclusion complex. *Carbohydr. Res.*, 198, 181.

Sandak, B., Wolfson, H. J., & Nussinov, R. (1998). Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins: Struct., Funct., Bioinf.*, 32, 159.

Sanner, M. F. (1999). Python: A programming language for software integration and development. *J. Mol. Graph. Model.*, 17, 57.

Schmid, G. (1991). *New trends in cyclodextrins and derivatives*. Paris.

Schneider, G. & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.*, 4, 649.

Schneider, H. (1991). Mechanismen der molekularen erkennung - untersuchung an organischen wirt-gast-komplexen. *Angew. Chem.*, 103, 1419.

Schneider, H.-J. & Yatsimirsky, A. (2000). *Principles and Methods in Supramolecular Chemistry*. John Wiley & Sons Ltd.

Schrader, T. (1996). Towards synthetic adrenaline receptors - strong binding of amino alcohols by bisphosphonates. *Angew. Chem. Int. Ed.*, 35, 2649.

Schrader, T. (1998). Toward synthetic adrenaline receptors: strong, selective, and biomimetic recognition of biologically active amino alcohols by bisphosphonate receptor molecules. *J. Org. Chem.*, 63, 264.

Schrader, T. & Hamilton, A. D. (2005). *Functional Synthetic Receptors*. Weinheim: Wiley-VCH.

Schrödinger, E. (1926a). Quantisierung als Eigenwertproblem (Dritte Mitteilung: Störungstheorie, mit Anwendung auf den Starkeffekt der Balmerlinien). *Ann. Phys.*, 80, 437.

Schrödinger, E. (1926b). Quantisierung als Eigenwertproblem (Erste Mitteilung). *Ann. Phys.*, 79, 361.

Schrödinger, E. (1926c). Quantisierung als Eigenwertproblem (Vierte Mitteilung). *Ann. Phys.*, 81, 109.

Schrödinger, E. (1926d). Quantisierung als Eigenwertproblem (Zweite Mitteilung). *Ann. Phys*, 79, 489.

Sharp, K. (2001). Entropy-enthalpy compensation: Fact or artifact? *Protein Sci.*, 10, 661–667.

Sheridan, R. P. & Kerarley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discov. Today*, 7, 903.

Shoichet, B., Leach, A., & Kuntz, I. (1999). Ligand solvation in molecular docking. *Proteins: Struct., Funct., Bioinf.*, 34, 4.

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*, 432, 862.

Slichenmyer, W. J., Rowinsky, E. K., Donehower, R. C., & Kaufmann, S. H. (1993). The current status of camptothecin analogs as antitumor agents. *J. Nat. Cancer Inst.*, 85, 271.

Smith, D. A. (2002). High-throughput screening–brains versus brawn. *Ernst Schering Res. Found. Workshop*, 37, 203.

Smithrud, D., Sanford, E. M., Chao, I., Ferguson, S. B., Carcanague, D. R., Evanseck, J. D., Houk, K. N., & Diederich, F. (1990). Solvent effects in molecular recognition. *Pure & Appl. Chem.*, 62, 2227.

Söntgen, O. (2003). *Entwicklung von Methoden zur Konformationsanalyse Supramolekularer Komplexe in Kraftfeldprogrammen*. PhD thesis, Johann Wolfgang Goethe-Universität Frankfurt.

Solov'ev, Varnek, & Wipff (2000). Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci.*, 40, 847.

Stanton, J. & Vincent, P. (2001). Tumor necrosis factor receptor 2. US 6673908.

Steed, J. W. & Atwood, J. L. (2000). *Supramolecular Chemistry*. Wiley.

Steffen, A. & Apostolakis, J. (2007). On the ease of predicting the thermodynamic properties of beta-cyclodextrin inclusion complexes. *Chem. Centr. J.*, 1, 29.

Steffen, A., Karasz, M., Thiele, C., Kämper, A., Lengauer, T., Wenz, G., & Apostolakis, J. (2007a). Similarity and QSPR based virtual screening technique for the identification of novel beta-cyclodextrin guest molecules. *New J. Chem.*, 31, 1941.

Steffen, A., Kämper, A., & Lengauer, T. (2006). Flexible docking of ligands into synthetic receptors using a two-sided incremental construction algorithm. *J. Chem. Inf. Model.*, 46, 1695.

Steffen, A., Thiele, C., Tietze, S., Strassnig, C., Kämper, A., Lengauer, T., Wenz, G., & Apostolakis, J. (2007b). Improved cyclodextrin based receptors for camptothecin by inverse virtual screening. *Chem. Eur. J.*, 13, 6801.

Stone, A. J. (2000). *Recent Theoretical and Experimental Advances in Hydrogen Bonded Clusters*, chapter Universal models of hydrogen bonding., (pp. 25–34). Kluwer, Boston.

Stouten, P. F. W., Frömmel, C., Nakamura, H., & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. simul.*, 10, 97.

Stuerzebecher, C.-S., Witt, W., Raduechel, B., Skuballa, W., & Vorbrueggen, H. (1996). Prostacyclins, their analogs or prostaglandins and thromboxane antagonists for treatment of thrombotic and thromboembolic syndromes. US 5523321.

Sussman, J. L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L., & Silman, I. (1991). Atomic structure of acetylcholinesterase from torpedo californica: a prototypic acetylcholine-binding protein. *Science*, 253, 872.

Suzuki, T. (2001). A nonlinear group contribution method for predicting the free energies of inclusion complexation of organic molecules with alpha- and beta-cyclodextrins. *J. Chem. Comput. Sci.*, 41, 1266.

Suzuki, T., Ishida, M., & Fabian, W. M. (2000). Classical QSAR and comparative molecular field analyses of the host-guest interaction of organic molecules with cyclodextrins. *J. Comput.-Aided Mol. Des.*, 14, 669.

Szejtli, J. (1980). Enhancement of stability and biological effect of cholecalciferol by beta-cyclodextrin complexation. *Pharmazie*, 35, 779.

Szejtli, J. (1984). Highly soluble β-cyclodextrin derivatives. *Stärke*, 36, 429.

Takimoto, C. H., Wright, J., & Arbuck, S. G. (1998). Clinical applications of the camptothecins. *Biochim. Biophys. Acta, Gene Struct. Expression*, 1400, 107.

