# Theoretical studies of protein-ligand binding

Misini Ignjatovic, Majda

2019

*Document Version:*
Publisher's PDF, also known as Version of record

# Theoretical studies of protein–ligand binding

MAJDA MISINI IGNJATOVIĆ | DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY

# Theoretical Studies of Protein–Ligand Binding

Majda Misini Ignjatović

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Science, Lund University, Sweden.
To be defended on 5th June 2019, at 9:15 in lecture hall B, Centre for Chemistry and Chemical Engineering.

*Faculty opponent*
Holger Gohlke, Heinrich-Heine-University, Duesseldorf

| | |
|---|---|
| Organization<br>LUND UNIVERSITY | Document name<br>DOCTORAL DISSERTATION |
| Centre for Chemistry and Chemical Engineering<br>P.O. Box 124, SE-221 00, Lund, Sweden. | Date of issue<br>2019–06–05 |
| Author(s)<br>Majda Misini Ignjatović | Sponsoring organization |


Title and subtitle
Theoretical studies of protein–ligand binding


Abstract

Understanding how drugs work is of great importance, since it can facilitate drug discovery, both time- and cost-wise. At the same time, it is important to have methods that can help predict how well does a potential drug molecule bind to its target. Computational methods can in many ways contribute to drug design process. In this thesis, we employ different computational approaches to study the binding of various ligands to galectin-3 protein, which is an excellent model system and an interesting therapeutic target. We study the effects of solvation thermodynamics, protein and ligand conformational entropy, as well as specific protein–ligand interactions. Our results indicate that accurate modelling of protein–ligand binding requires careful consideration of solvation and protein–ligand conformational entropy, since they contribute significantly to protein–ligand binding free energies. We also compared different methods used to study the water structure and thermodynamics in the protein–ligand binding site, where we showed that solvent-exposure of the binding site may dictate the choice of the method. Moreover, we participated in the D3R and SAMPL6 blind challenges, where we tested the performance of different methods used to estimate binding affinities. We have showed that the predictions of relative binding affinities improve if displaced water molecules are included in the free-energy perturbation calculations and if the ligand is treated by quantum mechanical methods.

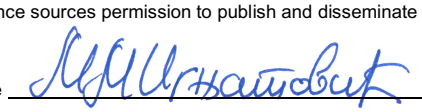


Key words
Protein–ligand binding, MD, GIST, GCMC, FEP, Solvation, Entropy, Water, QM/MM-FEP

Classification system and/or index terms (if any)

| | | |
|---|---|---|
| Supplementary bibliographical information | | Language<br>English |
| ISSN and key title | | ISBN<br>978-91-7422-662-1 |
| Recipient's notes | Number of pages 212 | Price |
| | Security classification | |




Signature  Date 2019-04-25

# Theoretical Studies of Protein–Ligand Binding

Majda Misini Ignjatović


LUND
UNIVERSITY

**Cover photo:**

Boris Ignjatović

**Funding information:**

This work was financially supported by the Knut and Alice Wallenberg foundation.



Faculty of Science
Division of Theoretical Chemistry

ISBN 978-91-7422-662-1 (print)
ISBN 978-91-7422-663-8 (pdf)

Printed in Sweden by Media-Tryck, Lund University
Lund 2019

*To Boris*

# Table of Contents

# List of Publications

I. M. Misini Ignjatović, O. Caldararu, G. Dong, C. Muñoz-Gutierrez, F. Adasme-Carreño, U. Ryde (2016) "Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations", J. Comp.-Aided Mol. Design, 30, 707-730; DO I 10.1007/s10822-016-9942-z.

II. O. Caldararu, M. A. Olsson, M. Misini Ignjatović, M. Wang, U. Ryde (2018) "Binding free energies in the SAMPL6 octa-acid host–guest challenge calculated with MM and QM methods", J. Comput.-Aided Mol. Design, 32, 1027-1046; DOI 10.1007/s10822-018-0158-2.

III. M. Misini Ignjatović, U. Ryde (2019) "Comparison of the GCMC and GIST methods to determine the water structure in protein binding sites", manuscript.

IV. R. Kumar, K. Peterson, M. Misini Ignjatović,[a] H. Leffler, U. Ryde, U. J. Nilsson, D. T. Logan (2019) "Substituted polyfluoroaryl interactions with an arginine side chain in galectin-3 are governed by steric-, desolvation and electronic conjugation effects", Org. Biomol. Chem., 17, 1081-1089; DOI 10.1039/C8OB02888E.

V. R. Kumar, M. Misini Ignjatović,[a] K. Peterson, M. Olsson, H. Leffler, U. Ryde, U. J. Nilsson, D. T. Logan (2019) "Structure and energetics of ligand–fluorine interactions with galectin-3 backbone and side-chain amides – insight into solvation effects and multipolar interactions", submitted to Chem. Sci.

VI. M. L. Verteramo, J. Wallerstein, M. Misini Ignjatović,[a] R. Kumar, V. Chadimová, H. Leffler, F. Zetterberg, D. T. Logan, U. Ryde, M. Akke, U. J. Nilsson (2019) "Structural and thermodynamic studies on halogen-bond interactions in ligand–galectin-3 complexes: electrostatics, solvation and entropy effects", manuscript.

VII. M. L. Verteramo, O. Stenström, M. Misini Ignjatović, O. Caldararu, M. A. Olsson, F. Manzoni, H. Leffler, E. Oksanen, D. T. Logan, U. J. Nilsson, U. Ryde, M. Akke (2018) "Interplay between conformational entropy and solvation entropy in protein ligand binding", J. Am. Chem. Soc., 141, 2012-2026; DOI 10.1021/jacs.8b11099.

VIII. J. Wallerstein, M. Misini Ignjatović,[a] R. Kumar, O. Caldararu, K. Peterson, H. Leffler, D. T. Logan, U. J. Nilsson, U. Ryde, M. Akke (2019) "Entropy–Entropy compensation between the conformational and solvent degrees of freedom finetunes affinity in ligand binding to galectin-3C", manuscript.

[a] Shared first author

# List of papers not included in this thesis

IX. M. Misini Ignjatović, P. Mikulskis, P. Söderhjelm, U. Ryde (2018), "Can MM/GBSA Calculations be Sped up by System Truncation?", J. Comput. Chem., 39, 361-372, DOI: 10.1002/jcc.25120.

X. M. Misini Ignjatović, M. Wang, U. Ryde (2019) "QM/MM free-energy perturbation for ligands binding to proteins", manuscript.

XI. M. Misini Ignjatović, L. Cao, M. A. Olsson, U. Ryde (2019) "Effect of quantum mechanical charges for free-energy perturbation calculations", manuscript.

# List of article contributions

I. I performed the free-energy simulations at the MM level for two out of three sets of ligands and all the GCMC simulations. I participated in the writing of the manuscript.

II. I performed all the MM→QM/MM free-energy simulations. I participated in the writing of the manuscript.

III. I participated in designing the study. I performed all the calculations in the manuscript. I wrote the first draft of the manuscript.

IV. I performed all the QM calculations. I participated in the writing of the manuscript.

V. I performed all the QM calculations. I participated in the writing of the manuscript.

VI. I performed all the QM calculations and MD simulations and GIST analysis. I participated in the writing of the manuscript.

VII. I performed the clustering of the MD trajectories with the unrestrained solute. I performed all subsequent MD simulations with the solute restrained, as well as the GIST post-processing of all trajectories with the solute restrained. I participated in the writing of the manuscript.

VIII. I performed all the MD simulations, conformational entropy calculations and GIST analysis. I participated in the writing of the manuscript.

# Popular science summary

Drug discovery and development is a time-consuming and expensive process. Years, sometimes decades, together with millions of dollars are invested in a compound before it reaches the drug market in the form of an effective and safe-to-use drug. Therefore, it is of great interest to reduce the time and costs needed along the way. The use of computational methods in drug design is one way in which this could be achieved.

To accomplish this, we need a thorough knowledge of how drugs fulfil their purpose. Whether we are talking about simple painkillers that can be found in every medicine cabinet or about drugs used to treat serious illnesses, such as those used in cancer treatment, the mechanism of action of a drug molecule is normally the same: Find the target molecule and inhibit its function. But how does this work? For sure, these target molecules are not fugitives and drugs are not bounty hunters. There are no "Wanted – dead or alive" posters hanging around our blood stream and there is no award waiting at the end of the drug molecule's journey. It is even more interesting than that.

In most cases, the target molecules are proteins. Compared to drug molecules, protein molecules are considered to be molecular giants. Although quite large, they do not have a complicated sequence – they are built out of small molecules called amino acids, connected together into long chains. There are twenty different amino acids that are used by nature to build proteins. We can imagine this as the language of nature, where amino acids are letters of an alphabet. Now, with that analogy, we can explain how it is possible that such, in principle very simple molecules, are so special that they are the first ones to look at when one is trying to cure a disease.

In the same way, a meaningful message is formulated by arranging letters into words and words into sentences: Putting together a number of amino acids in a specific sequence will produce a functional protein, capable of performing a certain function in an organism. But proteins are not just long molecular worms floating around living bodies. The amino acid sequence itself is not enough and one must "read between the lines". In this case, it means to look at the distinctive three-dimensional structure of the protein in question. The shape of

the protein is a direct consequence of the primary protein sequence and it determines what the protein function will be.

Proteins can have many roles in living organisms. They provide structure and support for cells, they bind and carry atoms and small molecules within cells and throughout the body, they act as messengers, transmitting signals between cells, tissues and organs, they speed up chemical reactions (enzymes), and they even protect living organisms from viruses and bacteria.

While performing their functions, proteins may interact with certain molecules, called ligands. A ligand can be anything between an atom and a macromolecule, and what happens with a ligand after it binds to a protein depends on the protein function. However, it is important to note that most proteins bind specific ligands or groups of ligands.

With this knowledge, we can now come back to where we started – drugs. Most drugs work by interacting with specific proteins, so that they either block the physiological function of the protein by disabling the binding of their natural ligands, or cause the protein to become active by mimicking the effect of the natural ligands. This explains the importance of proteins as drug targets. A protein with its particular function is typically part of an important process taking place in a living organism. Sometimes these processes get involved in disease pathways, meaning that, inhibiting or activating these particular targets can potentially eliminate the disease. Moreover, foreign disease-causing agents, such as bacteria, also have their own set of proteins that are crucial for their function, which means that targeting these proteins can eliminate the bacteria or disable their reproduction.

Designing a new drug is always a challenging task. There are protocols and rules to follow when one is trying to find a compound that will be able to bind to a specific target. Even before the preclinical stage of drug discovery, plenty of research is conducted in order to discover the drug target, find and synthetize possible drug candidates, and measure their binding affinities.

The use of computational methods can facilitate drug design in many ways. Simple and inexpensive methods, such as docking and scoring procedures, can enable screening of large compound libraries, which is one way of finding possible drug candidates. Another possibility is *de novo* design, where ligands are designed from scratch. Ideally, the choice of drug candidates should be narrowed down to a single compound, called lead compound, which is further optimized so that it fulfils all the criteria of a successful drug. This requires computational methods that employ higher levels of theory and therefore are

more expensive. There exists a plethora of methods used to estimate binding affinity of a drug molecule binding to a protein target. These methods are continuously being tested and improved. Computational methods can also be used to study the chemistry of protein–drug binding, which helps to better understand the binding process and thereby design better drugs in the future.

In this thesis, various methods were employed to study the binding of different drug candidates to the galectin-3 protein, a target that can potentially be used in cancer treatment. Specific protein–drug interactions were studied, as well as the role of water and conformational entropy of the protein and drug in drug binding. Moreover, we participated in two blind challenges (D3R and SAMPL6), in which we tested the performance of different methods used to estimate binding affinities.

# 1 Introduction

In this chapter, I will briefly describe the drug-design process and introduce basic ideas behind drug-design strategies. Next, I will give an overview of the most important concepts of protein–ligand binding, as proteins are in most cases used as drug targets. Finally, I will present the galectin-3 protein, which was used as the model system in most of the papers in this thesis.

## 1.1 Drug discovery and drug design

Drugs are compounds capable of interacting with a biological system in a way that produces a biological response. The response is not always beneficial for the biological system of interest. In fact, depending on how the system reacts, a compound can be anything between a medicine and a poison. There has been much discussion about the line between "good" and "bad" drugs, how safe it is to use certain drug compounds, and what are the side effects of their usage. Most of the time, the answer lies in the dosage. Taken in right amounts, a compound may, by interacting with specific macromolecules in a biological system, alter their function, and cause a desired biological response.

When designing new drugs that will be used to treat a disease, one should first choose a target macromolecule (receptor, enzyme or nucleic acids). This is not a trivial task, considering that it takes a great deal of understanding which macromolecules are involved in a disease. Knowing their structure, properties, and functions is crucial, since many processes in an organism may, in one way or another, depend on activating or inhibiting a potential target.

Once a target macromolecule has been selected, the next step is to find a lead compound. This compound can interact with the target, and, although the level of interaction may not be perfect and there may exist some side effects, the compound will serve as a start for the drug design and drug development process. There are numerous ways in which a lead compound can be found. Many medicines used today are developed from a lead compound obtained from a natural source, such as plants, microorganisms, animals, or even from venoms

and toxins. Another source is existing drugs, as well as libraries of synthetic compounds that never reached the pharmaceutical market. When possible, natural ligands or modulators can be used as lead compounds. In case of enzymes, it can be their substrates or products. Nowadays, if the target structure is well known, it is possible to use molecular modelling software to design drugs that can fit the desired binding site (*de novo* drug design).

After discovering a lead compound, the drug design process can begin. Ideally, the starting compound should be modified so to obtain a good level of activity and selectivity for its target. Of course, all side effects should be brought to a minimum. The new drug should be easily synthetized, chemically stable, non-toxic, and should have acceptable pharmacokinetic properties.

There are various strategies that can be used to optimize a lead compound. Many of them rely on identifying important interactions that exist between the target and the drug, which are responsible for the drug activity, as well as parts of the drug molecule that can be modified so that new, more favourable interactions with the target can be formed. The simplest approach is to look at the binding state and draw conclusions from the interactions present. However, it has become clear that sometimes this is insufficient, as other factors influence the binding. An example would be change in entropy of both the target and the drug molecule, upon formation of the complex. Another example is the influence of the surrounding water on the binding.

To conclude, the key to successful drug design is to understand the system in question. Knowing which factors are most relevant for the system increases the chance that the changes made on a lead compound will result in a new drug molecule with enhanced activity and better properties.

The drug discovery process is far from simple. It takes a long time to find a molecule that fulfils all the criteria described above. After optimizing a lead compound, many steps follow before the drug is out on the market. First comes preclinical stage, in which the drug is tested for toxicity on various tissues. After that come the clinical trials in several stages, where safety and dosage of the drug is tested on humans. Finally, if the drug meets all the requirements, it must be legalized and then the production may begin on a large-scale. Altogether, it is a lengthy and costly procedure. One way to reduce the time and costs of drug development is to use computational methods to design drugs. There are many computational methods that can be used for finding and optimizing a lead compound, ranging from simple and inexpensive docking and scoring procedures to computationally demanding but more accurate quantum-mechanics based methods.[1]

## 1.2 Protein–ligand binding

Any atom, ion, or a molecule capable of forming non-covalent interactions with a target macromolecule, is called a ligand. Drug molecules used in the treatment of different diseases, together with substances naturally binding to macromolecules in a living cell, belong to this large group of compounds. The principles of binding and forming a macromolecule–ligand complex are the same for both drugs and natural substances. Therefore, term "ligand" will be used instead of "drug" from now on. In this thesis, only protein targets have been studied, so the focus will be on proteins. I will briefly describe the most important concepts of protein–ligand binding, along with important interactions between the two molecules, and the role of surrounding water in the binding event.

### 1.2.1 Proteins as drug targets

I will start this part by explaining what makes proteins the most important drug targets in medicinal chemistry. Two main characteristics of a protein are its structure and function. Most proteins have a unique three-dimensional structure, resulting from the arrangement of the amino acids linked together into long chains through peptide bonds. There are 20 common amino acids found in humans. These protein building blocks can be non-polar, polar, or charged, and their number, ratio, and order in a protein chain, together with the rigidity of the peptide bond, determine the shape and the function of a protein. Proteins represent the cogwheels in all living organisms, having key roles in many biological processes, e.g. transport of molecules through the cell membrane (transport proteins), speeding up chemical reactions – catalysis (enzymes), and communication between cells (receptors).[2] If a certain biological process is involved in a disease, proteins that take part in that process become a potential drug target, since inhibiting these proteins may incapacitate the source of the disease. Whether these target proteins belong to our own cells or to a foreign source (e.g. virus, bacteria, or fungi), the idea is to prevent the protein from performing its function.