Tanimoto, T. T. (1957). *Internal Report 17th November*. Technical report, IBM.

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V., Radchenko, E., Zefirov, N. S., Makarenko, A. S., Tanchuk, V. Y., & Prokopenko, V. V. (2005). Virtual computational chemistry laboratory - design and description. *J. Comput.-Aid. Mol. Des.*, 19, 453.

Tietze, S. & Apostolakis, J. (2007). GlamDock: Development and validation of a new docking tool on several thousand protein-ligand complexes. *J. Chem. Inf. Model.*, (pp. accepted).

Todeschini, R. & Consonni, V. (2000). *Handbook of Molecular Descriptors*, volume 11 of *Methods and principles in medicinal chemistry*. Weinheim: Wiley-VCH.

Trumpp-Kallmeyer, S., Hoflack, J., Bruinvels, A., & Hibert, M. (1992). Modeling of g-protein-coupled receptors: application to dopamine, adrenaline, serotonin, acetylcholine, and mammalian opsin receptors. *J. Med. Chem.*, 35, 3448.

Ueda, H. & Perrin, J. H. (1986). A microcalorimetric investigation of the binding of flurbiprofen to cyclodextrins. *J. Pharm. Biomed. Anal.*, 4, 107.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.*, 52, 609.

Waldvogel, S. R., Fröhlich, R., & Schalley, C. A. (2000). First artificial receptor for caffeine: A new concept for the complexation of alkylated oxopurines. *Angew. Chem. Int. Ed.*, 39, 2472.

Walters, W. P. & Murcko, M. A. (2002). Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.*, 54, 255.

Wang, R. X., Gao, Y., & Lai, L. H. (2000). LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.*, 6, 498.

Wehrens, R. & Mevik, B.-H. (2007). *PLS: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR).* Technical report, R package version 2.0-1.

Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., & Shoichet, B. K. (2004). Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.*, 337, 1161.

Wei, D. Q., Zhang, R., Du, Q. S., Gao, W. N., Li, Y., Gao, H., Wang, S. Q., Zhang, X., Li, A. X., Sirois, S., & Chou, K. C. (2006). Anti-SARS drug screening by molecular docking. *Amino Acids*, 31, 73.

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

Wennmohs, F., Staemmler, V., & Schindler, M. (2003). Theoretical investigation of weak hydrogen bonds to sulfur. *J. Chem. Phys.*, 119, 3208.

Wenz, G. (1994). Cyclodextrins as building-blocks for supramolecular structures and functional units. *Angew. Chem. Int. Ed.*, 33, 803.

Wenz, G., Gruber, C., Keller, B., Schilli, C., Albuzat, T., & Müller, A. (2006a). Kinetics of threading alpha-cyclodextrin onto cationic and zwitterionic poly(bola-amphiphiles). *Macromolecules*, 39, 8021.

Wenz, G., Han, B. H., & Müller, A. (2006b). Cyclodextrin rotaxanes and polyrotaxanes. *Chem. Rev.*, 106, 782.

Yang, G.-F. & Huang, X. (2006). Development of quantitative structure-activity relationships and its application in rational drug design. *Curr. Pharm. Des.*, 12, 4601.

Zabel, V., Saenger, W., & Mason, S. A. (1986). Topography of cyclodextrin inclusion complexes.23. neutron-diffraction study of the hydrogen-bonding in beta-cyclodextrin undecahydrate at 120-k - from dynamic flip-flops to static homodromic chains. *J. Am. Chem. Soc.*, 108, 3664.

Zadmard, R. & Schrader, T. (2005). Nanomolar protein sensing with embedded receptor molecules. *J. Am. Chem. Soc.*, 127, 904.

A

Appendix A

A.1 FlexX

Our tool FlexR is based on the protein-ligand docking tool FlexX (Rarey et al., 1996a). The main data structures, algorithms and the underlying chemical modeling have been taken over for our work. The algorithms used for the structure prediction of protein-ligand complexes are summarized in the following sections.

A.1.1 Fragmentation and Base Fragment Selection

First, FlexX severs the ligand by cutting at each acyclic single bond. All obtained acyclic fragments are treated as rigid. For cyclic fragments with up to ten atoms multiple conformations are considered. These are generated automatically by Corina (Sadowski & Gasteiger, 1993). From these fragments FlexX selects so-called base fragments. Base fragments are single fragments or combinations of connected single fragments with preferably many directional interactions and a small number of discrete conformations. None of the base fragments is entirely contained in another base fragment.

A.1.2 Base Placement

The complex construction starts with the placement of the base fragments (Rarey et al., 1996b, 1997). For this step, FlexX employs two algorithms. The first, is called triangle matching. Here, FlexX superimposes triangles built on interaction centers of the ligand onto compatible triangles, which are derived from the interaction surfaces within the protein binding site (see Figure A.1, a). Two triangles are compatible if the corresponding interactions match, i. e. for example a hydrogen bond donor interacts with a hydrogen bond acceptor. The calculated transformation of the triangle is then applied for the entire base fragment. In the case, the base fragment has less than three interaction centers, or if too few

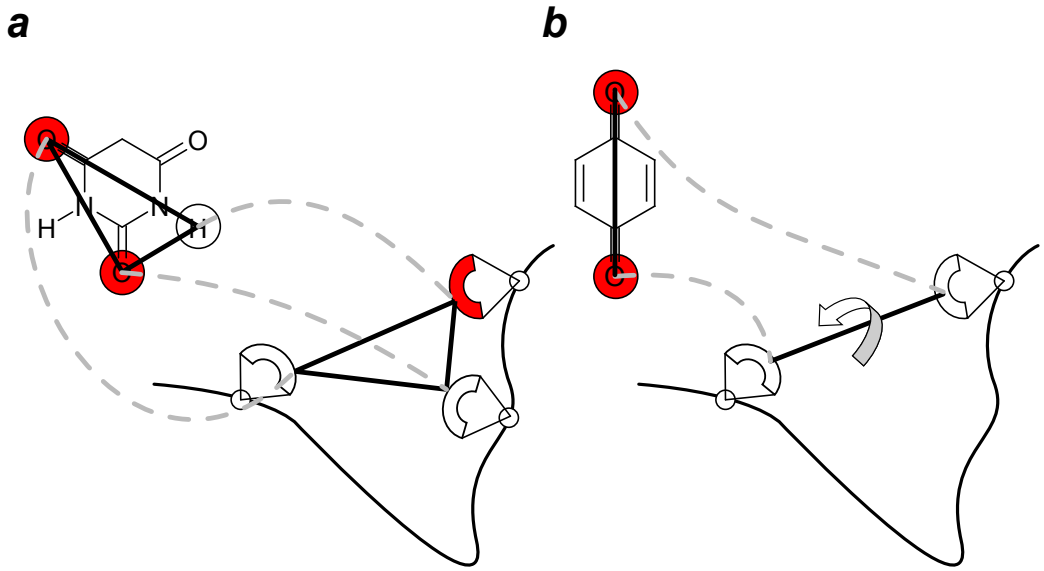


Figure A.1. Placement of the base fragments in FLEXX. a) The triangle matching algorithm matches triangles on the interaction centers of the base fragment with compatible triangles on the interaction surfaces of the binding site. b) The line matching algorithm matches pairs of interaction centers of the base fragment with compatible pairs of interaction points within the binding site. Due to geometric ambiguities the base fragment is rotated around the line axis.

placements were obtained with triangle matching, FLEXX uses a line matching algorithm. In line matching, FLEXX superimposes pairs of interaction centers of the base fragment with compatible pairs of interaction points within the protein binding site (see Figure A.1, b). Due to geometric ambiguities, the base fragment is rotated around the axis between the pair. Both algorithms employ a hash table to find compatible triangles or respectively pairs in an efficient manner. In order to reduce the number of base placements, FLEXX discards placements with steric clashes. Furthermore, the placements are geometrically clustered. In this way, FLEXX avoids the generation of too similar placements.

A.1.3 Incremental Complex Construction

After the base placement, FLEXX proceeds to the incremental complex construction phase (see Figure A.2). Here, all remaining fragments are consecutively placed in a precomputed order. Each fragment is added in a discrete number of low-energy torsion angles. The torsion angles are taken from the MIMUMBA database (Klebe & Mietzner, 1994). The FLEXX-scoring function scores each generated (partial) conformation of the ligand after a fragment was added. (Partial) conformations exhibiting steric clashes are discarded. After each round only the k best-scoring solutions submitted to a clustering procedure. The resulting

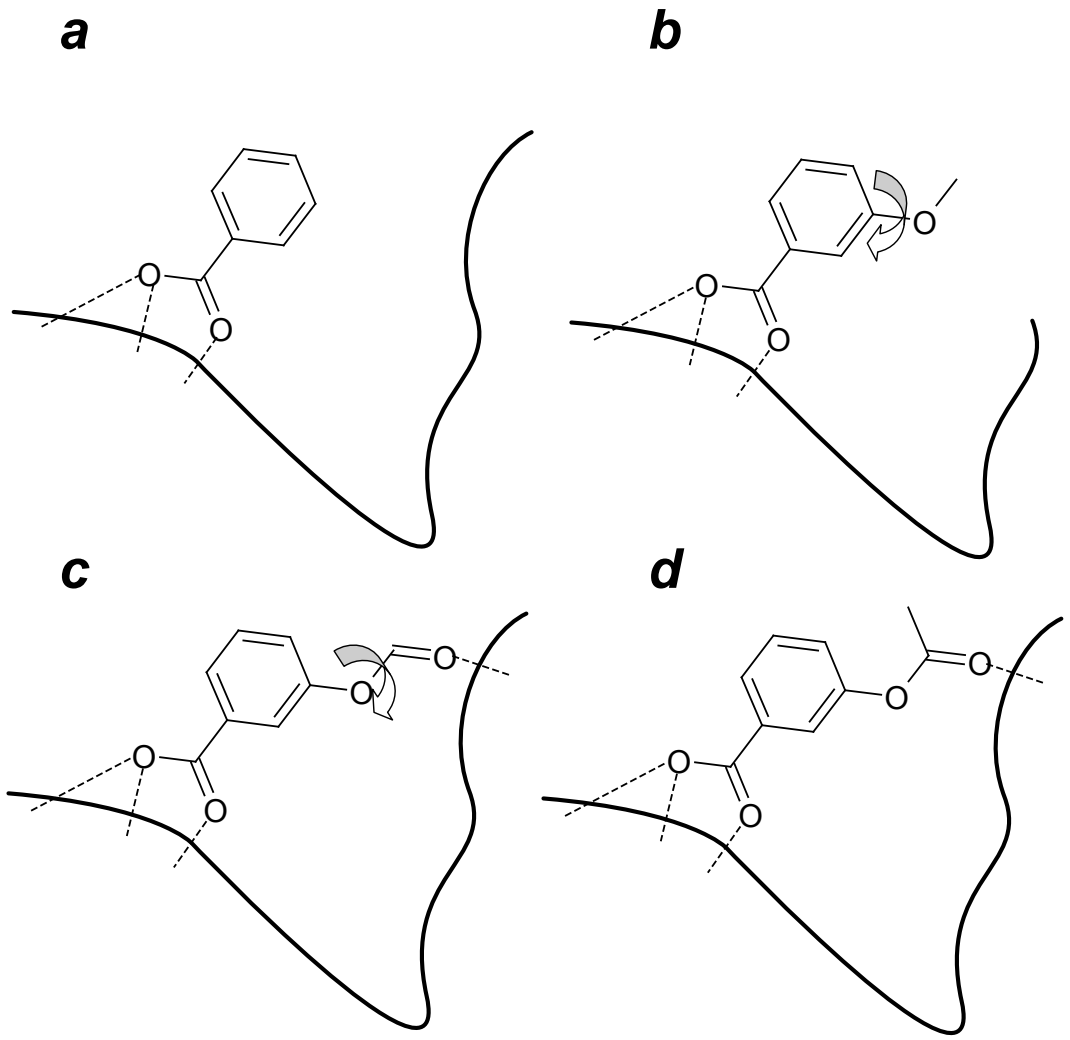


Figure A.2. The incremental complex construction. Starting from a given base fragment placement, FlexX incrementally constructs the ligand by consecutively adding the remaining fragments.

(partial) solutions enter into the next expansion step. This is repeated until the ligand is entirely built up. All generated complete conformations of the ligand are presented as the solution set.

B

Appendix B

The following table shows the selected descriptors of the QSPR model used in Chapter 4. The weights of the descriptors in the model as well as the selection order of the descriptors are given.