The binding of a ligand (L) to a protein (P) takes place in water solution, and results in a protein–ligand complex (PL). This can be described by the reaction:

$$P + L \rightleftharpoons PL$$

The binding reaction (forward ⇀) and unbinding reaction (backward ↽) both have their own kinetic rate constants, $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$, respectively. When the equilibrium state is reached, the two reactions are balanced and we can write:

$$k_{\mathrm{on}}[\mathrm{P}][\mathrm{L}] = k_{\mathrm{off}}[\mathrm{PL}]$$

where the square brackets indicate equilibrium concentrations. The binding constant ($K_{\mathrm{b}}$), and the dissociation constant ($K_{\mathrm{d}}$), can be calculated from the ratio:

$$K_{\mathrm{b}} = k_{\mathrm{on}}/k_{\mathrm{off}} = [\mathrm{PL}]/[\mathrm{P}][\mathrm{L}] = 1/K_{\mathrm{d}}$$

The binding takes place at conditions of a constant temperature and pressure, which means that it will only happen if the change in Gibbs free energy ($\Delta G_{\mathrm{bind}}$) at equilibrium is negative (the process is spontaneous). The more negative $\Delta G_{\mathrm{bind}}$ is, the more stable will the complex be. The relationship between $K_{\mathrm{b}}$ and $\Delta G_{\mathrm{bind}}$ is given by:

$$\Delta G_{\mathrm{bind}} = -RT \ln K_{\mathrm{b}} C$$

where $R$ is the ideal gas constant (8.314 J/mol/K), $T$ is the temperature in degrees of Kelvin, and $C$ is the standard concentration (M). The binding free energy of a protein–ligand complex can also be divided into enthalpy ($\Delta H$) and entropy ($T\Delta S$) components:

$$\Delta G_{\mathrm{bind}} = \Delta H_{\mathrm{bind}} - T\Delta S_{\mathrm{bind}}$$

This means that the binding free energy depends on changes in the enthalpy and the entropy of the system upon ligand binding. Changes in the enthalpy come from breaking and forming non-covalent interactions between the protein, the ligand, and solvent molecules. If the newly formed interactions in the system are stronger than the interactions that existed before the binding, $\Delta H$ will be negative, in favour of the binding, and vice versa. Intermolecular interactions will be discussed in the next section. The change in entropy reflects the change in the level of the order in the system, coming from the change in the solvent entropy ($\Delta S_{\mathrm{solv}}$), the conformational entropy of the protein and the ligand ($\Delta S_{\mathrm{conf}}$), and from the loss of translational and rotational degrees of freedom of the protein and ligand upon complex formation ($\Delta S_{\mathrm{trans/rot}}$):[3]

$$\Delta S = \Delta S_{\mathrm{solv}} + \Delta S_{\mathrm{conf}} + \Delta S_{\mathrm{trans/rot}}$$

Quite often, when comparing a series of homologous ligands binding to a protein, it is observed that ligands that exhibit more favourable interactions to

the protein, i.e. more favourable enthalpy, also experience lower mobility, which results in less favourable entropy. This phenomenon is called enthalpy–entropy compensation,[4] and it demonstrates itself through linear dependence of $\Delta H$ and $\Delta S$ for a series of homologous ligands binding to the same protein.

### 1.2.2 Protein–ligand interactions

There are several types of intermolecular interactions that can be found between a protein and a ligand. The number and types of these interactions in a binding site depend on the structure of the protein and ligand, and on the type of functional groups present in the binding site.

The strongest of all the intermolecular interactions is ionic interaction that exists between functional groups with opposite charges (e.g. between carboxylate and ammonium ions). The strength, which can be anywhere between 20 and 400 kJ/mol,[5] depends on the distance between the ions, but also the surrounding environment, with this interaction being stronger in a hydrophobic environment than in a polar one.

The hydrogen bond is defined as the interaction between an electron-rich heteroatom and electron-deficient hydrogen atom, bound covalently to another electronegative atom. The strength of a hydrogen bond can vary (6–20 kJ/mol),[6] and is determined by the geometry of the bond, as well as the atoms involved. Even weak hydrogen bonds vastly contribute to the binding enthalpy, since there can be plenty of them in a single binding site.

Although quite weak (2–4 kJ/mol), the Van der Waals interaction or London dispersion dominates between hydrophobic regions of a protein and a ligand when they are close enough to each other. They are a consequence of temporal fluctuations in the electron distribution, which gives rise to a temporary dipole, which may induce a dipole in the neighbouring regions.

Molecules that have a permanent dipole can form a dipole–dipole interaction with another permanent dipole, so that the two dipoles are aligned parallel to each other, but in opposite directions. Another possibility would be interaction with an ion to form an ion–dipole interaction, or an interaction with a non-polar group resulting in a dipole–induced dipole interaction.

Special cases of these interactions will be studied in this thesis. In particular, the cation–π interaction, which arises between an electron-rich π system and a cation. The strength of this interaction is rather significant, with energies that are quite often of the same order of magnitude as hydrogen bonds. For that reason, cation–π interactions are important in nature, mainly in protein systems,

where they play a role in protein structure, molecular recognition, and enzyme catalysis.

Organofluorine compounds are often used as drugs. These compounds have a highly polarized C–F bond, that can interact with parts of a protein that have a partial positive charge. One such example is the amide group. The partially negative fluorine atom may interact with the carbon atom from the C atom of the amide, forming fluorine–amide interaction (C–F···C=O).

Finally, halogen bonds are also studied in this thesis. These bonds are short-ranged interactions that may occur between an electron-rich atom (halogen bond acceptor) and a positive region of a polarized halogen atom (halogen bond donor). The reason the halogen atom is polarized is that it is covalently bound to another atom, which makes the electrostatic potential of the halogen unevenly distributed, with a partially positive region opposite the covalent bond, called the σ-hole. The formation of a σ-hole largely depends on the polarizability of the halogen, which increases with the size and the decreasing electronegativity of the halogen (i.e. F < Cl < Br < I). Moreover, the ability to withdraw electrons of the molecular group covalently bound to the halogen also contributes to the σ-hole formation.

### 1.2.3 Role of water in protein–ligand binding

Although often neglected, water plays an important role in protein–ligand binding. First, water is the medium where the binding process takes place, meaning that before the binding, water molecules completely surround and interact with both the protein and the ligand. During the binding event, water molecules that solvate the binding site are partly displaced into bulk. This process is called desolvation and contributes both enthalpically and entropically to the total free energy of the system.

From the enthalpic point of view, water molecules displaced from the solute surface lose all favourable interactions with the solute. On the other hand, they form a new hydrogen-bond network in bulk water, whose interactions may or may not be more favourable than the interactions these water molecules formed with the solute molecules. Entropically speaking, water molecules experience different levels of order upon displacement. For example, it has been shown that the density of water molecules in a hydrophobic cavity is lower than the one in bulk, so water molecules displaced from such a cavity, become more ordered in bulk, which opposes the binding. Conversely, if water molecules in a protein cavity are well localized, their displacement into bulk makes them less ordered, which lowers their entropy and favours the binding.

The sign and magnitude of the enthalpy and entropy contributions of desolvation to the total free energy of the system are determined by the size and the shape of the solute, as well as the type and strength of interactions that exist between water molecules and the solute.

Another way water molecules may contribute to protein–ligand binding is as individual water molecules that stay "trapped" at the protein–ligand binding interface. These water molecules are called bridging water molecules, since they connect parts of the protein and the ligand through hydrogen bonds. Usually, these molecules are tightly bound and highly ordered. This is entropically disfavouring, because a water molecule must pay high entropic penalty in order to stay in a single place. It was estimated that this penalty is around 8 kJ/mol at 300 K.[7] Such a water molecule is kept in place by a strongly favourable enthalpy, coming from the interactions with the solute molecules. From the ligand-design point of view, these water molecules are ideal for displacement by a new ligand. If we assume that the new ligand can form equally good interactions with the protein, the system would benefit from releasing this water molecule, since it would then not have to pay the entropic penalty. It is useful to know if such water molecules exist when a lead compound binds to a protein. If so, a new ligand may be designed so that it displaces the water molecule, but at the same time forms good interactions with the protein, without losing the favourable interactions that existed between it and the protein.[8,9] There are several methods that may predict if such water molecules exist and in the Methods part of this thesis, I will discuss some of them.

## 1.3 Galectin-3

Galectin-3 is a protein used as a model system in several papers in this thesis. For that reason, it will be briefly introduced here.

Galectins are a family of proteins from the class of lectins. Their main characteristic is the carbohydrate-recognition domain (CRD), capable of binding ligands containing a β-galactoside moiety.[10,11] The CRD consists of approximately 135 amino acid residues, folded into two β-sheets that form a slightly bent sandwich (Figure 1.1). The β-sheet on the convex side has five strands, whereas that on the concave side, which can bind the carbohydrate ligand, has six strands.

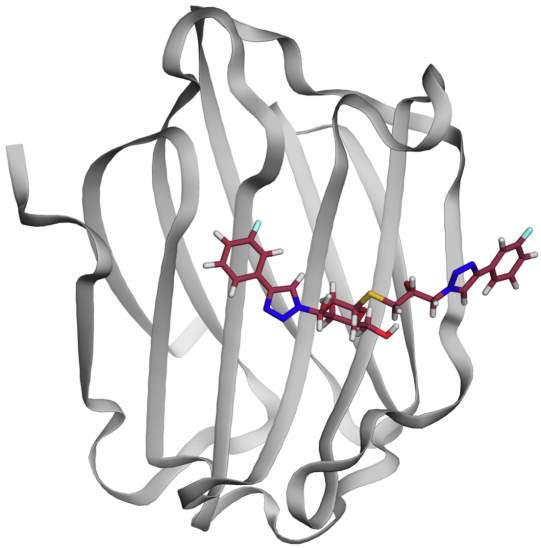


**Figure 1.1.** Galectin-3C in a complex with a ligand.

A total of 15 galectins have been found in mammals, 12 of which can be expressed in humans. Based on their structure, they have been classified into three types: prototypical, tandem-repeat, and chimeric galectins (Figure 1.2). The prototypical galectins contain a single CRD and may associate to form homo-dimers. The tandem-repeat galectins contain two distinct CRDs within one polypeptide chain, linked with a small peptide domain. Finally, the chimeric galectins have a single CRD and a very long amino-terminal domain, rich in proline, glycine, and tyrosine, which enable self-aggregation into oligomers consisting of up to five monomeric units. Galectin-3 is the only member of the chimera type of galectins. Depending on its concentration and the presence of the ligands, it can exist as a monomer or an oligomer. Oligomerization can sometimes be a problem when performing experiments involving galectin-3, so usually galectin-3C is used instead, in which the part responsible for oligomerization has been removed.

Galectins have many physiological functions. They are involved in several processes at the cellular level, such as cell signalling, adhesion, cell differentiation, migration, and autophagy. Through these processes, galectins regulate the organism's immune and inflammatory responses, but they are also involved in several diseases, such as fibrosis, cancer and heart disease. Galectin-3 has been shown to have great importance in cancer progression and metastasis, which makes it an interesting therapeutic target.[12]

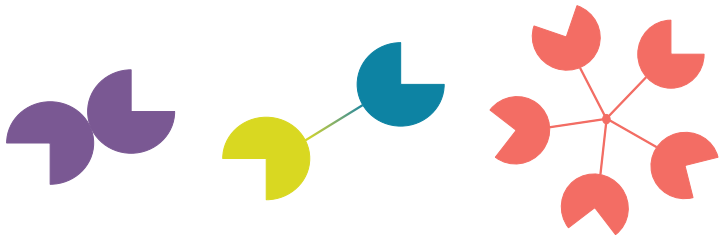


**Figure 1.2.** Different galectin types: a) prototypical galectins forming a dimer, b) tandem-repeat galectins, and c) five chimeric galectins forming a pentameric complex.

Galectin-3 is an excellent model system for the study of protein–ligand binding. It is biomedically relevant, but it is also experimentally well behaving (as galectin-3C) and its binding site allows studies of various types of protein–ligand interactions.

# 2 Molecular modelling

There are two different approaches to calculate the energy of a molecular system. In the first approach, called quantum mechanics, both electrons and nuclei are considered. On the other hand, the second approach, named molecular mechanics, treats atoms as a whole, meaning that electrons are neglected. In this chapter, I will present the basic concepts behind these two approaches, as well as the combination of the two.

## 2.1 Quantum mechanics

Quantum mechanical (QM) methods are based on the time-independent Schrödinger equation, introduced by Erwin Schrödinger in 1926:

$$\hat{H}\Psi = E\Psi$$

In this equation, $\hat{H}$ is the Hamiltonian operator for the system of interest, $\Psi$ is the wavefunction, which completely describes the system, and $E$ is the total energy of the system. For a three-dimensional chemical system, $\Psi$ is a function of the three Cartesian coordinates of all particles in the system, and the Hamiltonian operator is a sum of several energy terms:

$$\hat{H} = \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN} + \hat{T}_e + \hat{T}_N$$

The first three terms are the potential energy terms: $\hat{V}_{Ne}$ is the attractive potential between negatively charged electrons and positively charged nuclei, whereas $\hat{V}_{ee}$ and $\hat{V}_{NN}$ represent the repulsive potential energy terms for electrons and nuclei, respectively. The last two terms, $\hat{T}_e$ and $\hat{T}_N$, are the kinetic energy of the electrons and nuclei.[13]

### 2.1.1 Hartree–Fock theory

The Schrödinger equation can be solved exactly only for very simple systems, such as the hydrogen atom or hydrogen-like atoms. Even for the simplest molecule, $H_2^+$, the equation cannot be solved analytically, and various approximations and simplifications are used, in order to obtain a numerical solution. One such approximation is the Born–Oppenheimer approximation,[14] which allows electron and nuclear motions to be treated separately, since electrons are much lighter and faster compared to the nuclei.

The simplest QM approach used is Hartree–Fock (HF) theory,[15] which assumes that each electron moves in the average field of all other electrons (a mean-field approximation). In HF, the total *N*-electron wave function $\Psi$ is written as a Slater determinant:

$$\Psi^{\mathrm{SD}}(x_1, x_2, x_3, \ldots, x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_{\mathrm{N-1}}(x_1) & \varphi_{\mathrm{N}}(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_{\mathrm{N-1}}(x_2) & \varphi_{\mathrm{N}}(x_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \varphi_1(x_{\mathrm{N-1}}) & \varphi_2(x_{\mathrm{N-1}}) & \cdots & \varphi_{\mathrm{N-1}}(x_{\mathrm{N-1}}) & \varphi_{\mathrm{N}}(x_{\mathrm{N-1}}) \\ \varphi_1(x_{\mathrm{N}}) & \varphi_2(x_{\mathrm{N}}) & \cdots & \varphi_{\mathrm{N-1}}(x_{\mathrm{N}}) & \varphi_{\mathrm{N}}(x_{\mathrm{N}}) \end{vmatrix}$$

where $\varphi$ is a one-electron spin orbital, obtained by multiplying the spatial molecular orbital $\psi$ by a spin function $\alpha$ or $\beta$, representing the up- or down-spin of the electron:

$$\varphi = \psi(x) \cdot \begin{cases} \alpha(\omega) \\ \beta(\omega) \end{cases}$$

This approach satisfies the requirement of anti-symmetry for electrons, and obeys the Pauli exclusion principle. The molecular orbitals $\psi$ are obtained by a linear combination of $i$ atomic orbitals (LCAO) $\chi$:

$$\psi = \sum_i c_i \chi_i$$

where $c_i$ are coefficients, whose optimal values are determined using the variational principle, which states that the calculated ground-state energy $E$ of a system described by an approximate wave function is always larger than the ground state energy $E_0$ associated with the true wave function.

The set of atomic orbitals or basis functions used to construct molecular orbitals is known as the basis set. Usually, a linear combination of primitive Gaussian functions, known as contracted Gaussian-type orbitals (CGTO) are used.[16] Having one CGTO for each electron pair in the system is called minimal basis set and it is a required minimum for any calculation. However, it is better to use more than one CGTO, especially for valence electrons, giving split-valence basis sets, which are termed based on the number of CGTOs used as double-zeta (two CGTOs), triple-zeta (three CGTOs), *etc*.

On top of this, polarization and diffuse functions can be added to improve the accuracy. Polarization functions are GTOs of angular momentum $l + 1$. These functions allow molecular orbitals to be more asymmetric around nucleus, which is important for describing chemical bonding, because bonds are often polarized. Diffuse functions are GTOs with small exponents. These functions are important for describing anions, dipole moments, but also intra- and intermolecular bonding, since they accurately represent parts of the atomic orbitals that are distant from the atomic nuclei.

In principle, the larger the basis set, the better will the results be, since that increases accuracy of the calculations. However, it also increases the cost, so in practice, one should find a compromise between the accuracy and computational cost.

Even with an infinite basis set, the HF energy is still not exact, due to exclusion of electron correlation. There are many so-called post-HF methods, which describe electron–electron interactions, such as Møller–Plesset perturbation theory, configuration interaction methods, coupled cluster methods, and complete active space methods. An alternative to these rather costly "wave-function" methods is density functional theory, which instead uses electron density $\rho(r)$, a function of three Cartesian coordinates $r$, to characterize the system of interest.

### 2.1.2 Density functional theory

In 1964 Hohenberg and Kohn proved two theorems that enabled the use of density functional theory (DFT) for all systems. The first theorem, the existence theorem, states that the ground-state properties of a chemical system can be obtained from the electron density $\rho$, whereas the second theorem states that the method is variational, i.e. there exist a functional $F[\rho]$ that produces the ground-state energy $E$ which is minimized for the true ground state density $\rho$.[17] Proving these two theorems established DFT as a quantum chemical method, but there was still a lot of work to be done since exact formulation of $F[\rho]$ was

unknown. In 1965, Kohn and Sham developed a self-consistent field methodology that was based on fictitious system of non-interacting electrons, dividing the energy functional into several components:

$$E[\rho(r)] = T_e[\rho(r)] + V_{ne}[\rho(r)] + V_{ee}[\rho(r)] + \Delta T[\rho(r)] + \Delta V_{ee}[\rho(r)]$$

where $T_e$ is the kinetic energy of the non-interacting electrons, $V_{ne}$ is the nuclear–electron interaction, $V_{ee}$ is the classical electron–electron repulsion, $\Delta T$ is the correction to the kinetic energy of the electrons, and $\Delta V_{ee}$ represents all non-classical corrections to the electron–electron repulsion energy.[18]

The first three terms are computed in a similar way as in HF, with the largest difference in that the electron exchange energy that is included by definition in HF due to anti-symmetric Slater determinant wave function, is not included in DFT. Therefore, it must be included in the two corrective terms. The two terms also include electron correlation and together they give the exchange–correlation energy $E_{XC}$:

$$E_{XC}[\rho(r)] = \Delta T[\rho(r)] + \Delta V_{ee}[\rho(r)]$$

There are many approximate methods to calculate the exchange–correlation energy $E_{XC}$. The simplest is local-density approximation (LDA), which assumes that the density locally can be treated as a uniform electron gas. In order to account for the non-homogeneity of the true electron density, generalized gradient approximation (GGA) also includes the gradient of the density. So-called hybrid functionals also use a fraction of HF exchange. For example, the, B3LYP[19] functional can be expressed as:

$$E_{XC}^{\mathrm{B3LYP}} = (1-a)E_X^{\mathrm{LDA}} + \alpha E_X^{\mathrm{HF}} + b\Delta E_X^{\mathrm{B}} + (1-c)E_C^{\mathrm{LDA}} + cE_C^{\mathrm{LYP}}$$

with $a = 0.20$, $b = 0.72$, and $c = 0.81$.