Table B.1: Shown are the selected descriptors of the applied QSPR model used in Chapter 4, their weights in the model and the selection order.

ID	E-Dragon-ID	weight	selection order
1	ATS2p	-4.865	1
2	nCb-	3.448	10
3	Ms	2.975	5
4	nArOH	-2.968	4
5	BLTD48	2.663	34
6	R3u+	2.533	6
7	EEig11r	2.472	7
8	MLOGP2	-2.387	9
9	L2s	2.383	2
10	RDF100m	2.351	12
11	BELe6	-2.148	26

. . . continued on next page

Table B.1 ... continued from previous page

ID	E-Dragon-ID	weight	selection order
12	EEig06r	1.883	27
13	RTu	-1.799	39
14	Mor22u	-1.794	3
15	Mor03m	1.558	20
16	EEig10d	1.548	16
17	G3e	-1.483	41
18	EEig09x	1.442	28
19	SPI	-1.408	36
20	piPC07	-1.328	47
21	RDF050u	1.305	38
22	G2	-1.267	30
23	RDF050e	1.265	33
24	piPC08	1.22	44
25	X2	-1.211	50
26	Mor17p	1.182	8
27	S-107	-1.161	14
28	GGI5	-1.157	13
29	G3v	1.125	46
30	ATS6p	-1.059	54
31	BIC1	0.966	23
32	Mor19p	0.958	18
33	DISPe	-0.941	11

... continued on next page

Table B.1 . . . continued from previous page

ID	E-Dragon-ID	weight	selection order
34	C-031	0.909	35
35	Mor21v	0.888	22
36	R7u	-0.861	56
37	H5p	-0.842	37
38	IDET	0.835	66
39	nArNR2	0.827	43
40	nRCOOR	-0.774	32
41	H-049	0.762	15
42	EEig14r	0.724	49
43	MATS5v	0.705	19
44	HATS8p	0.69	55
45	C-037	0.674	45
46	Mor12m	-0.668	17
47	C-015	-0.663	21
48	Depressant-50	-0.614	52
49	Lop	0.613	42
50	T(O..F)	-0.557	29
51	HATS6u	0.552	51
52	SIC1	-0.491	24
53	IAC	-0.475	65
54	nPyrazoles	0.441	57
55	MAXDN	-0.436	31

. . . continued on next page

Table B.1 ... continued from previous page

ID	E-Dragon-ID	weight	selection order
56	ATS7m	-0.373	60
57	ATS7p	0.301	67
58	D/Dr08	0.297	53
59	BLI	-0.285	48
60	R6e+	-0.279	59
61	H7v	0.274	62
62	X1Av	-0.256	64
63	ATS8v	-0.252	63
64	ISH	-0.224	25
65	C-022	0.224	61
66	RTe+	-0.188	58
67	nPyrazines	-0.132	40
68	nArCOOR	-0.069	68

C

Appendix C

Table C.1 shows the assembled dataset for the generation of the QSPR models in Chapter 5. For each molecule the three thermodynamic parameters ΔG°, ΔH°, and $T\Delta S^\circ$ for the complex formation with β-cyclodextrin are listed. The data for each of the molecules was taken from (Rekharsky & Inoue, 1998).

Table C.1: The data used for the generation of the QSPR models in Chapter 5. All experimental values were taken from Rekharsky & Inoue (1998).

Molecule	ΔG° [kJ mol^{-1}]	ΔH° [kJ mol^{-1}]	$T\Delta S^\circ$ [kJ mol^{-1}]
(+-)-cis-2-methylcyclohexanol	-17.08	-9.9	7.2
(+-)-norphenylephrine	-8.65	-20.7	-12
(+-)-octopamine	-9.4	-15.86	-6.5
(+-)-trans-2-methylcyclohexanol	-16.38	-8.66	7.72
(-)-anisodamine	-13.3	-17.6	-4.3
(-)-anisodine	-10.6	-11.6	-1
(-)-atropine	-14.6	-19.5	-4.9
(-)-scopolamine	-12.9	-17.9	-5
(1-methylhexyl)ammonium	-10.7	2	12.7

. . . continued on next page

Table C.1 ... continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
(1R,2R)-(-)-pseudoephedrine	-10.49	-9.99	0.5
(1S,2R)-(+)-ephedrine	-9.95	-8.79	1.2
(2,5-dimethoxyphenethyl)-ammonium	-9.08	-9.39	-0.3
(2-methoxyphenethyl)ammonium	-5.15	-13.5	-8.3
(3,4-dihydroxyphenethyl)-ammonium	-8.58	-16.52	-7.9
(3,4-dimethoxyphenethyl)-ammonium	-6.52	-2	4.5
(3-methoxyphenethyl)ammonium	-10.39	-13.32	-2.9
(3-methylphenyl)acetate	-6.1	-11.5	-5.4
(3-phenylpropyl)ammonium	-11.29	-9.44	1.8
(4-hydroxyphenethyl)ammonium	-10.58	-13.8	-3.2
(4-methoxyphenethyl)ammonium	-10.78	-8.21	2.6
(4-methylphenethyl)ammonium	-11.11	-6.84	4.3
(4-methylphenyl)acetate	-9.17	-12.1	-2.9
(R)-(-)-2-butanol	-6.4	4.9	11.3
(R)-(-)-2-hexanol	-11.8	1.9	13.7
(R)-(-)-phenylephrine	-9.1	-21.9	-12.8
(S)-(+)-2-pentanol	-8.6	4.1	12.8
1-adamantaneacetate	-28.7	-33.1	-4.4
1-adamantaneammonium	-22.4	-22.1	0.3
1-adamantanecarboxylate	-25.7	-23.9	1.8
1-adamantylmethylammonium	-25.5	-17.2	8.6
1-adamantyltrimethylammonium	-20.5	-24.5	-4

... continued on next page

Table C.1 ... continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
1-benzylimidazole	-14.92	-15.9	-1
1-bicyclo[2.2.1]hept-2-enecarboxylate	-15.7	-7.5	8.1
1-bicyclo[2.2.1]heptanecarboxylate	-16.7	-8	8.8
1-bicyclo[2.2.2]octanecarboxylate	-21.9	-15.9	5.9
1-butanol	-6.9	4.3	11.2
1-butylimidazole	-12.5	-10.7	1.8
1-hexanol	-13.3	0.4	13.7
1-methylcyclohexanol	-17.47	-9.6	7.9
1-naphthaleneacetate	-24.8	-4.6	20.2
1-naphthalenesulfonate	-19.4	-6.2	13.2
1-pentanol	-10.3	4.6	14.9
1-phenylimidazole	-8	-39	-31
1-propanol	-3.7	6	9.8
2,2-dimethyl-1-propanol	-15.5	-8.8	6.3
2,3,6-naphthalenetrisulfonate	-12.7	-12.9	-0.3
2,6-naphthalenedisulfonate	-18.8	-11.7	7.1
2,7-naphthalenedisulfonate	-13.9	-28.2	-14.3
2-(4-aminophenyl)-ethyl-ammonium	-8.54	-8.7	-0.2
2-chlorophenol	-13.1	-19	-6
2-methylcyclohexanone	-15.7	-13.7	2.1
2-norbornaneacetate	-20.8	-10.7	10.2
2-propanol	-2.4	11.1	13.4

... continued on next page

Table C.1 . . . continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
3-(2-hydroxyphenyl)propionate	-10.89	-15.1	-4.2
3-(4-hydroxyphenyl)propionate	-14.11	-14.2	-0.1
3-O-methyldopamine	-3.6	-13.4	-9.8
3-chlorophenol	-13.1	-19	-5
3-methoxyphenylacetate	-9.02	-12.3	-3.2
3-methylcyclohexanol	-16.66	-8.74	7.93
3-nitrophenol	-13.9	-12.1	1.8
3-noradamantanecarboxylate	-21.1	-15.7	5.4
3-phenylbutanoate	-14.72	-9.41	5.3
4-O-methyldopamine	-9.78	-15.3	-5.5
4-amino-1-naphthalenesulfonate	-9.7	-10	0.3
4-benzylpiperidine	-18.83	-13.8	5.1
4-bromophenol	-16.7	-12.2	4.5
4-chlorophenol	-14.9	-11.9	3
4-hydroxycoumarin	-13.1	-12	1.1
4-iodophenol	-17	-16.1	0.9
4-methoxyphenylacetate	-10.51	-8.22	2.3
4-methylphenol	-13.7	-12.5	1.2
4-nitrophenol	-13.8	-13.4	0.3
4-phenylbutanoate	-15.06	-11.78	3.3
5-methylresorcinol	-9.8	-21.2	-11.2
6-[(4-tert-butylphenyl)-amino]-2-	-27.1	-25.3	1.8

. . . continued on next page

Table C.1 ... continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
naphthalenesulfonate			
D-glucose	1.3	1.6	0.3
L-alpha-O-benzylglycerol	-12.03	-9.2	2.8
L-phenylalanine	-7.2	-9	-1.8
L-phenylalanineamide	-7.7	-9	-1.3
L-tryptophan	-13.3	-0.8	12.5
L-tyrosine	-8.7	-3.8	4.9
N-methylphenethylammonium	-7.59	-7.3	0.3
acenocoumarin	-14.7	-15.5	-0.7
adiphenine	-19.6	-31.9	-12.3
amobarbital	-17.7	-17.8	-0.1
aspartame	-12	-11.7	0.3
barbital	-13.9	-11.5	2.4
benzene	-11.6	-3.5	8.1
benzoate	-6.86	-10.5	-3.6
benzylalcohol	-7.7	-13.8	-6.2
bromodiphenhydramine	-19	-25.4	-6.4
butabarbital	-16.9	-33.2	-16.4
butethal	-16.8	-9.8	7
butylbarbituricacid	-14.8	-15.8	-1.0
butylthiobarbituricacid	-16.4	-20.4	-3.9
chlorcyclizine	-19.4	-22.8	-3.4

... continued on next page

Table C.1 . . . continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
chlorpromazine	-22.4	-26.8	-4.5
cinnarizine	-19.8	-17.3	2.4
cis-1,2-cyclohexanediol	-13.9	-9.8	4.2
cis-4-methylcyclohexanol	-18.07	-9.5	8.55
cyclizine	-17.6	-28.8	-11.2
cyclobarbital	-17.9	-20.2	-4.8
cyclobutanol	-6.5	3.7	10.2
cycloheptanol	-19.08	-12.37	6.7
cyclohexanol	-16.2	-4.9	11.1
cyclohexanone	-15.5	-11.7	3.9
cyclooctanol	-20.8	-16.4	4.4
cyclopentanol	-12.76	-4.56	8.2
di-2-(1-adamantyl)ethylhydrogen-phosphate	-30.5	-29.3	1.3
dicumarol	-20.5	-40.4	-20
diphenhydramine	-17.5	-29.4	-11.9
diphenidol	-17	-33.7	-16.7
diphenylpyraline	-19.1	-27.8	-8.7
ethylthiobarbituricacid	-14.0	-15.5	-0.3
flufenamicacid	-18.1	-11.4	6.7
flurbiprofen	-18.8	-23.3	-4.5
heptanoate	-14.2	1.8	15.9

. . . continued on next page

Table C.1 ... continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
heptylbarbituricacid	-19.8	-32.0	-12.2
hexanoate	-9.54	5.5	15
hexobarbital	-16.4	-24.5	-8.2
hexylammonium	-10.4	2.5	12.9
hexylthiobarbituricacid	-20.0	-29.6	-10.2
hydroquinone	-11.7	-17.1	-5.4
hydroxyzine	-19.2	-24.2	-5
imidazole	-1.6	-16	-14
meclizine	-19.1	-22.6	-3.4
mephobarbital	-18.1	-39.7	-21.6
methapyriline	-14.5	-15.5	-1
methylorange (anion)	-18.8	-15.9	3.1
methylred (anion)	-20.5	-19.6	0.9
niflumicacid	-15.5	-19	-3.5
octylammonium	-15	-2	13
orphenadrine	-17.7	-31.3	-13.5
pentanoate	-5.3	8	13
pentobarbital	-18	-23.2	-5.2
pentylthiobarbituricacid	-19.2	-23.6	-6.7
phenethylammonium	-7.43	-6.9	0.6
phenobarbital	-18.3	-31.2	-12.9
phenol	-11.3	-12.2	-1.2