## 2.2 Molecular mechanics

Molecular mechanics (MM) methods are based on a “ball and spring” model of a molecule, in which atoms are treated as balls, connected by bonds, represented by springs. The electronic structure of the molecule is neglected and the energy of the molecule is described by a force field.

The force field is an empirical potential energy function that gives the energy of a molecule as a function of the Cartesian coordinates of all atoms. A standard force field for biomolecular simulations usually consists of a sum of five terms:

$$U_{\text{total}} = U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}} + U_{\text{vdw}} + U_{\text{el}}$$

The first three terms describe the internal energy of the molecule, coming from all bonds, angles, and dihedrals present in the molecule, whereas the last two terms are the non-bonded terms.

In most variants of MM, covalent bonds are represented by springs, so the first term, $U_{\text{bonds}}$, uses the harmonic energy potential to describe covalent bond stretching around the equilibrium bond length, $r_0$:

$$U_{\text{bonds}} = \sum k_b (r - r_0)^2$$

where $k_b$ is the spring force constant for a given bond type $b$. Similarly, the bond angle term, $U_{\text{angles}}$, is also described by a harmonic potential:

$$U_{\text{angles}} = \sum k_a (\theta - \theta_0)^2$$

where $k_a$ is the angle force constant for angle $a$ involving three bonded atoms, and $\theta_0$ is the equilibrium angle. The dihedral term, $U_{\text{dihedrals}}$, uses a periodic function to describe torsion angle rotation around a bond:

$$U_{\text{dihedrals}} = \sum \frac{V_n}{2} [1 + \cos(n\phi - \delta)]$$

where, $V_n$ is the corresponding force constant, $n$ is the periodicity of the torsion angle $\phi$, and $\delta$ is the phase shift.

The fourth term is the van der Waals energy, which describes interactions between atoms that are not covalently bonded. At large interatomic distances, this term should be equal to zero, whereas at very short distances it should be strongly repulsive. However, at intermediate distances, where atoms are close to each other, but their electron clouds are not overlapping, this term should be slightly negative, due to induced dipole–dipole interactions, resulting from electron correlation. This behaviour is well described by the Lennard-Jones potential, which consists of two parts, a short-range repulsive term and a long-range attractive term:

$$U_{\text{vdw}} = \sum_{i<j} 4\varepsilon_{ij} \left[ \frac{\sigma_{ij}}{r_{ij}^{12}} - \frac{\sigma_{ij}}{r_{ij}^{6}} \right]$$

where $-\varepsilon_{ij}$ corresponds to the depth of the potential energy curve, $\sigma_{ij}$ is the distance between two atoms at which the potential energy is zero, and $r_{ij}$ is the distance between the two atoms. Finally, the last term describes the Coulomb electrostatic interaction energy between two atoms with partial atomic charges of $q_i$ and $q_j$:

$$U_{\text{el}} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon\varepsilon_0 r_{ij}}$$

where $\varepsilon_0$ is permittivity of vacuum, $\varepsilon$ is relative permittivity of the given medium (typically set to unity in atomistic simulations), and $r_{ij}$ is the inter-atomic distance.

All constants that appear in the expressions above are either obtained from experimental data or computed using high-level QM calculations. In this thesis, the AMBER ff14SB[20] force field was used for proteins, the generalized AMBER force field[21] (GAFF and GAFF2) was used for small organic compounds, and several force fields were employed for water molecules (TIP3P, TIP4P, TIP4P-Ew, and OPC)[22–24].

In MM, it is important to assign good partial charges to all atoms, since electrostatics contribute a lot to the non-bonded energy term. Charges assigned to the small organic compounds were calculated using the restrained electrostatic potential (RESP) approach.[25]

## 2.3 The combined QM/MM approach

The QM/MM approach is a way to apply the accurate, but expensive QM method on large chemical systems such as a protein–ligand complex.[26] The idea is that a small part of such system, usually containing the binding site (the ligand and its surroundings), is treated with QM methods, whereas the rest of the protein and solvent are treated with MM methods. In this way, the system is divided into a small QM part (subsystem 1) and large MM part (subsystem 2), as shown in Figure 2.1.

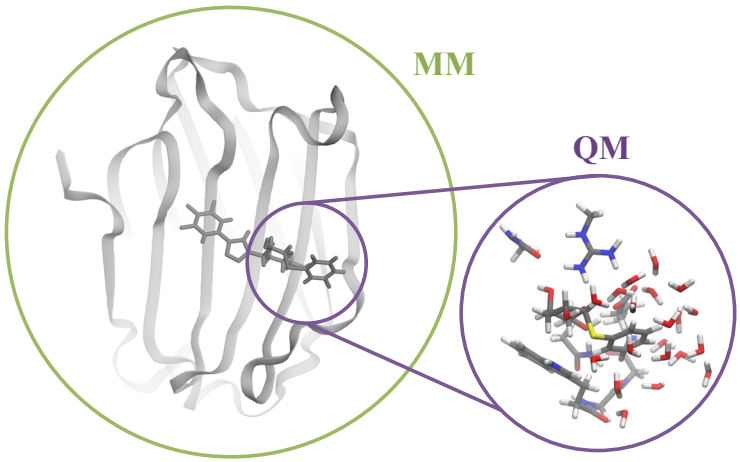



**Figure 2.1.** QM/MM system: the small QM part (purpule) and the large MM part (green).

There are several ways the QM/MM energy $E^{\mathrm{QM/MM}}$ can be calculated. A simple and intuitive approach is to calculate the MM energy of the entire system $E^{\mathrm{MM}}_{1+2}$, subtract the MM energy of the subsystem 1 $E^{\mathrm{MM}}_{1}$, and add the QM energy of subsystem 1 $E^{\mathrm{QM}}_{1}$:

$$E^{\mathrm{QM/MM}} = E^{\mathrm{MM}}_{1+2} - E^{\mathrm{MM}}_{1} + E^{\mathrm{QM}}_{1}$$

In this approach, all interactions between the subsystems 1 and 2 are treated at the MM level, which is fine for van der Waals interactions, but is somewhat problematic for electrostatic interactions, so an alternative is to treat the electrostatic interactions between the two subsystems with QM, by including the MM point charges in the QM calculation and turning off the corresponding interactions in the MM calculations by zeroing the charges in the subsystem 1:

$$E^{\mathrm{QM/MM}} = E^{\mathrm{MM}}_{1+2,q_1=0} - E^{\mathrm{MM}}_{1,q_1=0} + E^{\mathrm{QM}}_{1,q_2}$$

where $q_1$ and $q_2$ are the MM charges in the subsystems 1 and 2, respectively. The first approach is called mechanical embedding and the second electrostatic embedding.

# 3 Sampling methods

The energy of a molecule depends on its coordinates, as was discussed in the previous chapter. Since proteins and flexible ligands can adopt many, more or less similar conformations, it is necessary to calculate the total energy of the system as the average over the ensemble of conformations of the system in equilibrium. In addition, entropy of the system can also be calculated from the conformational ensemble. For these reasons, it is necessary to sample these possible conformations for the molecules involved to obtain reliable free energies of the protein–ligand binding. There are two approaches used to explore the conformational space of a molecule, molecular dynamics and Monte Carlo simulations. Both methods have advantages and disadvantages that will be discussed in this chapter.

## 3.1 Molecular dynamics

Molecular dynamics (MD) simulations employ the laws of classical mechanics, in particular Newton's second law[27], to move all atoms in the system of interest:

$$F_i = m_i a_i = m_i \frac{d^2 r_i(t)}{dt^2}$$

where $F_i$ is the force acting on an atom $i$ with mass $m_i$, and $a_i$ is acceleration, which also can be expressed as the second derivative of the position $r_i(t)$ of the atom $i$ with respect to time $t$.[28] The forces acting on each atom can be calculated from the derivative of the potential energy $U$ with respect to the position $r(t)$:

$$F(t) = -\frac{dU}{dr(t)}$$

MD simulations typically start by assigning initial positions $r(t)$ and velocities $v(t)$ to all atoms in a system. Then, it is possible to calculate the

corresponding positions and velocities at time $t + \Delta t$, by integrating the Newton's equations of motion, where $\Delta t$ is the time step used in the simulations. For a very small time step, $\Delta t$ $(0.5 - 1$ fs), it is possible to solve the equations of motion numerically, where the positions of the atoms in the system are approximated by a Taylor expansion:

$$r(t + \Delta t) = r(t) + \frac{\partial r(t)}{\partial t}(\Delta t) + \frac{1}{2}\frac{\partial^2 r(t)}{\partial t^2}(\Delta t)^2 + \cdots$$

or:

$$r(t + \Delta t) = r(t) + v(t)(\Delta t) + \frac{1}{2}a(\Delta t)^2 + \cdots$$

where $a$ is acceleration obtained from the calculated force for each atom. After calculating the new positions of all atoms, it is possible to update the energy and forces and iterate the procedure as many times as needed (Figure 3.1). This way, an MD simulation gives a trajectory, which shows how positions and velocities of all atoms vary with time.

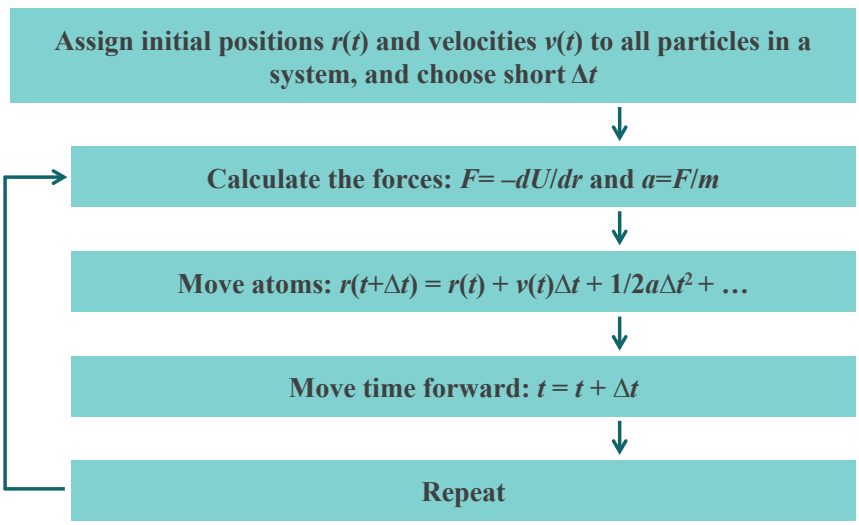



**Figure 3.1.** A simple MD algorithm.

There are several ways in which a MD simulation can be made more efficient, in terms of speed and computational demand. For instance, disabling vibrations of all bonds involving hydrogen atoms, by constraining them to their equilibrium values (e.g. using the SHAKE algorithm[29]), enables the use of a larger time step (2 fs, rather than 0.5 fs), which speeds up the simulations.

Another commonly used approach is to reduce the number of calculated non-bonded interactions, by employing a cut-off distance, beyond which the non-bonded interactions are not calculated every time step. While this works fine for the short-ranged van der Waals interactions, it is a problem for long-ranged electrostatic interactions. If periodic boundary conditions are used, in which the simulation box is replicated infinitely in all directions, long-range electrostatic interactions can be treated by Ewald summation.[30]

## 3.2 Metropolis Monte Carlo simulations

The term Monte Carlo (MC) applies to all simulation techniques that use stochastic methods to generate new configurations of a system of interest, i.e. are based on random sampling. This means that in MC, an ensemble average is obtained, rather than a time average.

In the canonical ensemble (the number of particles $N$, the volume $V$, and the temperature $T$ are constant), the ensemble average of a given property $\langle A \rangle$ can be obtained from the following:

$$\langle A \rangle = \frac{\int A(r^N) e^{-U(r^N)/k_B T} dr^N}{\int e^{-U(r^N)/k_B T} dr^N}$$

where $k_B$ is the Boltzmann constant, $U$ is the potential energy, and $r^N$ denotes the configuration of an $N$-particle system (i.e., the positions of all $N$ particles).[31] The probability density of finding the system in configuration $r^N$ is:

$$\rho(r^N) = \frac{e^{-U(r^N)/k_B T}}{\int e^{-U(r^N)/k_B T} dr^N}$$

where the denominator is the configurational integral. If one can randomly generate $M$ points in configuration space according to this equation, then $\langle A \rangle$ can be expressed as:

$$\langle A \rangle \approx \frac{1}{M} \sum_{i=1}^{M} A(r_i^N)$$

One way to obtain such configurations is to generate them using Metropolis algorithm:[32] First, an initial configuration $r_i$ is defined, having the potential

energy $U(r_i)$. Then, a new trial configuration $r_j$ is generated, by adding a random displacement to one atom. The potential energy of this new configuration is $U(r_j)$. Now, if $U(r_j) \leq U(r_i)$ the trial move is accepted. Otherwise, a random number between 0 and 1, $\rho$, is generated and the trial move is accepted if

$$\rho < e^{-[U(r_j)-U(r_i)]/k_B T}$$

The Metropolis algorithm is illustrated in Figure 3.2. Possible MC moves involve translations and rotations of the particles in the system. It is also possible to use "unphysical" moves, as is discussed in the next section.

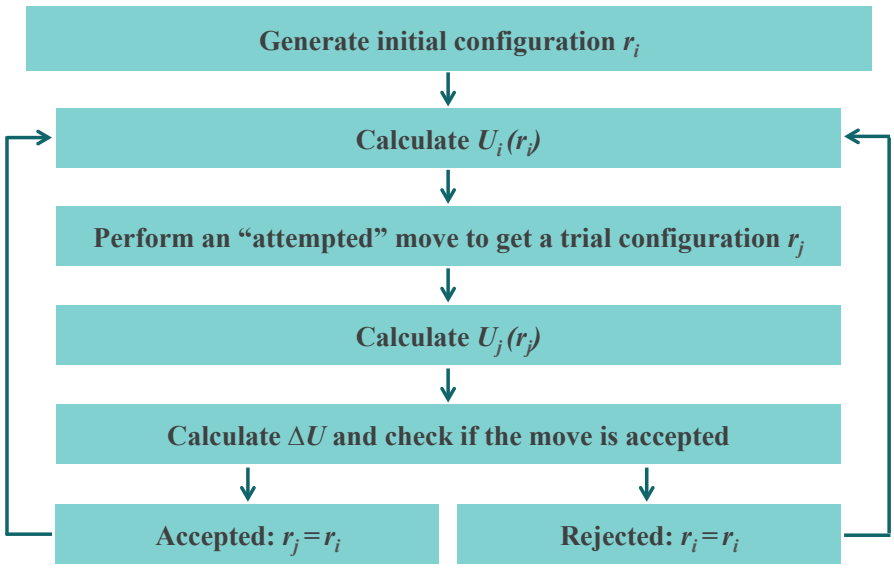



**Figure 3.2.** A simple MC algorithm.

### 3.2.1 Grand canonical Monte Carlo simulations

In this thesis, grand canonical Monte Carlo (GCMC) simulations, as implemented by Essex and coworkers[33] in the ProtoMS software package,[34] were employed to study water structure and energetics around protein–ligand binding sites.

In these simulations, rotations and translations of water molecules in a solvated protein–ligand system are performed. In addition, attempts are made to insert and delete water molecules in a region around the binding site. To do this, the region is coupled to an ideal-gas reservoir of water molecules and water molecules are allowed to exchange between the two (Figure 3.3, left).

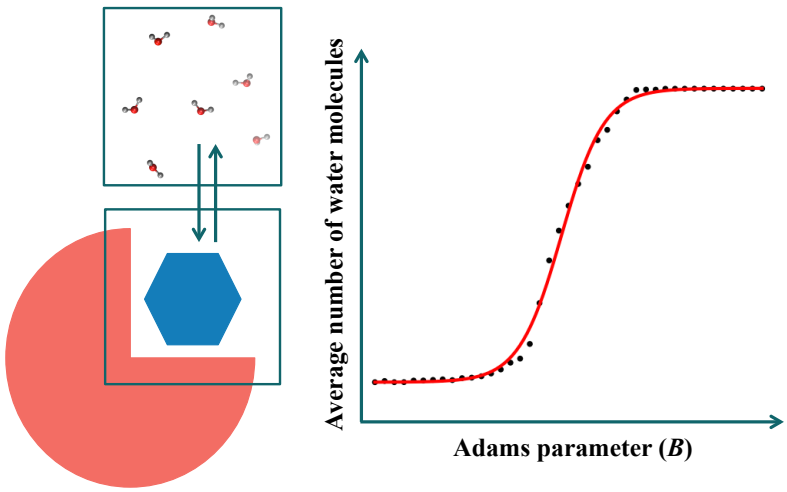



**Figure 3.3.** Left: Schematic diagram of water molecules in an ideal-gas reservoir coupled to a region within the protein. Right: GCMC titration curve showing the average number of inserted water molecules at different values of the Adams parameter.

The chemical potential of a system at constant temperature $T$ and volume $V$, where the number of molecules of only one species can vary, is defined as:

$$\mu = \left(\frac{\partial F(N,V,T)}{\partial N}\right)_{T,V}$$

where $N$ is the instantaneous number of molecules in the system and $F$ is the Helmholtz free energy of the system. This means that a chemical equilibrium between the region with the chemical potential $\mu_{\text{region}}$ coupled to the gas reservoir with potential $\mu_{\text{reservoir}}$ will be established when $\mu_{\text{region}} = \mu_{\text{reservoir}}$.