... continued on next page

Table C.1 . . . continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
phenprocoumon	-16.2	-13.6	2.6
phenylacetate	-7.1	-7.5	-0.4
phenylpropionate	-12.27	-7.6	4.7
piroxicam	-11.2	-10.5	0.7
proadifen	-16.9	-29.5	-12.6
propylbarbituricacid	-12.8	-11.6	1.3
propylthiobarbituricacid	-14.1	-16.4	-2.2
prostaglandin E2	-18.7	-19.3	-0.6
resorcinol	-11.6	-18.2	-6.5
secobarbital	-18.6	-25.4	-6.7
sulfadiazine	-15.7	-24	-8.3
sulfadimethoxine	-15.92	-19.1	-3.2
sulfaethidole	-18	-14.6	3.5
sulfaisoxazole	-15.95	-28	-12
sulfamerazine	-14.44	-16.5	-2.1
sulfamethizole	-17.7	-27.6	-9.9
sulfamethoxazole	-15.81	-22.5	-6.7
sulfapyridine	-15.2	-33.9	-18.7
sulfathiazole	-19.24	-29.8	-10.5
sulfathidole	-18.1	-35.4	-17.3
sulfoisomidine	-12.84	-15.6	-2.8
terfenadine	-24.4	-20	4.4

. . . continued on next page

Table C.1 ... continued from previous page

Molecule	ΔG°	ΔH°	$T\Delta S^\circ$
thenyldiamine	-13.2	-14.8	-1.6
thiopental	-19.6	-25.7	-6.3
thiophenobarbital	-20.5	-34.4	-14.0
trans-1,2-cyclohexanediol	-11.4	-4.4	7.2
trans-2-hydroxycinnamic acid	-14.7	-23.0	-8.9
trans-2-methylcinnamic acid	-14.9	-12.1	3.5
trans-3-hydroxycinnamate	-13.5	-21.3	-8.0
trans-3-methylcinnamate	-14.4	-18.8	-4.6
trans-4-hydroxycinnamate	-14.9	-20.5	-5.6
trans-4-methylcinnamate	-15.2	-17.6	-1.9
trans-4-methylcyclohexanol	-19	-9.1	9.9
triprolidine	-13.8	-13.8	0
tropicamide	-15	-25	-10
valerolactam	-8	-3.7	4.2
warfarin	-16.3	-11.6	4.8

In order to investigate the influence of structural changes to thermodynamical parameters we clustered the β-cyclodextrin guest molecules presented in Rekharsky & Inoue (1998) and used in Chapter 5 according to molecular similarity as calculated by FUZZEE. Clusters were built using a similarity threshold of 0.7 with a complete linkage algorithm. In this way all structures within a cluster have a similarity of 0.7 or higher and thus are structurally closely related compounds. In Table C.2 the clusters as well as the mean values for ΔG°, ΔH° and $T\Delta S^\circ$ together with their standard deviations for all molecules of the cluster are given.

Table C.2: The generated structural clusters are shown as used in the discussion of Section 5.3. Only clusters containing multiple data points are shown. The experimental data is taken from reference Rekharsky & Inoue (1998).

Cluster-ID 1

Molecules		
	(+-)-cis-2-methylcyclohexanol	(+-)-trans-2-methylcyclohexanol
	(R)-(-)-2-butanol	(R)-(-)-2-butanol
	(R)-(-)-2-hexanol	(S)-(+)-2-pentanol
	1-hexanol	1-methylcyclohexanol
	1-pentanol	1-propanol
	2,2-dimethyl-1-propanol	2-propanol
	3-methylcyclohexanol	cis-4-methylcyclohexanol
	cyclobutanol	cycloheptanol
	cyclohexanol	cyclooctanol
	cyclopentanol	trans-4-methylcyclohexanol

$\varnothing(\Delta G^\circ) = -14.81 \pm 5.55 \quad \varnothing(\Delta H^\circ) = -3.26 \pm 6.81 \quad \varnothing(T\Delta S^\circ) = 11.55 \pm 5.23$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 2

Molecules (+-)-norphenylephrine (1R,2R)-(-)-pseudoephedrine
(1S,2R)-(+)-ephedrine (R)-(-)-phenylephrine

$\varnothing(\Delta G^\circ) = -9.62 \pm 0.74$ $\varnothing(\Delta H^\circ) = -15.39 \pm 5.95$ $\varnothing(T\Delta S^\circ) = -5.74 \pm 6.62$

Cluster-ID 3

Molecules (+-)-octopamine (3,4-dihydroxyphenethyl)-ammonium
3-O-methyldopamine

$\varnothing(\Delta G^\circ) = -7.63 \pm 2.4$ $\varnothing(\Delta H^\circ) = -13.91 \pm 4.42$ $\varnothing(T\Delta S^\circ) = -6.25 \pm 3.92$

Cluster-ID 4

Molecules (-)-anisodamine (-)-anisodine
(-)-atropine (-)-scopolamine

$\varnothing(\Delta G^\circ) = -12.92 \pm 1.45$ $\varnothing(\Delta H^\circ) = -16.92 \pm 3.06$ $\varnothing(T\Delta S^\circ) = -4.0 \pm 1.7$

Cluster-ID 5

Molecules (1-methylhexyl)ammonium hexylammonium
octylammonium

$\varnothing(\Delta G^\circ) = -12.03 \pm 2.57$ $\varnothing(\Delta H^\circ) = 0.83 \pm 2.47$ $\varnothing(T\Delta S^\circ) = 12.87 \pm 0.15$

Cluster-ID 6

Molecules (2,5-dimethoxyphenethyl)-ammonium
2-methoxyphenethyl)ammonium
(3-methoxyphenethyl)ammonium

$\varnothing(\Delta G^\circ) = -8.86 \pm 2.22$ $\varnothing(\Delta H^\circ) = -11.62 \pm 1.83$ $\varnothing(T\Delta S^\circ) = -2.76 \pm 3.24$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 7

Molecules (3,4-dimethoxyphenethyl)-ammonium

4-O-methyldopamine

$\varnothing(\Delta G^\circ) = -9.0 \pm 2.2 \quad \varnothing(\Delta H^\circ) = -10.24 \pm 7.2 \quad \varnothing(T\Delta S^\circ) = -1.4 \pm 5.24$