The reservoir in the grand canonical ensemble is not explicitly considered and is completely defined by the chemical potential.

Instead of chemical potential, Essex and coworkers use the Adams formulation of GCMC[35] to calculate the acceptance probabilities for inserting a molecule from the gas reservoir and deleting a particle, respectively:

$$P_{\text{insert}} = \min\left[1, \frac{1}{N+1} e^{B} e^{-\Delta U/k_B T}\right]$$

and:

$$P_{\text{delete}} = \min\left[1, N e^{-B} e^{-\Delta U/k_B T}\right]$$

where $k_B$ is Boltzmann's constant, $\Delta U$ is change in energy caused by the trial move, $N$ is the number of water molecules in the system, and $B$ is the Adams parameter,[33,35] calculated as:

$$B = \frac{\mu'(N)}{k_B T} + \ln N$$

where $\mu'$ is the excess chemical potential, i.e. the difference between the chemical potential of a given species $\mu$ and that of an ideal gas $\mu_{\text{ideal}}$ under the same conditions:

$$\mu'(N,V,T) = \mu(N,V,T) - \mu_{\text{ideal}}(N,V,T)$$

Simulating at a constant $B$ guarantees that the simulation is run at a constant $\mu$, and like the temperature, $B$ must be set prior to running the GCMC simulation. The parameter $B$ influences the probability that a water molecule is inserted or deleted, and therefore, the number of inserted water molecules directly depends on $B$.

By performing GCMC simulations at different $B$ values, a virtual titration is performed. From the titration curve showing the number of inserted water molecules as a function of $B$ (Figure 3.3, right), it is possible to calculate the free energy of transfer of $N$ water molecules from the gas reservoir to the region of interest:

$$\Delta F_{\text{trans}}(N_i \to N_f) = k_B T \left( N_f B_f - N_i B_i + \ln \frac{N_i!}{N_f!} - \int_{B_i}^{B_f} N(B) dB \right)$$

where $N_i$ and $N_f$ are initial and final number of water molecules, respectively, and the corresponding $B$ values, $B_i$ and $B_f$. This method is called grand

canonical integration (GCI).[33] To improve the precision, the GCMC titration data is fitted to a logistic equation:

$$N(B)_{\text{approx}} = \sum_{i=1}^{m} \frac{n_i}{1 + e^{(\omega_{0i} - \omega_i B)}}$$

where $m$, $n$, $\omega_{0i}$ and $\omega_i$ are fitted parameters.

To obtain the free energy of transfer of $N$ water molecules from bulk water to the region of interest, i.e. the binding free energy, $\Delta F_{\text{bind}}$, one use the thermodynamic cycle shown in Figure 3.4, giving the equation:

$$\Delta F_{\text{bind}}(N) = \Delta F_{\text{trans}}(N) - \Delta F_{\text{hyd}}(N)$$

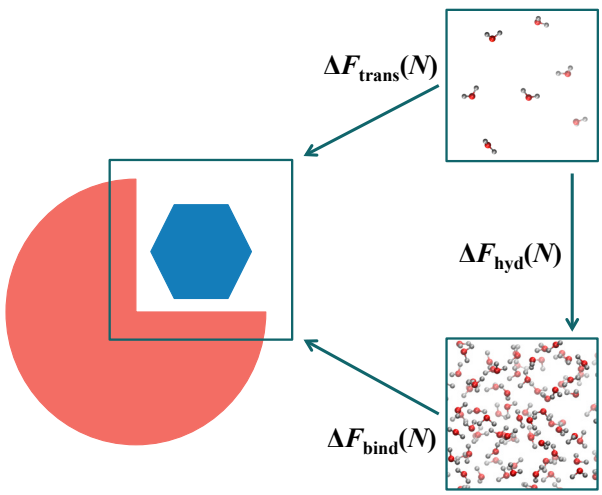



**Figure 3.4.** Thermodynamic cycle used to calculate the binding free energy of $N$ water molecules to a GCMC region of interest.

where $\Delta F_{\text{hyd}}(N)$ is the free energy of hydration of $N$ water molecules, obtained by multiplying the number of water molecules with excess chemical potential of a single water molecule ($\mu'_{\text{hyd}}$), which can be taken from experiments or from simulations of the particular water model used.[33] Finally, the number of water molecules that minimizes $\Delta F_{\text{bind}}$ ($N_{\text{opt}}$) and the corresponding $B$ value that produces the optimal occupancy of a cavity are given by:

$$\mu'_{\text{prot}}(N_{\text{opt}}) = \mu'_{\text{hyd}}$$

$$B_{\text{opt}} = \frac{\mu'_{\text{hyd}}}{k_B T} + \ln N_{\text{opt}}$$

# 4 Thermodynamics of protein–ligand binding

In the introduction part of this thesis, I briefly discussed the importance of considering both the enthalpy and entropy when discussing protein–ligand binding free energies. The focus of this chapter will be calculations of entropy, since the enthalpy part was covered in the previous chapters. As mentioned before, the change in entropy may come from the change in the solvent entropy ($\Delta S_{\mathrm{solv}}$), the conformational entropy of the protein and the ligand ($\Delta S_{\mathrm{conf}}$), and from the loss of translational and rotational degrees of freedom of the protein and ligand upon complex formation ($\Delta S_{\mathrm{trans/rot}}$):

$$\Delta S = \Delta S_{\mathrm{solv}} + \Delta S_{\mathrm{conf}} + \Delta S_{\mathrm{trans/rot}}$$

In this thesis, the conformational entropy of the protein and the ligand ($\Delta S_{\mathrm{conf}}$) and the solvent entropy around the binding site ($\Delta S_{\mathrm{solv}}$) were calculated from MD generated trajectories. The theory behind these calculations will be introduced below.

## 4.1 Conformational entropy of the protein and the ligand

There are many ways in which conformational entropy can be calculated. In this thesis, we used dihedral-distribution histogramming (DDH).[36,37] In this approach, the Cartesian coordinates of the protein are converted to internal (bond, angle, and torsion) coordinates. The entropy contributions from the bond and angle fluctuations are small and essentially constant during ligand binding,[38] so only the dihedral angles are used to calculate the conformational entropy. The distribution for each dihedral angle $i$ is then approximated by a discrete histogram and the entropy is calculated from:

$$S_i = \frac{R}{2} - R \ln N_{\text{bin}} - R \sum_{j=1}^{N_{\text{bin}}} p_i(j) \ln p_i(j)$$

where $R$ is the gas constant, $p_i(j)$ is the probability that the dihedral angle is found in bin $j$, and $N_{\text{bin}}$ is the number of bins.[39] The first two terms are normalization factors, giving the entropy of a free rotor $R/2$ for a uniform distribution. These terms cancel for relative entropies. The number of bins used in this thesis is 72 (5° in each bin).[38]

## 4.2 Solvent thermodynamics

The majority of methods used to analyse the thermodynamics of the solvent involved in protein–ligand binding, are based on inhomogeneous solvation theory (IST).[40] This theory uses statistical thermodynamics to extract information about the solvent around the solute, using MD generated trajectories.

According to IST, the solvation entropy $\Delta S_{\text{solv}}$, of a solute may be divided into contributions from solute–water correlations $\Delta S_{\text{sw}}$ and water–water correlations $\Delta S_{\text{ww}}$:

$$\Delta S_{\text{solv}} = \Delta S_{\text{sw}} + \Delta S_{\text{ww}}$$

Here, $\Delta S_{\text{solv}}$ is approximated, so that it accounts for only the solute–water term:

$$\Delta S_{\text{solv}} \approx \Delta S_{\text{sw}} \equiv -\frac{k_B \rho^0}{8\pi^2} \int g_{\text{sw}}(r, \omega) \ln g_{\text{sw}}(r, \omega)\, dr d\omega$$

where $k_B$ is Boltzmann's constant, $\rho^0$ is the number density of bulk solvent, and $g_{\text{sw}}(r, \omega)$ is the solute–water pair-correlation function in the solute frame of reference, where $r$ is the location of the water oxygen relative to the solute, and $\omega$ is Euler angles in the solute frame of reference. By rewriting $g_{\text{sw}}(r, \omega)$ as the product of a translational distribution function $g_{\text{sw}}(r)$ and an orientational distribution function conditioned on the position $g_{\text{sw}}(\omega|r)$, the solute–water entropy term can be further divided into translational and orientational terms:

$$\Delta S_{\text{sw}} = \Delta S_{\text{sw}}^{\text{trans}} + \Delta S_{\text{sw}}^{\text{orient}}$$

where

$$\Delta S_{\mathrm{sw}}^{\mathrm{trans}} \equiv -k_B \rho^0 \int g_{\mathrm{sw}}(r) \ln g_{\mathrm{sw}}(r)\, dr$$

and

$$\Delta S_{\mathrm{sw}}^{\mathrm{orient}} \equiv \rho^0 \int g_{\mathrm{sw}}(r) S^{\omega}(r) dr$$

$$S^{\omega}(r) \equiv -\frac{k_B}{8\pi^2} \int g_{\mathrm{sw}}(\omega|r) \ln g_{\mathrm{sw}}(\omega|r)\, d\omega$$

The solvation energy $\Delta E_{\mathrm{solv}}$ also consists of solute–water $\Delta E_{\mathrm{sw}}$ and water–water $\Delta E_{\mathrm{ww}}$ terms:

$$\Delta E_{\mathrm{solv}} = \Delta E_{\mathrm{sw}} + \Delta E_{\mathrm{ww}}$$

The $\Delta E_{\mathrm{sw}}$ term is calculated in the following way:

$$\Delta E_{\mathrm{sw}} = \rho^0 \int g_{\mathrm{sw}}(r) \Delta E_{\mathrm{sw}}(r) dr$$

$$\Delta E_{\mathrm{sw}}(r) \equiv -\frac{1}{8\pi^2} \int g_{\mathrm{sw}}(\omega|r) U_{\mathrm{sw}}(r, \omega) d\omega$$

where $U_{\mathrm{sw}}(r, \omega)$ is the solute–water interaction potential. Similarly, the $\Delta E_{\mathrm{ww}}$ term is calculated as:

$$\Delta E_{\mathrm{ww}} = \rho^0 \int g_{\mathrm{sw}}(r) \Delta E_{\mathrm{ww}}(r) dr$$

$$\begin{aligned} \Delta E_{\mathrm{ww}}(r) \equiv \left(\frac{1}{8\pi^2}\right)^2 \rho^0 \int g_{\mathrm{sw}}(\omega|r) \\ \times \left[g_{\mathrm{sw}}(r', \omega') g_{\mathrm{ww}}(r, \omega, r', \omega') - g_{\mathrm{ww}}^0(r, \omega, r', \omega')\right] \\ \times U_{\mathrm{ww}}(r, \omega, r', \omega') d\omega\, dr' d\omega' \end{aligned}$$

where $g_{\mathrm{ww}}(r, \omega, r', \omega')$ is the pair distribution function between water molecules with spatial and orientational coordinates $(r, \omega)$ and $(r', \omega')$ close to the solute, $g_{\mathrm{ww}}^0(r, \omega, r', \omega')$ is the corresponding quantity in bulk water, and $U_{\mathrm{ww}}(r, \omega, r', \omega')$ is the water–water interaction potential.

### 4.2.1 Grid inhomogeneous solvation theory

In this thesis, the grid inhomogeneous solvation theory (GIST), developed by Gilson and coworkers[41], is used. This implementation of IST discretizes the equations of IST onto a three-dimensional grid placed over a region of interest (Figure 4.1). The grid is divided into small boxes, called voxels and indexed with $k$, for which the thermodynamic quantities are calculated. The spatial integrals that appear in the IST expressions are replaced by sums over the voxels of the grid. This is an approximation, that becomes exact when the volume of the voxels $V_k \to 0$. This means that the smaller the voxels, the better the approximation.

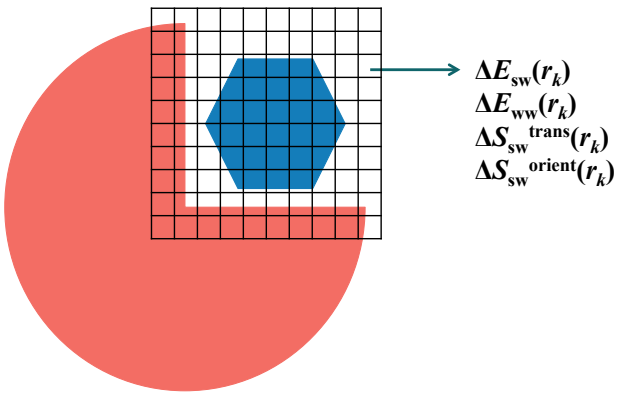



**Figure 4.1.** Schematic diagram of of GIST's gridded water properties in a binding site.

In GIST, the total translational solvation entropy of a region $R$ is given by:

$$\Delta S_{\mathrm{sw}}^{R,\mathrm{trans}} \approx \sum_{k \in R} \Delta S_{\mathrm{sw}}^{\mathrm{trans}}(r_k)$$

$$\Delta S_{\mathrm{sw}}^{\mathrm{trans}}(r_k) \approx k_B \rho^0 V_k g(r_k) \ln g(r_k)$$

$$g(r_k) \equiv \frac{N_k}{\rho^0 V_k N_f}$$

where $r_k$ is the location of the voxel $k$, $N_k$ is the number of water molecules within voxel $k$ summed across all frames $N_f$.

Similarly, the total orientational entropy of a region $R$ is:

$$\Delta S_{\mathrm{sw}}^{R,\mathrm{orient}} \approx \sum_{k \in R} \Delta S_{\mathrm{sw}}^{\mathrm{orient}}(r_k)$$

$$\Delta S_{\mathrm{sw}}^{\mathrm{orient}}(r_k) \approx \rho^0 V_k g(r_k) S^{\omega}(r_k)$$

$$S^{\omega}(r_k) = \frac{-k_B}{N_k}\left(-\gamma + \sum_{i=1}^{N_k} \ln\left[\frac{g(\omega_i|r_k)}{N_k}\right]\right)$$

where $\gamma$ is Euler's constant. The corresponding energy terms are calculated in the following way:

$$\Delta E_{\mathrm{sw}}^{R} \approx \sum_{k \in R} \Delta E_{\mathrm{sw}}(r_k)$$

$$\Delta E_{\mathrm{ww}}^{R} \approx \sum_{k \in R} \Delta E_{\mathrm{ww}}(r_k) - \frac{1}{2}\sum_{k \in R}\sum_{\substack{l \neq k \\ l \in R}} E_{\mathrm{ww}}(r_k, r_l)$$

Finally, the free energy of solvation $\Delta G(r_k)$ for voxel $k$ is:

$$\Delta G(r_k) = \Delta E_{\mathrm{total}}(r_k) - T\Delta S_{\mathrm{sw}}^{\mathrm{total}}(r_k)$$

where:

$$\Delta E_{\mathrm{total}}(r_k) \equiv \Delta E_{\mathrm{sw}}(r_k) + \Delta E_{\mathrm{ww}}(r_k)$$

and:

$$\Delta S_{\mathrm{sw}}^{\mathrm{total}}(r_k) \equiv \Delta S_{\mathrm{sw}}^{\mathrm{trans}}(r_k) + \Delta S_{\mathrm{sw}}^{\mathrm{orient}}(r_k)$$

# 5 Free-energy calculations

The difference in affinity between two structurally similar ligands binding to a protein can be calculated using *alchemical* free energy calculations. These calculations employ unphysical (alchemical) intermediates in order to estimate free energies of various physical processes. Some examples of these calculations are the free energy of transfer of a small molecule from gas to water (free energy of solvation), the absolute binding free energy of a ligand binding to a protein, the free energy of a mutation of a protein side chain, or a modification of a ligand bound to a protein (relative binding affinity).

## 5.1 Thermodynamic cycle

Free energy calculations are based on thermodynamic cycles, in which the end-states are defined. For the calculations of relative binding affinity, the thermodynamic cycle is shown in Figure 5.1. It contains four end-states, two in which the either ligand A or ligand B is bound to the protein, and two in which ligand A or ligand B is free in solution, far from the protein. The upper and lower horizontal reactions (arrows) represent the binding of either ligand to the protein, i.e. the absolute binding affinities for each ligand. These are rather hard to calculate explicitly. If we are only interested in the relative binding affinity of the two ligands to the protein, we can instead study the two vertical reactions in the figure. These correspond to the difference in free energy of converting ligand A to ligand B both when they are bound to the protein and when they are free in solution. Since the total free energy of going around a thermodynamic cycle vanishes, the relative binding affinity is given by the equation

$$\Delta\Delta G_{\mathrm{bind}} = G^{\mathrm{B}}_{\mathrm{bind}} - G^{\mathrm{A}}_{\mathrm{bind}} = G^{\mathrm{bound}}_{\mathrm{A}\to\mathrm{B}} - G^{\mathrm{free}}_{\mathrm{A}\to\mathrm{B}}$$

These calculations are much less demanding. Of course, the absolute binding affinities in this case remain unknown. However, for the purpose of drug-design, one is mainly interested in the difference in binding affinity between two ligands, which tells us which ligand is the stronger binder.

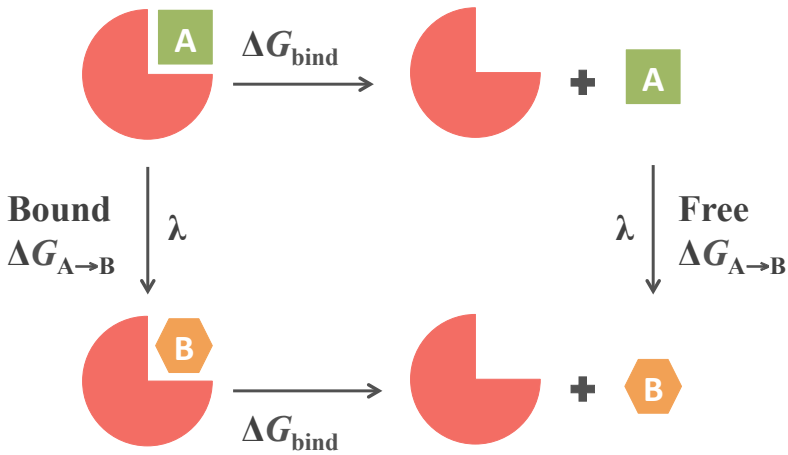



**Figure 5.1.** Thermodynamic cycle used for the calculations of relative binding affinities.

## 5.2 Free-energy fundamentals

For two states, A and B, with their respective potentials, $U_A$ and $U_B$, the potential energy difference is:

$$\Delta U = U_B - U_A$$

and the free energy difference in the canonical ensemble is directly related to the ratio of probabilities of the two states through their partition functions:

$$\Delta F_{A\rightarrow B} = -k_B T(\ln Q_B - \ln Q_A) = -k_B T \ln \frac{Q_B}{Q_A} = -k_B T \ln \frac{\int e^{-U_B(x)/k_B T} dx}{\int e^{-U_A(x)/k_B T} dx}$$

where $\Delta F$ is the Helmholtz free energy difference between state B and state A, $Q_A$ and $Q_B$ are the canonical partition functions, $T$ is the temperature, and $x$ indicates that the potential energy depends on the coordinates of the particles in the system.

The simplest way to calculate the free energy from simulations is to use the Zwanzig relationship[42]:

$$\Delta F_{A\rightarrow B} = -k_B T \ln\langle e^{-(U_B - U_A)/k_B T}\rangle$$

in which the energy is calculated from the ensemble average over configurations sampled from the reference state A. This method is also known as free-energy perturbation (FEP), which might be confusing sometimes, as the

abbreviation FEP is also used for all free-energy methods. Therefore, it is better to denote this method as exponential averaging. The method is exact, but its main drawback is that the calculations converge only when the difference between the two states is small. For most studied systems, this is not the case.

To overcome this problem, the transformation from A to B is usually broken into several intermediate states by introducing a coupling parameter $\lambda$, which can have values from 0 to 1, and calculating the potential energy as a linear combination of the two end-state potentials $U_{\mathrm{A}}$ and $U_{\mathrm{B}}$:

$$U(\lambda) = (1-\lambda)U_{\mathrm{A}} + \lambda U_{\mathrm{B}}$$

By running the simulations using different modified potentials (gradually increasing the λ values), we alchemically transform ligand A to ligand B.

There are several other ways to estimate the free energy of mutating ligand A to ligand B. One of the most commonly used methods is thermodynamic integration (TI),[43] in which the free energy is calculated by integration the ensemble-averaged derivative of the potential energy with respect to $\lambda$:

$$\Delta F_{\mathrm{A}\to\mathrm{B}} = -k_B T \ln Q(\lambda)$$

$$\frac{\partial F}{\partial \lambda} = -\frac{k_B T}{Q}\frac{\partial Q}{\partial \lambda} = \frac{\partial U}{\partial \lambda}$$

$$\Delta F = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda$$

Another approach is the Bennett acceptance ratio (BAR) method.[44,45] This method requires information from two states in order to estimate free energy differences. Free energies are calculated based on the equation:

$$e^{-(\Delta F - C)/k_B T} = \frac{\langle f[(U_{\mathrm{B}} - U_{\mathrm{A}} - C)/k_B T]\rangle_{\mathrm{A}}}{\langle f[(U_{\mathrm{A}} - U_{\mathrm{B}} + C)/k_B T]\rangle_{\mathrm{B}}}$$

where $f(x) = (1 + e^{x/k_B T})^{-1}$ is the Fermi function, and $C$ is a constant that is iteratively calculated until the above ensemble averages, denoted by angular brackets, are equal. Given a fixed length of each simulation, fewer intermediate states are required for BAR than for TI to give equivalent level of statistical precision.

The multistate Bennett acceptance ratio (MBAR) method[46] is also used in this thesis. It is an extension to the BAR method, in which data from all $\lambda$ states are used to estimate free energy differences.

## 5.3 QM/MM free-energy perturbation

The accuracy of free energy calculations largely depends on the choice of the potential energy function. Usually, an MM potential is used, since it allows extensive sampling, which is another important factor determining the accuracy of these calculations. However, the MM potential often fails to properly describe some types of interactions that can be found in a protein–ligand system. Therefore, it is essential to develop free energy methods that use energy potentials that can deal with a broader spectrum of interactions and capture most of the chemistry involved in protein–ligand binding, such as a QM/MM potential.[47] Simply replacing the MM potential with QM/MM energy potential improves the free-energy estimates, but it also increases the cost of sampling by many orders of magnitude.[48,49] Estimating the difference in relative binding affinity of two ligands A and B binding to a protein using this approach ($\Delta\Delta G^{\mathrm{QM/MM}}$) corresponds to the upper purple arrow in Figure 5.2.

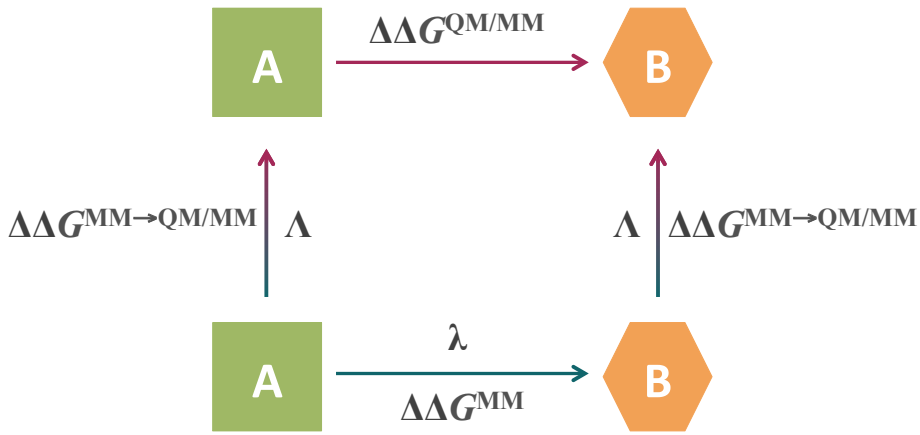



**Figure 5.2.** Thermodynamic cycle used for the QM/MM free energy calculations of relative binding affinities.

Another possibility to calculate $\Delta\Delta G^{\mathrm{QM/MM}}$ is to use the reference-potential method in which the free energy is first estimated at the MM level, $\Delta\Delta G^{\mathrm{MM}}$ (blue arrow), and then additional free-energy calculations are employed at the end-points to calculate the free-energy of going from the MM to the QM/MM potential $\Delta\Delta G^{\mathrm{MM\rightarrow QM/MM}}$ (vertical arrows).[50,51] According to the thermodynamic cycle in Figure 5.2, $\Delta\Delta G^{\mathrm{QM/MM}}$ can then be calculated from:

$$\Delta\Delta G_{\mathrm{A\rightarrow B}}^{\mathrm{QM/MM}} = -\Delta\Delta G_{\mathrm{A}}^{\mathrm{MM\rightarrow QM/MM}} + \Delta\Delta G_{\mathrm{A\rightarrow B}}^{\mathrm{MM}} + \Delta\Delta G_{\mathrm{B}}^{\mathrm{MM\rightarrow QM/MM}}$$

In analogy with the use of the λ coupling parameter above, the MM→QM/MM free energies can be calculated based on the energy function:

$$E(\Lambda) = (1 - \Lambda)E^{\mathrm{MM}} + \Lambda E^{\mathrm{QM/MM}}$$

where $E^{\mathrm{MM}}$ is the MM energy, $E^{\mathrm{QM/MM}}$ is the QM/MM energy and Λ is a coupling parameter going from 0 to 1. This is called the reference-potential approach with QM/MM sampling, RPQS.[52,53]

# 6 Summary of the papers

The papers in this thesis can be divided into two main groups:

- Development and application of MM- and QM/MM-FEP methods to calculate protein–ligand binding affinities, and testing of the methods to determine water structure in protein–ligand binding site (Papers I, II, III).
- Application of computational methods to study protein–ligand binding on Galectin-3 model system, with focus on effects of solvation, entropy, and specific protein–ligand interactions (Papers IV–VIII). These projects involve extensive experimental studies of the binding, which were performed by our colleagues in a large interdisciplinary collaboration. Our calculations were performed to understand and explain the trends found experimentally.

## 6.1 Paper I

**Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations**

In this paper, binding affinities of three sets of ligands binding to the heat-shock protein 90 in the D3R grand challenge blind test competition were estimated using four different methods (docking, MM/GBSA, QM/MM, and FEP). The results were rather disappointing, with poor and often negative correlation and Kendall's tau values for most of the methods and ligand sets. The mean absolute deviations (MADs) for FEP calculations were 4–15 kJ/mol with maximum errors of up to 26 kJ/mol.

For one of the sets, the problem could be traced to the displacement of one or two water molecules by one of the ligands. After employing GCMC calculations to deduce which water molecules dissociate for the various ligands, and rerunning the FEP simulations with those water molecules included in the perturbations, the results improved for that set, giving a perfect correlation ($R$ = 1.0) and lower MADs (4–5 kJ/mol), as shown in Table 6.1.

**Table 6.1. Performance of FEP calculations of relative binding free energies (based on two crystal structures) compared to experimental results (MAD and maximum error, Max, in kJ/mol).**

| | FEP without water | | FEP with water | |
|---|---|---|---|---|
| | **2WI7** | **3FT5** | **2WI7** | **3FT5** |
| MAD | 14.2 ± 1.0 | 5.3 ± 0.8 | 4.8 ± 1.3 | 3.7 ± 1.3 |
| $R$ | –0.81 ± 0.07 | 0.59 ± 0.10 | 1.0 ± 0.04 | 1.0 ± 0.04 |
| $\tau$ | –0.33 ± 0.33 | 0.33 ± 0.48 | 0.33 ± 0.43 | 0.33 ± 0.43 |
| Max | 26.0 ± 1.8 | 11.2 ± 1.8 | 6.1 ± 1.8 | 4.1 ± 1.8 |

## 6.2 Paper II

**Binding free energies in the SAMPL6 octa-acid host–guest challenge calculated with MM and QM methods**

In this paper, we have estimated free energies for the binding of eight carboxylate ligands to two variants of the octa-acid deep-cavity host (OAH and OAM) in the SAMPL6 blind-test challenge, using different methods: FEP at the MM level, FEP at the QM/MM level obtained with the reference-potential approach with QM/MM sampling (RPQS) at the PM6-DH+/MM level, as well as energies from QM/MM optimised structures at the PM6-DH+/MM and DFT/MM levels of theory.

The results were quite satisfying because for the first time we were able to improve MM-FEP results for the octa-acid host with QM/MM methods and the results were among the best five submissions to the competition. The RPQS method performed the best, with MADs of 2.4–5.0 kJ/mol, excellent correlation of 0.81–0.93, and Kendall's tau of 0.79–0.86 (Figure 6.1).

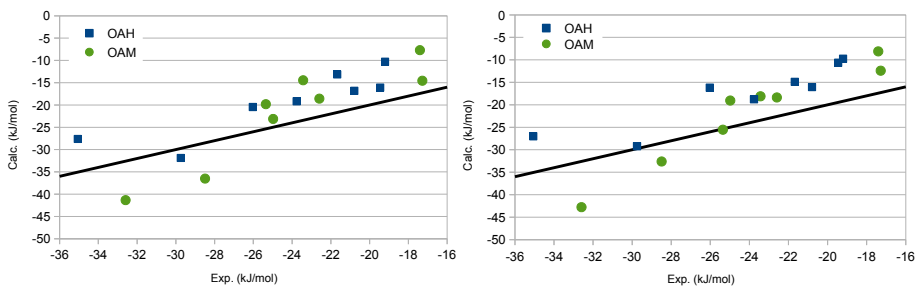



**Figure 6.1.** Comparison of the experimental and calculated absolute affinities obtained with the MM-FEP (left) and QM/MM-FEP (right) methods. The line shows the perfect correlation.

## 6.3 Paper III

**Comparison of the GCMC and GIST methods to determine the water structure in protein binding sites**

In this paper, we test the performance of two methods used to determine the water structure in protein–ligand binding sites: GIST and GCMC. We compare how well the predictions of the two approaches agree for two cases: a system with a buried binding site (ferritin) and a system with a solvent-exposed binding site (galectin-3).

Our results indicate that GCMC calculations can be recommended for buried binding sites, for which the equilibration of water molecules with bulk may be slow (Figure 6.2). However, for solvent-exposed sites, GCMC gives poor water densities and the results should always be compared to MD results and the Adams parameter should be selected to reproduce water densities observed in MD.

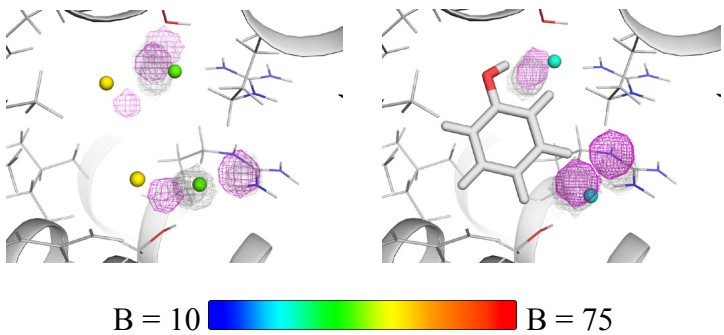


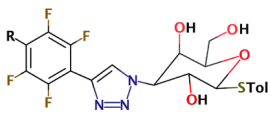

**Figure 6.2.** Comparison of MD densities (grey; 10 separate simulations) and GCMC densities (magenta) for apo (left) and phenol-bound (right) ferritin started with crystal water molecules included. The GCMC results are from the simulations giving N closest to Nopt, viz. that with B = –8 for apo and B = –11 for the simulation with phenol. Densities are shown for an isovalue of 0.6. Crystal water molecules are shown as spheres, coloured based on their B-factors.

## 6.4 Paper IV

**Substituted polyfluoroaryl interactions with an arginine side chain in galectin-3 are governed by steric-, desolvation and electronic conjugation effects**

In this paper, we studied the binding of a series of 2,3,5,6-tetrafluorophenyl derivatives with different para substituents (Figure 6.3, left) to galectin-3. The compound with fluorine as the para substituent showed the highest binding affinity and any replacement of the fluorine in the para position led to a drop in the affinity. This was assumed to be due to fluorine–amide interaction with the backbone amide of Ser237–Gly238. However, the QM interaction energy between the backbone of Ser237–Gly238 and this ligand was not larger than for some of the other ligands.

Instead, we showed that the relative affinities seem to be determined by other effects. First, the pocket beneath Arg144 is not large enough to fit bulkier groups (steric effects). Second, the solvation energy decreases strongly in the series OH–$NH_2$–$N_3$–F, implying that the desolvation penalty also decreases in this series, closely following the affinities of these ligands. Finally, we also showed that the solvation effect is partly counteracted by the interaction energy of the substituted tetrafluorophenyl group with Arg144, which becomes less favourable in this series.




| R | $K_d$ (μM) |
|---|---|
| OH | 23 ± 1.7 |
| F | 3.4 ± 0.21 |
| $N_3$ | 8.5 ± 1.2 |
| $NH_2$ | 11 ± 0.6 |
| OMe | 18 ± 2.1 |
| OEt | 88 ± 12 |
| NHMe | 18 ± 0.9 |
| $NMe_2$ | 40 ± 3.3 |
| Pyrrolidine | > 300 |

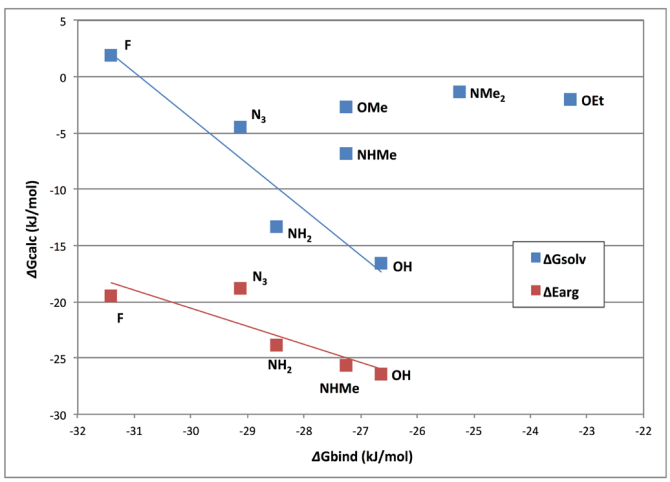



**Figure 6.3.** Left: Ligands used in this study and their $K_d$ values obtained by fluorescence polarisation assay. Right: Solvation free energies (blue squares) and interaction energies between Arg144 and the substituted tetraphenyl group (red squares).

## 6.5 Paper V

**Structure and energetics of ligand–fluorine interactions with galectin-3 backbone and side-chain amides – insight into solvation effects and multipolar interactions**

In this paper, we performed structural and theoretical analyses of galectin-3 ligands containing fluorinated phenyltriazolyl-thiogalactosides in order to study fluorine–amide interactions in the galectin-3 binding pocket and to attempt to correlate these with binding affinity as measured by fluorescence polarisation. We concluded that the binding of all ligands in this study is not governed by fluorine–amide interactions, but is instead determined by other effects, in particular dispersion, desolvation and polar interactions with other parts of the ligand. We show that the fluorine group is not more important than the hydrogen atoms on the benzene ring, but that at the same time, fluorine slightly but significantly decreases the solvation energy of the ligand which promotes the binding to a hydrophobic site (Figure 6.4). Thus, fluorine–amide interactions in protein–ligand interactions cannot simply be predicted on geometrical considerations alone but require careful consideration of the energetic components.