Cluster-ID 8

Molecules	(3-methylphenyl)acetate	(4-methylphenyl)acetate
	1-naphthaleneacetate	3-phenylbutanoate
	4-phenylbutanoate	phenylacetate
	phenylpropionate	trans-2-methylcinnamicacid
	trans-3-methylcinnamate	trans-4-methylcinnamate

$\varnothing(\Delta G^\circ) = -13.41 \pm 4.82 \quad \varnothing(\Delta H^\circ) = -10.79 \pm 4.17 \quad \varnothing(T\Delta S^\circ) = 2.71 \pm 6.83$

Cluster-ID 9

Molecules	(3-phenylpropyl)ammonium	(4-methylphenethyl)ammonium
	N-methylphenethylammonium	phenethylammonium

$\varnothing(\Delta G^\circ) = -9.49 \pm 1.84 \quad \varnothing(\Delta H^\circ) = -7.45 \pm 1.21 \quad \varnothing(T\Delta S^\circ) = 2.05 \pm 1.35$

Cluster-ID 10

Molecules (4-hydroxyphenethyl)ammonium (4-methoxyphenethyl)ammonium

$\varnothing(\Delta G^\circ) = -10.68 \pm 0.36 \quad \varnothing(\Delta H^\circ) = -11.54 \pm 3.1 \quad \varnothing(T\Delta S^\circ) = -0.86 \pm 3.44$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 11

Molecules 1-adamantaneacetate 1-adamantanecarboxylate
1-bicyclo[2.2.2]octanecarboxylate 2-norbornaneacetate
3-noradamantanecarboxylate heptanoate
hexanoate pentanoate
1-bicyclo[2.2.1]hept-2-enecarboxylate
1-bicyclo[2.2.1]heptanecarboxylate

$\varnothing(\Delta G^\circ) = -20.99 \pm 5.91$ $\varnothing(\Delta H^\circ) = -16.87 \pm 11.36$ $\varnothing(T\Delta S^\circ) = 4.1 \pm 6.52$

Cluster-ID 12

Molecules 1-adamantaneammonium 1-adamantylmethylammonium

$\varnothing(\Delta G^\circ) = -23.95 \pm 2.19$ $\varnothing(\Delta H^\circ) = -19.65 \pm 3.46$ $\varnothing(T\Delta S^\circ) = 4.45 \pm 5.87$

Cluster-ID 13

Molecules 1-adamantyltrimethylammonium

$\varnothing(\Delta G^\circ) = -19.0 \pm 2.12$ $\varnothing(\Delta H^\circ) = -21.6 \pm 4.1$ $\varnothing(T\Delta S^\circ) = -2.6 \pm 1.98$

Cluster-ID 14

Molecules 1-benzylimidazole 1-phenylimidazole

$\varnothing(\Delta G^\circ) = -11.46 \pm 4.89$ $\varnothing(\Delta H^\circ) = -27.45 \pm 16.33$ $\varnothing(T\Delta S^\circ) = -16.0 \pm 21.21$

Cluster-ID 15

Molecules 1-naphthalenesulfonate 4-amino-1-naphthalenesulfonate

$\varnothing(\Delta G^\circ) = -14.55 \pm 6.86$ $\varnothing(\Delta H^\circ) = -8.1 \pm 2.69$ $\varnothing(T\Delta S^\circ) = 6.75 \pm 9.12$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 16

Molecules 2,3,6-naphthalenetrisulfonate 2,6-naphthalenedisulfonate

$\varnothing(\Delta G^\circ) = -15.75 \pm 4.31$ $\varnothing(\Delta H^\circ) = -12.3 \pm 0.85$ $\varnothing(T\Delta S^\circ) = 3.4 \pm 5.23$

Cluster-ID 17

Molecules 2-chlorophenol 3-chlorophenol

4-bromophenol 4-chlorophenol

4-iodophenol 4-methylphenol

benzylalcohol phenol

$\varnothing(\Delta G^\circ) = -14.02 \pm 3.18$ $\varnothing(\Delta H^\circ) = -14.37 \pm 2.97$ $\varnothing(T\Delta S^\circ) = -0.18 \pm 4.91$

Cluster-ID 18

Molecules 2-methylcyclohexanone cyclohexanone

$\varnothing(\Delta G^\circ) = -15.6 \pm 0.14$ $\varnothing(\Delta H^\circ) = -12.7 \pm 1.41$ $\varnothing(T\Delta S^\circ) = 3.0 \pm 1.27$

Cluster-ID 19

Molecules 3-(2-hydroxyphenyl)propionate 4-nitrophenol

trans-2-hydroxycinnamicacid

$\varnothing(\Delta G^\circ) = -14.17 \pm 1.79$ $\varnothing(\Delta H^\circ) = -18.9 \pm 11.78$ $\varnothing(T\Delta S^\circ) = -4.7 \pm 10.32$

Cluster-ID 20

Molecules 3-(4-hydroxyphenyl)propionate L-tyrosine (protonated)

trans-4-hydroxycinnamate

$\varnothing(\Delta G^\circ) = -12.44 \pm 2.77$ $\varnothing(\Delta H^\circ) = -11.3 \pm 7.54$ $\varnothing(T\Delta S^\circ) = 1.23 \pm 5.22$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 21

Molecules 3-methoxyphenylacetate 4-methoxyphenylacetate

$\varnothing(\Delta G^\circ) = -9.77 \pm 1.05 \quad \varnothing(\Delta H^\circ) = -10.26 \pm 2.88 \quad \varnothing(T\Delta S^\circ) = -0.45 \pm 3.89$

Cluster-ID 22

Molecules 3-nitrophenol hydroquinone

$\varnothing(\Delta G^\circ) = -12.8 \pm 1.56 \quad \varnothing(\Delta H^\circ) = -14.6 \pm 3.54 \quad \varnothing(T\Delta S^\circ) = -1.8 \pm 5.09$

Cluster-ID 23

Molecules 5-methylresorcinol resorcinol

$\varnothing(\Delta G^\circ) = -10.93 \pm 0.99 \quad \varnothing(\Delta H^\circ) = -20.07 \pm 1.63 \quad \varnothing(T\Delta S^\circ) = -9.07 \pm 2.38$