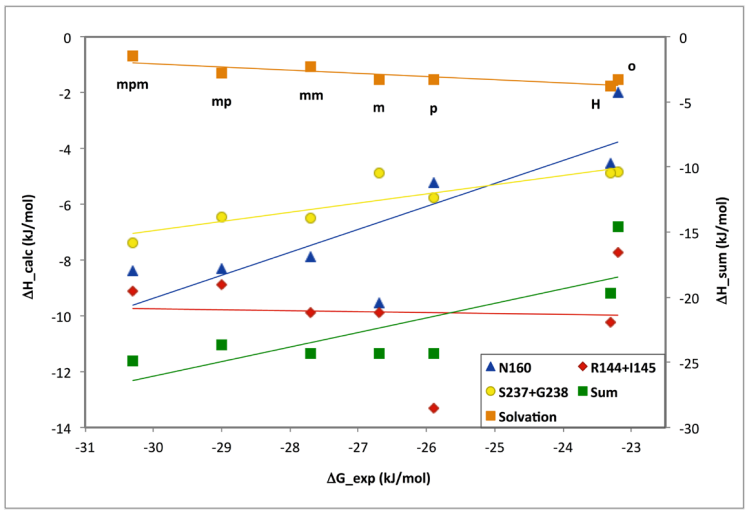



**Figure 6.4.** Calculated QM interaction and solvation energies for the seven substituted benzene groups plotted on the *y*-axis against the experimental binding free energy on the *x*-axis. Interaction energies were calculated for three amino-acid models: the sidechain of Asn160, the backbone of Arg144–Ile145 and the backbone of Ser237–Gly238. Sum is the sum of these three interaction energies minus the solvation energy of the ligand and it is shown on the right-hand-side *y*-axis. Best-fit lines are shown in the same colour as the symbols for each set of data.

## 6.6 Paper VI

**Structural and thermodynamic studies on halogen-bond interactions in ligand–galectin-3 complexes: electrostatics, solvation and entropy effects**

In this paper, we investigate the variation in binding affinity between galecin-3 and a systematically varied series of halogen-containing ligands, using experimental and theoretical methods such as isothermal titration calorimetry (ITC), competitive fluorescence polarization, X-ray crystallography, NMR spectroscopy, MD simulations, and QM calculations.

The QM calculations show that the binding enthalpy can be explained by interactions with Gly182 and other nearby residues, as well as the desolvation penalty (Figure 6.5, left). The change in entropy seems to be related to the number of water molecules displaced by the ligands: The ligand with H retains two water molecules, one of which is displaced by the ligands with F, Cl and Br, whereas the one with I displaces both molecules (Figure 6.5, right).

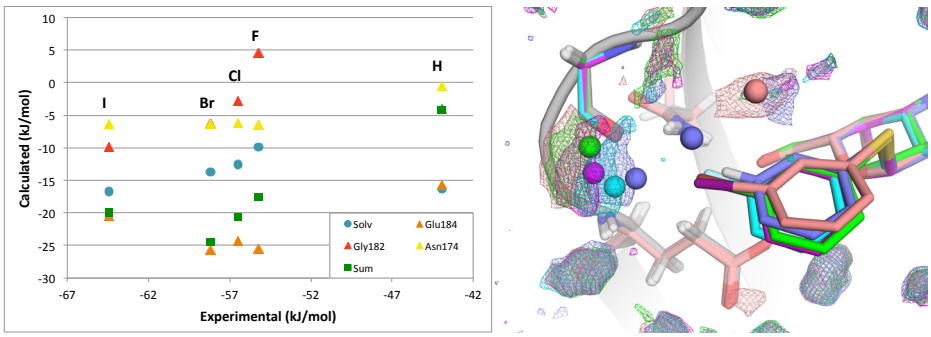



**Figure 6.5.** Left: Comparison of the experimental $\Delta H$ from ITC with the calculated solvation free energy and QM interaction energies between the ligands and three nearby residues for the five ligands H, F, Cl, Br, and I. The sum of the three interaction energies minus the solvation free energy (Sum) shows a fair correlation to the experimental binding enthalpy ($R$ = 0.86). Right: Superposition of the crystal structures and the water densities from the MD simulations for the five galectin-3C–ligand complexes, focused on the variable part of the ligands. The isodensity level is the same in all figures, five times the bulk density. The variable water molecule in the crystal structures are shown as balls. The structures and densities are color coded by H (slate), F (cyan), Cl (magenta), Br (green), and I (salmon pink).

## 6.7 Paper VII

**Interplay between conformational entropy and solvation entropy in protein–ligand binding**

In Paper III, we study conformational entropy and solvation entropy contributions to the difference in free energy of binding of a pair of diastereomeric ligands (called R and S) to galectin-3, using a combination of ITC, X-ray crystallography, NMR relaxation, and molecular dynamics simulations.

To study solvation thermodynamics, we applied GIST analysis, as implemented by Gilson and coworkers in the cpptraj module of Amber software.[41] For both ligands, we first performed clustering of the unrestrained MD simulations that were used for the conformational entropy calculations. Subsequently we performed MD simulations for each identified cluster, in which we kept the protein restrained toward the starting crystal structure, and the ligand toward the conformation that best represented the cluster, and we analysed these simulations with GIST (Figure 6.6).

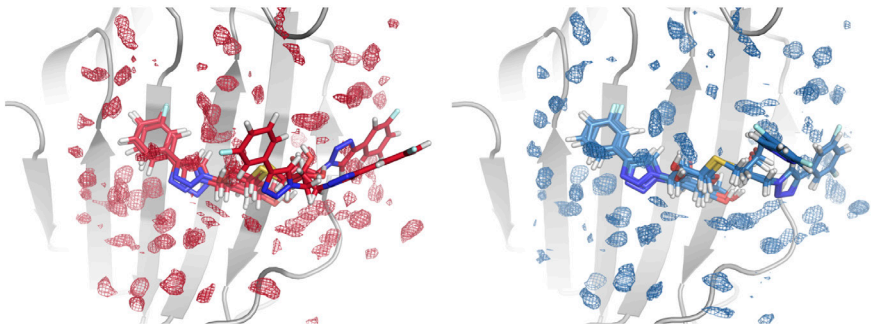


**Figure 6.6.** Differences in solvation around the binding site. Regions with high density of water relative to bulk water (six times the bulk water density) are represented as red mesh for R-galectin-3C (left) and blue mesh for S-galectin-3C (right).

We showed that the difference in solvent entropy of the two complexes amounts to only 3 ± 2 kJ/mol, whereas the difference in conformational entropy is 10 ± 5 kJ/mol, both in favour of the S-complex. The net contribution from conformational and solvent entropy of 13 ± 5 kJ/mol is greater than the overall entropy difference determined by ITC, 3 ± 1 kJ/mol (Figure 6.7), but the difference is not significant at the 95% confidence level. We conclude that conformational entropy dominates over solvation entropy in dictating the difference in the overall entropy of binding.

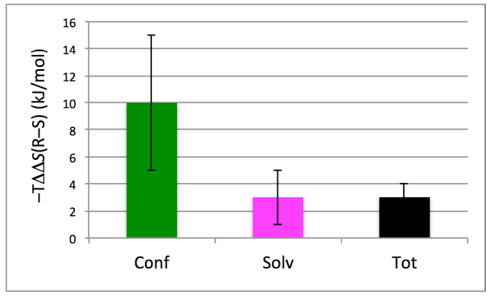



**Figure 6.7.** Entropy contributions to the differential binding of ligands R and S to galectin-3C. The bars indicate contributions from conformational entropy (green), solvation (magenta), and total entropy of binding determined by ITC (black). Error bars indicate the standard error (one standard deviation).

## 6.8 Paper VIII

**Entropy–Entropy compensation between the conformational and solvent degrees of freedom finetunes affinity in ligand binding to galectin-3C**

In this paper, we characterize the effects of minor changes in ligand structure on ligand affinity to the carbohydrate recognition domain of galectin-3. We employ a congeneric series of ligands with a fluorophenyl-triazole moiety, where the fluorine varied between the ortho, meta, and para positions. We used a combination of ITC, X-ray crystallography, NMR relaxation, and computational approaches including conformational entropy and GIST analyses of MD trajectories, to study how various entropic contributions to binding might vary between slightly different protein–ligand complexes.

Our results show that minor differences between protein–ligand complexes in their overall binding thermodynamics might encompass greater differences among individual contributions, including a case of entropy–entropy compensation between the protein conformational and solvent degrees of freedom.

# 7 Conclusions

In this thesis we employed several computational methods in order to investigate experimentally observed differences in binding affinity of various ligands binding to the galectin-3C protein. We studied the effects of solvation thermodynamics, protein and ligand conformational entropy change upon binding, as well as the significance of specific protein–ligand interactions, namely cation–π, fluorine–amide and halogen-bond interactions. Furthermore, we participated in two blind challenges, where we tested the performance of different free energy methods used to estimate protein–ligand binding affinities.

We show that, in order to better understand protein–ligand binding and to be able to accurately predict binding affinities, it is not enough to take into account only the contributions coming from protein–ligand interactions. In fact, it is equally important to consider the surrounding solvent, since it can in many ways contribute to the free energy of binding. For instance, introducing functional groups in the ligand that could form stronger interactions with the protein might lead to higher desolvation penalties that can significantly affect the binding affinities. Furthermore, we show that the dynamics of the solvent around the binding site, together with the conformational entropy of both protein and ligand, give crucial contributions to binding thermodynamics. Finally, we show that the predictions of relative binding affinities may improve if displaced water molecules are included in the free-energy perturbation calculations.

Finally, we compare different methods used to compare water structure and energetics, and we conclude that the GCMC simulations perform better for buried binding sites, whereas for solvent-exposed sites, MD simulations give more reliable results.

# 8 References


1. Patrick GL. *An Introduction to Medicinal Chemistry*.; 2013. doi:10.1017/CBO9781107415324.004.
2. Berg, Jeremy M, Tymoczko JL, Gatto GJ, Stryer L. *Biochemistry 8th Edition*.; 2015. doi:10.1007/978-3-8274-2989-6.
3. Du X, Li Y, Xia YL, et al. Insights into protein–ligand interactions: Mechanisms, models, and methods. *Int J Mol Sci*. 2016. doi:10.3390/ijms17020144.
4. Dunitz JD. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem Biol*. 1995. doi:10.1016/1074-5521(95)90097-7.
5. Klebe G. The foundations of protein–ligand interaction gerhard klebe. *Foundations*. 2009. doi:10.1007/978-90-481-2339-1_6.
6. Davis AM, Teague SJ. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew Chemie - Int Ed*. 1999. doi:10.1002/(SICI)1521-3773(19990315)38:6<736::AID-ANIE736>3.0.CO;2-R.
7. Dunitz JD. The entropic cost of bound water in crystals and biomolecules. *Science (80- )*. 1994. doi:10.1126/science.264.5159.670.
8. Snyder PW, Lockett MR, Moustakas DT, Whitesides GM. Is it the shape of the cavity, or the shape of the water in the cavity? *Eur Phys J Spec Top*. 2014. doi:10.1140/epjst/e2013-01818-y.
9. Ladbury JE. Just add water! The effect of water on the specificity of protein- ligand binding sites and its potential application to drug design. *Chem Biol*. 1996. doi:10.1016/S1074-5521(96)90164-7.
10. Ajit Varki, Richard D Cummings, Jeffrey D Esko, Hudson H Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W Hart and MEE, Varki A, Cummings RD, et al. *Essentials of Glycobiology, 2nd Edition*.; 2009. doi:10.1016/S0962-8924(00)01855-9.
11. Johannes L, Jacob R, Leffler H. Galectins at a glance. *J Cell Sci*. 2018;131:208884. doi:10.1242/jcs.208884.
12. Chou FC, Chen HY, Kuo CC, Sytwu HK. Role of galectins in tumors

and in clinical immunotherapy. *Int J Mol Sci*. 2018. doi:10.3390/ijms19020430.

13. Atkins PW, Friedman R. Molecular Quantum Mechanics, fifth edition. *Oxford Universiy Press*. 2011.

14. Born M, Oppenheimer R. Zur Quantentheorie der Molekeln. *Ann Phys*. 1927. doi:10.1002/andp.19273892002.

15. Hartree DR. The Wave Mechanics of an Atom with a Non-Coulomb Central Field Part I Theory and Methods. *Math Proc Cambridge Philos Soc*. 1928. doi:10.1017/S0305004100011919.

16. Boys SF. Electronic wave functions I. A general method of calculation for the stationary states of any molecular system. *Proc R Soc London*. 1950. doi:10.1098/rspa.1950.0036.

17. Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev*. 1964. doi:10.1103/PhysRev.136.B864.

18. Jensen F. *Introduction to Computational Chemistry*. 3rd ed. Chichester: John Wiley & Sons, Ltd; 2017.

19. Becke AD. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A*. 1988. doi:10.1103/PhysRevA.38.3098.

20. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*. 2015;11(8):3696-3713. doi:10.1021/acs.jctc.5b00255.

21. Wang J, Wang W, Kollman PA, Case DA. Development and testing of a general amber force field. *J Comput Chem*. 2004;25:1157-1174. doi:10.1002/jcc.20035.

22. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79(2):926-935. doi:10.1063/1.445869.

23. Horn HW, Swope WC, Pitera JW, et al. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys*. 2004;120(20):9665-9678. doi:10.1063/1.1683075.

24. Izadi S, Anandakrishnan R, Onufriev A V. Building water models: A different approach. *J Phys Chem Lett*. 2014. doi:10.1021/jz501780a.

25. Bayly CI, Cieplak P, Cornell WD, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem*. 1993;97(40):10269-10280. doi:10.1021/j100142a004.

26. Warshel A, Levitt M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol*. 1976. doi:10.1016/0022-2836(76)90311-9.

27. Newton I. *Philosophiae Naturalis Principia Mathematica*.; 2016. doi:10.5479/sil.52126.39088015628399.

28. Leach AR. *Molecular Modelling: Principles and Applications (2nd Edition)*.; 2001. doi:10.1016/S0097-8485(96)00029-0.

29. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*. 1977;23(3):327-341. doi:10.1016/0021-9991(77)90098-5.

30. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys*. 1993;98(12):10089. doi:10.1063/1.464397.

31. Hill TL, Gillis J. *An Introduction to Statistical Thermodynamics*. *Phys Today*. 1961;14(3):62-64. doi:10.1063/1.3057470.

32. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953. doi:10.1063/1.1699114.

33. Ross GA, Bodnarchuk MS, Essex JW. Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J Am Chem Soc*. 2015;137(47):14930-14943. doi:10.1021/jacs.5b07940.

34. Woods CJ, Michel J, Bodnarchuk MS, et al. ProtoMS 3.2. 2016.

35. Adams DJ. Chemical potential of hard-sphere fluids by Monte Carlo methods. *Mol Phys*. 1974;28:1241-1252. doi:10.1080/00268977400102551.

36. Edholm O, Berendsen HJC. Entropy estimation from simulations of non-diffusive systems. *Mol Phys*. 1984;51(4):1011-1028.

37. Trbovic N, Cho JH, Abel R, Friesner RA, Rance M, Palmer AG. Protein side-chain dynamics and residual conformational entropy. *J Am Chem Soc*. 2009;131(2):615-622. doi:10.1021/ja806475k.

38. Diehl C, Genheden S, Modig K, Ryde U, Akke M. Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J Biomol NMR*. 2009;45:157-169. doi:10.1007/s10858-009-9356-5.

39. Genheden S, Ryde U. Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys Chem Chem Phys*. 2012;14(24):8662-8677. doi:10.1039/c2cp23961b.

40. Lazaridis T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J Phys Chem B*. 2002. doi:10.1021/jp9723574.

41. Nguyen CN, Young TK, Gilson MK. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J Chem Phys*. 2012;137:044101. doi:10.1063/1.4733951.

42. Zwanzig RW. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J Chem Phys*. 1954;22(8):1420-1426. doi:10.1063/1.1740409.

43. Kirkwood JG. Statistical Mechanics of Fluid Mixtures. *J Chem Phys*. 1935;3(5):300.

44. Bennett CH. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J Comput Phys*. 1976;22:245-268.

45. Shirts MR, Bair E, Hooker G, Pande VS. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys Rev Lett*. 2003. doi:10.1103/PhysRevLett.91.140601.

46. Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys*. 2008;129:124105 (10 pages). doi:10.1063/1.2978177.

47. Ryde U, Söderhjelm P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem Rev*. 2016;116:5520-5566. doi:10.1021/acs.chemrev.5b00630.

48. Rathore RS, Reddy RN, Kondapi AK, Reddanna P, Reddy MR. Use of quantum mechanics/molecular mechanics-based FEP method for calculating relative binding affinities of FBPase inhibitors for type-2 diabetes. *Theor Chem Acc*. 2012;131(2):1096; 10 pages. doi:10.1007/s00214-012-1096-z.

49. Reddy MR, Erion MD. Relative binding affinities of fructose-1,6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method. *J Am Chem Soc*. 2007;129(30):9296-9297. doi:10.1021/ja072905j.

50. Muller RP, Warshel A. Ab initio calculations of free energy barriers for chemical reactions in solution. *J Phys Chem*. 1995. doi:10.1021/j100049a009.

51. Rod TH, Ryde U. Quantum mechanical free energy barrier for an enzymatic reaction. *Phys Rev Lett*. 2005;94:138302 (4 pages). doi:10.1103/PhysRevLett.94.138302.

52. Olsson MA, Ryde U. Comparison of QM/MM Methods To Obtain

Ligand-Binding Free Energies. *J Chem Theory Comput*. 2017;13:2245-2253. doi:10.1021/acs.jctc.6b01217.

53. Steinmann C, Olsson MA, Ryde U. Relative ligand-binding free energies calculated from multiple short QM/MM MD simulations. *J Chem Theory Comput*. 2018;14:3228-3237.

# 9 Acknowledgements

I would like to begin by thanking **Ulf Ryde**, my supervisor, for all help and support over these four years. You are patient, positive, and always available. I am glad I had you as my supervisor.

I am very thankful to all former and current members of the Ryde group: **Paulius**, for your help and tips in the first weeks of my PhD studies. **Francesco**, for your humour and personality, both which I miss very much. **Martin**, for helping me learn free-energy methods. **Geng**, for stories about China and all the plants in our office. **Octav**, for introducing me to so many board games, as well as for the coins you remember to bring me from your travels. I really appreciate that. **Lili**, for making tasty Chinese dishes for the group, and for showing up in my office and making me laugh every day. **Erik**, for encouraging me to learn Dan… Swedish. Do not give up on me, one day jag ska tala svenska med dig! **Justin**, for many interesting chats and for sharing cake with me every time you bake something nice. I also thank **Pär** and **Esko**, for joining our Journal Clubs. I learnt a lot from you. I want to thank other members and all the visitors. It was nice to meet you all and do research with some of you.

To people involved in the DECREC project: **Mikael**, who is also my co-supervisor, **Hakon**, **Ulf N.**, **Derek**, **Maria Luisa**, **Kristoffer**, **Olof**, **Johan**, and **Rohit**, thank you very much for your efforts in experiments you performed with galectin-3 and its ligands, and for interesting and fun meetings over the years.

I want to say thank you to everyone in the **Teokem** division, for providing a pleasant working environment. I really enjoyed our fika and lunch times, the seminars, the courses, the parties, the Thursday evenings. I am very grateful to everyone who was a part of the teokem/fyskem coffee-room often full of people from so many different countries. Thank you all for bringing a piece of your culture with you and making me a part of the world every day.

To the Misini family: my parents **Suzana** and **Zaim**, my siblings **Sandra** and **David**, and my nephew **Leonardo**, thank you for loving and supporting me unconditionally. To my in-laws: **Valerija**, **Vladan** and **Sanja**, thank you for making me feel like I was always a part of your family.

Finally, **Boris**, my partner in life, thank you for the cover photo!

# Paper I

J Comput Aided Mol Des (2016) 30:707–730
DOI 10.1007/s10822-016-9942-z

# Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations

**Majda Misini Ignjatović[1] · Octav Caldararu[1] · Geng Dong[1] · Camila Muñoz-Gutierrez[2] · Francisco Adasme-Carreño[2] · Ulf Ryde[1]**

Received: 2 June 2016 / Accepted: 17 August 2016 / Published online: 26 August 2016


**Abstract** We have estimated the binding affinity of three sets of ligands of the heat-shock protein 90 in the D3R grand challenge blind test competition. We have employed four different methods, based on five different crystal structures: first, we docked the ligands to the proteins with induced-fit docking with the Glide software and calculated binding affinities with three energy functions. Second, the docked structures were minimised in a continuum solvent and binding affinities were calculated with the MM/GBSA method (molecular mechanics combined with generalised Born and solvent-accessible surface area solvation). Third, the docked structures were re-optimised by combined quantum mechanics and molecular mechanics (QM/MM) calculations. Then, interaction energies were calculated with quantum mechanical calculations employing 970–1160 atoms in a continuum solvent, combined with energy corrections for dispersion, zero-point energy and entropy, ligand distortion, ligand solvation, and an increase of the basis set to quadruple-zeta quality. Fourth, relative binding affinities were estimated by free-energy simulations, using the multi-state Bennett acceptance-ratio approach. Unfortunately, the results were varying and rather poor, with only one calculation giving a correlation to the experimental affinities larger than 0.7, and with no consistent difference in the quality of the predictions from the various methods. For one set of ligands, the results could be strongly improved (after experimental data were revealed) if it was recognised that one of the ligands displaced one or two water molecules. For the other two sets, the problem is probably that the ligands bind in different modes than in the crystal structures employed or that the conformation of the ligand-binding site or the whole protein changes.

**Keywords** Ligand-binding affinity · Induced-fit docking · MM/GBSA · QM/MM · Big-QM · Free-energy perturbation · Continuum solvation · Bennett acceptance ratio · D3R grand challenge · Blind-test competition

## Introduction

One of the prime challenges of computational chemistry is to predict the free energy for the binding of small molecules to biomacromolecules. Many biological functions are exerted by the binding of substrates or inhibitors to enzymes or effectors to receptors, and the prime aim of drug development is to find small molecules that bind strongly to the target receptor, but with a small effect on other biosystems. Consequently, much effort has been spent to develop methods with this aim, ranging from simple docking and scoring approaches, via end-point

Majda Misini Ignjatović, Octav Caldararu, Geng Dong, Camila Muñoz-Gutierrez and Francisco Adasme-Carreño have contributed approximately equal to the investigation: MMI performed the FES simulations of sets 1 and 3, as well as the GCMC calculations; OC performed the FES calculations on set 2; GD performed the QM/MM calculations; CMG and FAD performed the docking and MM/GBSA calculations.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-016-9942-z) contains supplementary material, which is available to authorized users.

✉ Ulf Ryde
Ulf.Ryde@teokem.lu.se

[1] Department of Theoretical Chemistry, Lund University, Chemical Centre, P. O. Box 124, 221 00 Lund, Sweden

[2] Centro de Bioinformática y Simulación Molecular, Facultad de Ingeniería, Universidad de Talca, 2 Norte 685, Talca, Chile

methods, such as MM/GBSA (molecular mechanics combined with generalised Born and solvent-accessible surface area solvation) and linear interaction energies (LIE), to strict free-energy simulation (FES) methods [1–4].

Numerous studies have evaluated the performance of various binding-affinity methods, e.g. docking [5, 6], MM/GBSA [7, 8], and FES methods [9–11]. The conclusion has typically been that docking methods can rapidly find the correct binding pose among several other poses, but that they have problems to correctly rank the affinities of a set of ligands to the same protein. MM/GBSA calculations typically give a better ranking of the ligands and an understanding of energy terms involved in the binding, but often vastly overestimate energy differences and the results strongly depend on the employed continuum-solvation model [2, 12]. Large-scale tests of FES calculations have given rather impressive results for relative binding affinities of similar ligands to the same protein, with mean absolute deviations (MAD) of 4–6 kJ/mol [9–11]. However, the comparisons have been primarily directed to small changes in the ligands and the performance is uneven, with very good results for some proteins, but quite poor performance for other proteins, occasionally with errors of over 20 kJ/mol.

Comparisons of different approaches for the same test case are less common and often half-hearted in the meaning that the authors are experts or developers of one approach and include other methods mainly to show that they are worse [10, 13, 14]. In this respect, blind-test competitions are important to judge the true performance of different approaches, allowing experts to provide predictions that are not biased by the experimental results. In the SAMPL4 octa-acid host–guest challenge for binding affinities, FES methods gave the best results (the root-mean-squared deviation, RMSD, was 5 kJ/mol and the correlation coefficient, $R^2$, was 0.9), although docking gave results of only slightly worse quality (RMSD = 6 kJ/mol, $R^2 = 0.8$) [15–17]. However, this test case was ideal for FES calculations with quite small differences between the ligand and a conserved net charge. For the cucurbit [7] uril host, the results were worse and more varying, but a FES-based approach still gave the best results RMSD = 12 kJ/mol, $R^2 = 0.8$, whereas docking gave poor results (RMSD = 33 kJ/mol, $R^2 = 0.1$) [15, 17]. The results for the SAMPL3 host–guest systems were even worse, with either RMSD and $R^2$ both low, e.g. 6 kJ/mol and 0.4 for the MM/GBSA-like solvent interaction energy (SIE) approach [18], or both high, e.g. 47 kJ/mol and 0.8 for FES [19].

For protein systems, the results have been even worse. For the HIV integrase binding-affinity challenge in SAMPL4, a SIE approach was pointed out as best with a mean absolute deviation (MAD) of 5 kJ/mol, but it gave a negative correlation ($R = -0.3$) [20, 21]. Docking calculations gave positive correlation ($R = 0.5$–$0.6$), but the MAD was high (76–113 kJ/mol), because a raw docking score was employed [22]. An MM/PBSA approach gave a lower MAD, 16 kJ/mol, and a positive correlation ($R = 0.4$) [20]. The reason for these poor results was that all eight experimental binding affinities were within 4 kJ/mol.

A similar problem applied to the trypsin challenge in SAMPL3, where the experimental range of the 17 ligands was only 9 kJ/mol (and 13 within 4 kJ/mol). Unfortunately, no overview article was published for this test case, so it is hard to reach any unbiased conclusions. A comparison of five methods indicated that none of them gave any useful correlation ($R^2 < 0.02$), but LIE gave a correct ranking of all ligands for which both the experimental and computational estimates were statistically significant [14]. Docking with the Glide software gave the lowest MAD (3 kJ/mol) and also the best discrimination between binders and non-binders (the area under the receiver-operating-characteristic curve, AUC, was 0.8). LIE gave a slightly larger MAD (4 kJ/mol), but a poorer-than-random AUC (0.3). MM/PBSA and MM/GBSA gave large MAD (20 and 16 kJ/mol), but reasonable AUC (0.7).

In this article, we present a comparison of four different approaches to calculate absolute or relative binding affinities for three sets of similar ligands to the heat-shock protein 90 (HSP90) within the drug-design data resource (D3R) 2015 grand challenge [23]. HSP90 is a conserved chaperone protein that is expressed ubiquitously in high concentration [24], in particular in cancer cells [25, 26] and therefore of large interest as a multiple-oncogenic-pathway therapeutics [27–30]. We have performed docking with the Glide software [31], MM/GBSA scoring with single minimised structures with the Prime software [32], and FES calculations of relative affinities. In addition, we have made an attempt to perform combined quantum and molecular mechanics (QM/MM) scoring with an approach similar to that developed by Grimme and coworkers for host–guest systems [33, 34] combined with our big-QM approach to obtain stable QM/MM energies for proteins [35].

## Methods

Relative binding free energies for three sets of ligands binding to HSP90 were estimated as a part of the D3R Grand Challenge 2015 [23]. Sets 1, 2, and 3 consist of five, four, and ten ligands, respectively and involve chemically similar ligands, which allow for the calculation of relative binding free energies by alchemical FES methods. The 19 ligands are shown in Fig. 1. The FES calculations employed four additional reference ligands, which are also

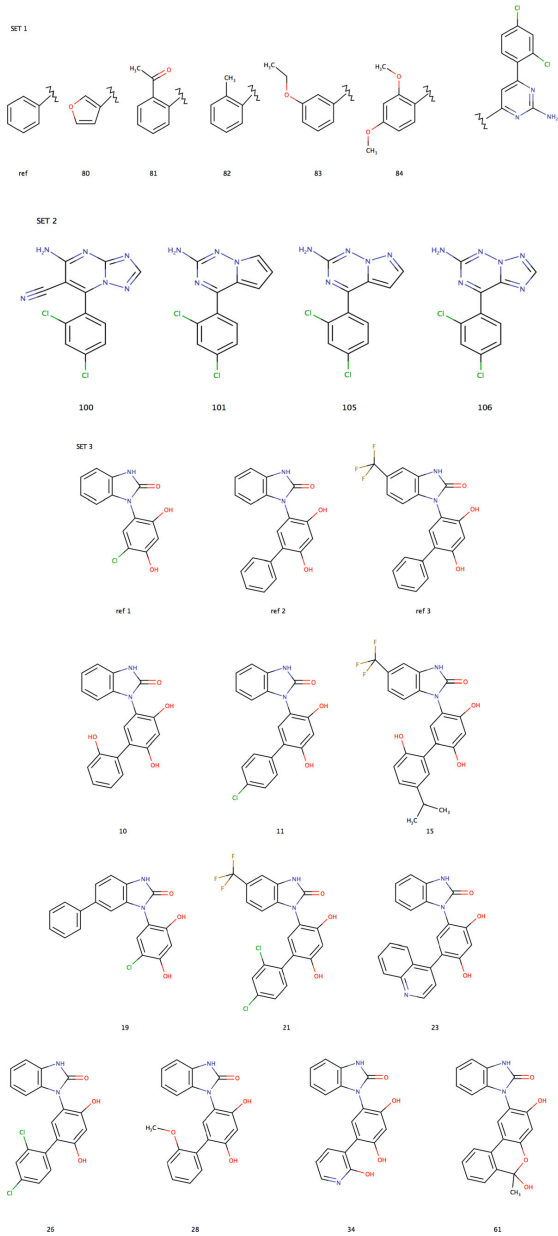



**Fig. 1** Structures of all ligands from sets 1, 2, and 3, considered in this study. The additional reference ligands that were employed for sets 1 and 3 are also shown. The numbering of ligands is the same as in the HSP90 D3R grand challenge data set. Ligands of sets 1 and 3 are shown in conformation 1

shown in the figure. Four methods were used to estimate the binding affinities, viz. docking, MM/GBSA, QM/MM, and FES. They are described in separate sections below.

The studies were based on five protein crystal structures (PDB files 3VHA [36], 2WI7 [37], 3FT5 [38], 3OW6 [39], and 4YKR [40]), which are described in Table 1. They were selected based on the quality of the structure, the conformation of the entrance of the ligand-binding pocket (closed, semi-closed, or open [38]) and the similarity of the co-crystallised ligand with the ligands in the various sets. The ligands in the crystal structures are shown in Figure S1 in the supplementary material. The 3VHA structure was obtained at 1.4 Å resolution and it contains a ligand that is quite similar to those in set 1. It was the only structure used for the set 1 calculations and it was also used for some set 2 calculations. However, the ligand in 2WI7 is more similar to the set 2 ligands, although the resolution is rather poor, 2.5 Å. The ligand in 3FT5 is also similar to the set 2 ligands, but it is much smaller and the binding pocket is in the closed conformation. The resolution is intermediate (1.9 Å). For set 3, two structures were employed, 3OW6 and 4YKR. They are of similar resolution (1.8 and 1.6 Å, respectively) and contain similar ligands of a proper scaffold (the ligand is slightly smaller in the 3OW6 structure).

## Docking calculations

The docking calculations were set up with the Schrödinger 2015-2 suite of software [41]. They were based on the 3VHA [36] structure for set 1 and 2, and the 3OW6 [39] structure for set 3. The 4YKR [40] structure was also tested for set 3, but no reasonable docked structures could be obtained for ligands **15** and **61**. After the experimental results were revealed, docking calculations were also performed with the 2WI7 crystal structures for set 2 [37]. The protein preparation wizard module was employed for preparing the protein structures [41]. Crystal water molecules more than 5 Å away from the ligand were removed prior to the hydrogen-bond optimisation and protein minimisation stages. The hydrogen-bond network was optimised at pH 7 by sampling Asn and Gln rotamers, hydroxyls, thiols, and water orientations. The protonation states for Asp, Glu, and His were derived from PropKa 3.1 [42, 43]. The protonation states employed for the His residues are shown in Table 1.

According to the recommended protein preparation protocol [44], the prepared structures were then relaxed by means of a restrained molecular minimisation using the Impact refinement module using the OPLS 2005 force field [45], with heavy atoms restrained to remain within a RMSD of 0.30 Å from the initial coordinates. This allows hydrogen atoms to be freely minimised and heavy atoms can move to relax strained bonds, angles, and steric clashes. After a closer inspection of the hydrogen-bond network in the ligand-binding site, three (3OW6) or four (3VHA and 2WI7) water molecules were identified that form at least one hydrogen bond to either the protein or the ligand. These water molecules were kept in the calculations, whereas the remaining crystal water molecules were deleted. For set 2, one of the four crystal-water molecules (called Wat2 below) made steric clashes with one of the ligands. In the calculations with the 3VHA structure, this water molecule was deleted when docking all four ligands, whereas with the 2WI7 structure, Wat2 was deleted only for ligand **100** and was kept for the other three ligands.

The ligand structures were built using the Maestro visualisation software [46] and then prepared with the LigPrep module [47], in which the ionisation and tautomeric states at pH 7 were predicted using Epik [48]. Finally, an energy minimisation in gas phase using Macromodel [49] with the OPLS 2005 force field [45] was performed.

All docking calculations were performed using the Glide software [31]. Initial docking studies using the standard-precision (SP) mode with default parameters for grid and pose generation were unable to produce poses that fitted into the binding site for the tested inhibitors, probably because the binding cavity is too tight to fit molecules larger than the co-crystallised ligands. Scaling down the van der Waals radii of non-polar protein atoms, a crude approach to allow steric clashes during docking, did not produce better results. Therefore, we employed the

**Table 1** Description of the protein structures used in this study and protonation states of the His residues

| Crystal structure | Resolution (Å) | State | His protonation | | | | Set | Ref. |
|---|---|---|---|---|---|---|---|---|
| | | | 77 | 154 | 189 | 210 | | |
| 3VHA | 1.39 | Semi-closed | HIP | HIP[a] | HIP | HIE | 1, 2 | [36] |
| 2WI7 | 2.50 | Open | HIP | HIE | HIP | HIE | 2 | [37] |
| 3FT5 | 1.90 | Closed | HIP | HIE | HIP | HIE | 2 | [38] |
| 3OW6 | 1.80 | Semi-closed | HIP | HID | HIP | HIE | 3 | [39] |
| 4YKR | 1.61 | Closed | HIP | HIE | HIP | HIE | 3 | [40] |

[a] HID in the docking and QM/MM calculations

induced-fit docking (IFD) workflow [50, 51] to generate alternative conformations of the receptor suitable to bind the studied ligands, by allowing the protein to undergo sidechain or backbone movements during the docking.

The IFD procedure has four steps: (1) initial Glide docking using a softened-potential (van der Waals scaling of 0.5) into a rigid receptor to generate an ensemble of poses; (2) sampling of protein conformations using the sidechain prediction module Prime [32], followed by a structure minimisation of each protein–ligand complex; (3) redocking of the ligands into low energy induced-fit structures from the previous step using default Glide settings (no scaling of van der Waals interactions); and (4) estimation of the binding energy of the optimised protein–ligand complexes.