Cluster-ID 24

Molecules L-phenylalanine L-phenylalanineamide

$\varnothing(\Delta G^\circ) = -8.22 \pm 3.58 \quad \varnothing(\Delta H^\circ) = -9.42 \pm 2.78 \quad \varnothing(T\Delta S^\circ) = -1.2 \pm 5.47$

Cluster-ID 25

Molecules acenocoumarin warfarin

$\varnothing(\Delta G^\circ) = -15.5 \pm 1.13 \quad \varnothing(\Delta H^\circ) = -13.55 \pm 2.76 \quad \varnothing(T\Delta S^\circ) = 2.05 \pm 3.89$

Cluster-ID 26

Molecules adiphenine proadifen

$\varnothing(\Delta G^\circ) = -18.25 \pm 1.91 \quad \varnothing(\Delta H^\circ) = -30.7 \pm 1.7 \quad \varnothing(T\Delta S^\circ) = -12.45 \pm 0.21$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 27

Molecules	amobarbital	barbital
	butabarbital	butethal
	butylbarbituricacid	cyclobarbital
	heptylbarbituricacid	hexobarbital
	mephobarbital	pentobarbital
	phenobarbital	propylbarbituricacid
	secobarbital	

$\varnothing(\Delta G^\circ) = -17.11 \pm 1.78 \quad \varnothing(\Delta H^\circ) = -24.07 \pm 9.57 \quad \varnothing(T\Delta S^\circ) = -7.05 \pm 8.46$

Cluster-ID 28

Molecules benzoate

$\varnothing(\Delta G^\circ) = -6.33 \pm 0.75 \quad \varnothing(\Delta H^\circ) = -12.75 \pm 3.18 \quad \varnothing(T\Delta S^\circ) = -6.2 \pm 3.68$

Cluster-ID 29

Molecules	bromodiphenhydramine	diphenhydramine
	diphenylpyraline	orphenadrine

$\varnothing(\Delta G^\circ) = -18.33 \pm 0.84 \quad \varnothing(\Delta H^\circ) = -28.48 \pm 2.5 \quad \varnothing(T\Delta S^\circ) = -10.13 \pm 3.19$

Cluster-ID 30

Molecules	butylthiobarbituricacid	ethylthiobarbituricacid
	hexylthiobarbituricacid	pentylthiobarbituricacid
	propylthiobarbituricacid	thiopental
	thiophenobarbital	

$\varnothing(\Delta G^\circ) = -17.69 \pm 2.81 \quad \varnothing(\Delta H^\circ) = -23.66 \pm 6.89 \quad \varnothing(T\Delta S^\circ) = -6.23 \pm 4.71$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 31

Molecules chlorcyclizine cinnarizine

cyclizine meclizine

$\varnothing(\Delta G^\circ) = -19.06 \pm 0.85$ $\varnothing(\Delta H^\circ) = -23.12 \pm 4.11$ $\varnothing(T\Delta S^\circ) = -4.06 \pm 4.85$

Cluster-ID 32

Molecules cis-1,2-cyclohexanediol trans-1,2-cyclohexanediol

$\varnothing(\Delta G^\circ) = -12.65 \pm 1.77$ $\varnothing(\Delta H^\circ) = -7.1 \pm 3.82$ $\varnothing(T\Delta S^\circ) = 5.7 \pm 2.12$

Cluster-ID 33

Molecules di-2-(1-adamantyl)ethylhydrogenphosphate

$\varnothing(\Delta G^\circ) = -25.65 \pm 6.86$ $\varnothing(\Delta H^\circ) = -22.7 \pm 9.33$ $\varnothing(T\Delta S^\circ) = 3.0 \pm 2.4$

Cluster-ID 34

Molecules flufenamicacid niflumicacid

$\varnothing(\Delta G^\circ) = -16.8 \pm 1.84$ $\varnothing(\Delta H^\circ) = -15.2 \pm 5.37$ $\varnothing(T\Delta S^\circ) = 1.6 \pm 7.21$

Cluster-ID 35

Molecules flurbiprofen

$\varnothing(\Delta G^\circ) = -19.8 \pm 1.41$ $\varnothing(\Delta H^\circ) = -19.1 \pm 5.94$ $\varnothing(T\Delta S^\circ) = 0.7 \pm 7.35$

Cluster-ID 36

Molecules imidazole

$\varnothing(\Delta G^\circ) = -2.0 \pm 0.57$ $\varnothing(\Delta H^\circ) = -15.5 \pm 0.71$ $\varnothing(T\Delta S^\circ) = -13.5 \pm 0.71$

. . . continued on next page

Table C.2 . . . continued from previous page

Cluster-ID 37

Molecules methapyriline, thenyldiamine, triprolidine

$\varnothing(\Delta G^\circ) = -13.83 \pm 0.65$ $\varnothing(\Delta H^\circ) = -14.7 \pm 0.85$ $\varnothing(T\Delta S^\circ) = -0.87 \pm 0.81$

Cluster-ID 38

Molecules methylorange (anion)

$\varnothing(\Delta G^\circ) = -19.8 \pm 1.41$ $\varnothing(\Delta H^\circ) = -17.8 \pm 2.69$ $\varnothing(T\Delta S^\circ) = 2.3 \pm 1.13$

Cluster-ID 39

Molecules sulfadiazine, sulfaisoxazole, sulfamerazine, sulfamethizole, sulfamethoxazole, sulfapyridine, sulfathiazole, sulfoisomidine

$\varnothing(\Delta G^\circ) = -15.98 \pm 1.92$ $\varnothing(\Delta H^\circ) = -22.91 \pm 5.04$ $\varnothing(T\Delta S^\circ) = -6.94 \pm 3.84$

Lebenslauf

Persönliche Daten

Name:	Andreas Steffen
Anschrift:	Svangatan 17B
	41668 Göteborg
	Schweden
Geburtsdatum und -ort:	20. September 1977 in Osnabrück
Familienstand:	ledig

Schulische Ausbildung

1984 – 1988	*Stadtschule*, Wunstorf
1988 – 1990	*Orientierungsstufe Nord*, Wunstorf
1984 – 1988	*Hölty-Gymnasium*, Wunstorf

Studium

1998 – 2003	*Philipps-Universität*, Marburg und
	Université Paris-Sud XI in Paris,
	Fach: Pharmazie
2004 – 2007	Promotionsstudium am
	Max-Planck-Institut für Informatik

Saarbrücken, der 17. Januar 2008

Andreas Steffen