The IFD standard protocol was employed, generating up to 20 poses per ligand on each iteration. The docking grid was generated for the co-crystallised ligands. The OPLS 2005 force field [45] was used for the minimisation stage, in which residues within 5 Å of each ligand pose were optimised. Pose rescoring was performed with the SP docking mode. All other parameters were set to their default values. Finally, the obtained docking poses were visually inspected, filtering out those that did not adopt a similar position and orientation as the reference inhibitors. Only the most favourable docking pose for each ligand was selected for structural analysis.

### Pose rescoring with MM/GBSA

All docking poses were rescored with the MM/GBSA approach, as implemented in the Prime program in the Schrödinger software suite [32, 41]. It employed a single minimised protein–ligand structure, thus establishing an efficient approach to rapidly refine and rescore docking results. We employed the variable dielectric solvent model VSGB 2.0 [52], which includes empirical corrections for modelling directionality of hydrogen-bond and π-stacking interactions. This approach has been shown to give good binding free energies for a wide range of protein–ligand complexes [53]. Residues within 5.0 Å of the ligand were allowed to relax during the MM minimisation of the complex, keeping the rest of the structure fixed.

### QM/MM scoring

The docked structures were also rescored using a QM/MM approach, developed as a combination of the QM-cluster approach for the study of the binding in host–guest systems by Grimme and coworkers [33, 34] and the big-QM approach developed in our group to obtain stable QM/MM energies in proteins [35]. The QM/MM calculations employed the docked structures, but the first four residues in the protein for sets 1 and 2 were deleted (Pro11–Glu14, because they are hanging free in solution, without any interactions with the remainder of the protein) and a MOPS buffer molecule, far from the ligand-binding site, was also deleted. The docked structure was solvated in a sphere of water molecules with a radius of 37 Å, centred on the geometric centre of the protein, giving a total of ~18,600 atoms. Hydrogen atoms and water molecules were optimised with a 120 ps simulated annealing calculation with an initial temperature of 370 K, followed by a minimisation using the Amber software [54].

#### *QM/MM calculations*

The QM/MM calculations were performed with the ComQum software [55, 56]. In this approach, the protein and solvent are split into two subsystems: System 1 (the QM system) was relaxed by QM methods. For sets 1 and 2, it consisted of the ligand, as well as Asn51, Ser52, Asp54, Ala55, Lys58, Asp93, Gly95, Ile96, Gly97, Met98, Asp102, Asn106, Leu107, Phe138, Tyr139, Val150, Thr152, His154, Thr184, and Val186. For set 3, the QM system included residues Leu48, Ile49, Asn51, Ser52, Asp54, Ala55, Lys58, Asp93, Ile96, Gly97, Met98, Asn106, Leu107, Lys112, Gly135, Val136, Gly137, Phe138, Tyr139, Val148, Val150, Thr152, Thr184, and Val186. In both cases, the six water molecules closest to the ligand were also included, giving a total of ~280 and ~320 atoms, respectively. The two QM systems are shown in Fig. 2a, b. System 2 consisted of the remaining part of the protein and the solvent. It was kept fixed at the original docked coordinates.

In the QM calculation, System 1 was represented by a wavefunction, whereas all the other atoms were represented by an array of partial point charges, one for each atom, taken from MM libraries. Thereby, the polarisation of the QM system by the surroundings is included in a self-consistent manner (electrostatic embedding). When there is a bond between systems 1 and 2 (a junction), the hydrogen link-atom approach was employed: the QM system was capped with hydrogen atoms (hydrogen link atoms, HL), the positions of which are linearly related to the corresponding carbon atoms (carbon link atoms, CL) in the full system [55, 57]. All atoms were included in the point-charge model, except the CL atoms [58].

The total QM/MM energy in ComQum is calculated from [55, 56]

$$E_{\mathrm{QM/MM}} = E^{\mathrm{HL}}_{\mathrm{QM1+ptch2}} + E^{\mathrm{CL}}_{\mathrm{MM12},q_i=0} - E^{\mathrm{HL}}_{\mathrm{MM1},q_i=0} \quad (1)$$

where $E^{\mathrm{HL}}_{\mathrm{QM1+ptch2}}$ is the QM energy of the QM system truncated by HL atoms and embedded in the set of point charges modelling system 2 (but excluding the self-energy

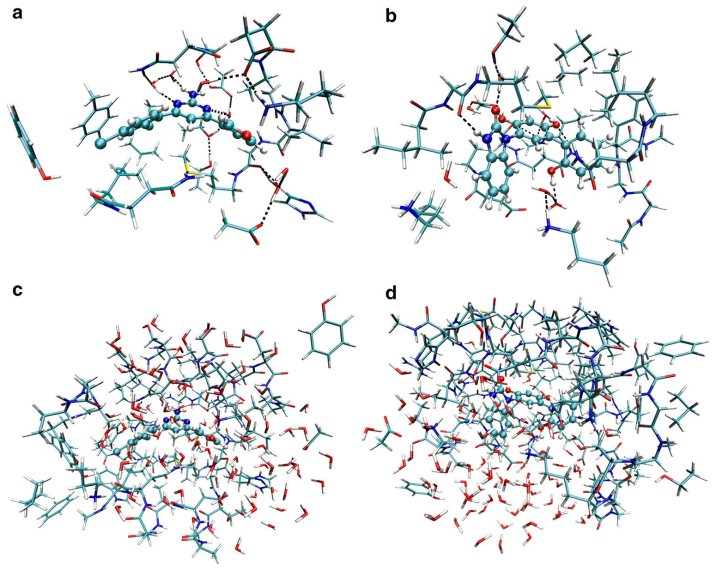



**Fig. 2** The QM systems used in the QM/MM optimisations for sets 1 and 2 (**a**), and set 3 (**b**), as well as in the big-QM calculations (**c**, **d**). The ligand is shown in *ball*-and-*sticks* representation

of the point charges). $E^{HL}_{MM1,q1=0}$ is the MM energy of the QM system, still truncated by HL atoms, but without any electrostatic interactions. Finally, $E^{CL}_{MM12,q1=0}$ is the classical energy of all atoms in the system with CL atoms and with the charges of the QM system set to zero (to avoid double counting of the electrostatic interactions). By this approach, which is similar to the one used in the ONIOM method [59], errors caused by the truncation of the QM system should cancel.

The geometry optimisations were continued until the energy change between two iterations was less than 2.6 J/mol ($10^{-6}$ a.u.) and the maximum norm of the Cartesian gradients was below $10^{-3}$ a.u. The QM calculations were carried out using Turbomole 7.0 software [60]. The geometry optimisations were performed using the TPSS [61] functional in combination with def2-SV(P) [62] basis set, including empirical dispersion corrections with the DFT-D3 approach [63]. The MM calculations were performed with the Amber software [54], using the Amber ff14SB force field [64].

### *Big-QM calculations*

Previous studies have shown that QM/MM energies strongly depend on the size of the studied QM system [58, 65]. To avoid this problem, we have developed the big-QM approach to obtain converged energies [35]: we constructed a very large QM system, consisting of all residues with at least one atom within 7.5 Å of the ligand in any of the studied structures. Thus, the QM system was the same for all ligands. For sets 1 and 2 residues 22, 26, 47–59, 61, 62, 78, 91–108, 112, 135–139, 141, 142, 148–155, 162, 180, and 182–187, as well as the 79 closest water molecules were included, in total ~970 atoms. For the set 3 ligands, the QM system consisted of residues 22, 26, 29, 44, 45, 47–59, 61, 62, 77, 78, 90–99, 102–113, 115, 131–142, 148–155, 162, 180, and 182–188, as well as the 80 closest water molecules, in total ~1160 atoms. Both systems included the single buried charged group in the protein, Asp93. The ligand is not covalently connected to the protein, so it does not form any junction to the protein (in the standard big-QM approach, all buried charges in the protein should be included and junctions should be moved two residues away from the minimal QM system [35]). The QM systems are shown in Fig. 2c, d. The big-QM calculations were performed on coordinates from the QM/MM optimisation. Two sets of big-QM calculations were performed. In the first, a point-charge model of the surroundings was included, because this gave the fastest calculations in our previous tests [35]. In the second approach, we performed the calculation without the point-

charge model, but included instead a conductor-like screening model (COSMO) [66, 67] continuum solvent with a dielectric constant of 80. In both cases, the calculations were performed at the TPSS/def2-SV(P) level of theory and they employed the multipole-accelerated resolution-of-identity J approach [68].

*Additional energy terms*

To the big-QM energy, we added the DFT-D3 dispersion correction, calculated for the same big-QM system with Becke–Johnson damping [69], third-order terms, and default parameters for the TPSS functional using dftd3 program [70].

Moreover, we added a correction for increasing the basis set from def2-SV(P) to def2-QZVP [71], calculated for the QM system used in the QM/MM geometry optimisations with the TPSS method and including a point-charge model of the surroundings:

$$\Delta E_{\text{bsc}} = E(\text{TPSS/def2-QZVP}) - E(\text{TPSS/def2-SV(P)}) \quad (2)$$

Thermal corrections to the Gibbs free energy at 298 K and 1 atm pressure ($G_{\text{therm}}$; including zero-point vibrational energy (ZPE) entropy, and enthalpy corrections) were calculated by an ideal-gas rigid-rotor harmonic-oscillator approach [72] from vibrational frequencies calculated at the MM level. These were obtained for truncated systems in which only residues and water molecules within 12 Å of the ligand were included in the calculations. Moreover, residues and water molecules more than 8 Å from the ligand were kept fixed in the calculations and they were ignored when the frequencies were calculated. Such an approach is employed in MM/PBSA calculations [73] and it has been found to give reliable results [74]. To obtain more stable results, low-lying vibrational modes were treated by the free-rotor approximation, using the interpolation model suggested by Grimme and $\omega_0 = 100\ \text{cm}^{-1}$ [33].

For all energy terms, interaction energies were calculated, i.e. separate calculations were performed for the complex, for the protein without the ligand, and for the isolated ligand:

$$\Delta E_{\text{int}} = E(\text{complex}) - E(\text{protein}) - E(\text{ligand}) \quad (3)$$

The protein calculations were always done using the geometry of the complex after removal of the ligand. For the free ligand, we did two sets of calculations. The first was single-point calculations on the QM/MM structures of the complex, whereas in the second approach, we optimised the geometry of the ligand at the TPSS/def2-SV(P) level of theory in a COSMO continuum solvent with a dielectric constant of 80. This allowed for the calculation of the relaxation energy of the ligand (i.e. the difference in the TPSS/def2-QZVP energy of ligand when optimised in the complex or isolated in the COSMO solvent).

Several approaches were tested to calculate the solvation energy of the complex. In particular, we tested the QM/MM-PBSA and -GBSA approaches [75], using Poisson–Boltzmann (PB) or generalised Born (GB) solvation energies of the whole protein–ligand complex after removal of the water molecules. However, this gave strongly varying energies with large differences between the PB and GB results. Therefore, we decided to simply use big-QM calculations performed in a COSMO solvent with a dielectric constant of 80. Such calculations were performed on both the complex and the protein without the ligand. More accurate solvation energies of the ligand (including also non-polar effects) were calculated with the COSMO-RS (real solvent) approach [76, 77] using the COSMOTHERM software [78]. These calculations were based on two single-point QM calculations at the BP/TZVP level of theory, either in vacuum and with an infinite dielectric constant.

Consequently, the final binding free energies involved six energy terms: the big-QM energies in the COSMO solvent, the basis-set correction, the DFT-D3 dispersion energy, the $\Delta G_{\text{therm}}$ free-energy corrections, the relaxation energy of the ligand, and the solvation free-energy correction for the ligand:

$$\Delta G_{\text{bind}} = \Delta G_{\text{BQ}} + \Delta E_{\text{bsc}} + \Delta E_{\text{disp}} + \Delta G_{\text{therm}} + \Delta E_{\text{L,rlx}} + \Delta\Delta G_{\text{L,solv}} \quad (4)$$

## FES calculations

Relative binding free energies were also estimated by FES calculations. These were set up independently, using slightly different methods. For set 1, the 3VHA structure was used [36], whereas for set 2, two crystal structures were employed: 2WI7 and 3FT5 [37, 38]. The ligand pose in 3FT5 is rotated 180° around C–$NH_2$ bond relative to that in 2WI7. We also tried to start the simulations from the protein structure of 3FT5, but with the ligand in the orientation found in structure 2WI7 (3FT5/2WI7). For set 3, the 4YKR structure was used [40]. The structures were protonated using the leap module of Amber 14 [54]. The protonation of His residues was determined by investigating the surroundings, the hydrogen-bond network and the solvent accessibility of each residue (Table 1). The assignment agreed for three of the His residues in all structures. However, for His154, we used a varying assignment, because the crystal structures show that the $N^{\delta 1}$ atom interacts either with the backbone O atom of

Asn155 or the backbone N atom of Asp156. In the 3VHA structure this residue is solvent exposed and forms a water-bridged interaction with Glu-62 and it was therefore assumed to be doubly protonated to reduce the net negative charge of the protein. All Glu and Asp residues were assumed to be negatively charged and all Lys and Arg residues positively charged, whereas the other residues were neutral. This assignment was checked by the PropKa software [42, 43].

All crystal-water molecules were kept in the calculations, except in set 2, for which one water molecule was deleted to avoid steric clashes with the cyano group in ligand **100**. However, after submission of the results, we run additional calculations with set 2, keeping all crystal-water molecules or deleting one (3FT5) or two (2WI7) water molecules by FES before the **101** → **100** perturbation. The protein–ligand complex and the free ligand were solvated in a truncated octahedral box of TIP3P water molecules [79], extending 10 Å from the protein and the ligand, respectively.

The proteins were described with the Amber14SB force field [64] and no counter ions were added to the system. All ligands were manually built into the corresponding protein structure and were described with general Amber force field [80]. Charges were obtained with the restrained electrostatic potential method [81]: the ligands were optimised with the semiempirical AM1 method, followed by a single-point calculation at the Hartree–Fock/6-31G* level to obtain the electrostatic potentials, sampled with the Merz–Kollman scheme [82]. These calculations were performed with the Gaussian 09 software [83]. The potentials were then used by antechamber to calculate the charges. A few missing parameters were obtained with the Seminario approach [84]: the geometry of the ligands was optimised at TPSS/def2-SV(P) level, followed by a frequency calculation using aoforce module of Turbomole 7.01 [60]. From the resulting Hessian matrix, parameters for the missing angles and dihedrals were extracted with the Hess2FF program [85]. These parameters are given in Tables S1 and S2 in the supplementary material.

After submission of the results, it was discovered that the structures of the set 1 ligands were strange, with a tetrahedral $-NH_2$ group, accepting hydrogen bonds from the protein and water molecules (Figure S2 in the supplementary material). This was traced back to a missing improper torsion for this group. By adding this torsion with a force constant of 10 kcal/mol/rad2 (cf. Table S2), more reasonable structures were obtained.

In order to estimate the relative binding free energy between two ligands, $L_1$ and $L_2$, $\Delta\Delta G^\circ_{\text{bind}} = \Delta G^\circ_{\text{bind}}(L_2) - \Delta G^\circ_{\text{bind}}(L_1)$, we employed a thermodynamic cycle that relates $\Delta\Delta G^\circ_{\text{bind}}$ to the free energy of alchemically transforming $L_1$ into $L_2$ when they are either bound to the protein, $\Delta G^\circ_{\text{bound}}$, or free in solution, $\Delta G^\circ_{\text{free}}$ [86],

$$\Delta\Delta G^\circ_{\text{bind}} = \Delta G^\circ_{\text{bind}}(L_2) - \Delta G^\circ_{\text{bind}}(L_1) = \Delta G^\circ_{\text{bound}} - \Delta G^\circ_{\text{free}}. \quad (5)$$

After dividing the transformation of $L_1$ to $L_2$ into a discrete number of states, described by a coupling parameter λ, multi-state Bennett acceptance-ratio method (MBAR) was used to calculate $\Delta G_{\text{bound}}$ and $\Delta G_{\text{free}}$ [87], using the pyMBAR software [88]. Energies were also calculated with Bennett acceptance ratio (BAR) [89], thermodynamic integration (TI) [90], and exponential averaging (EA) [91]. Separate calculations for the ligand free in water and bound to the protein and 13 intermediate states were used (λ = 0.00, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, and 1.00). The electrostatic and van der Waals interactions were perturbed simultaneously in each simulation using soft-core potentials for both types of interactions [92, 93].

For all ligands in set 1 and ligands **10**, **15**, **21**, **23**, **26**, **28**, and **34** in set 3, there are two possible orientations of the modified ring system. No flipping of this ring was observed during the simulations in the protein. Therefore, we run two independent perturbations starting from the two different conformations, in order to enhance the sampling. The resulting dihedral angles in the simulations and the docked structures are shown in Table S3 in the Supplementary material. Ligand **61** in set 3 has two possible configurations (*R* and *S*) and we studied both (experimentally, the racemate was studied [23]).

The alchemical perturbation simulations were performed in the following way [10]: the system at each lambda value was subjected to 100 cycles of steepest-descent minimisation, with all atoms, except water molecules and hydrogen atoms, restrained to their start position with a force constant of 418 kJ/mol/Å$^2$. This was followed by 50 ps NPT simulation and a 500 ps NPT equilibration without any restraints. Finally, a 1 ns production simulation was run. Energy differences for MBAR were sampled every 10 ps.

All minimisations and simulations were performed with the pmemd module of Amber14 [54, 94]. The temperature was kept constant at 300 K using a Langevin thermostat with a collision frequency of 2.0 ps$^{-1}$ [95] and the pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm with a relaxation time of 1 ps [96]. Long-range electrostatics were treated by particle-mesh Ewald method [97]. The cutoff for the van der Waals interactions was set to 8 Å. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm [98], so that a time step of 2 fs could be used.

### GCMC calculations

To determine the number of water molecules in the binding site of the set 2 ligand, we employed grand canonical Monte Carlo (GCMC) calculations, as implemented by Essex and coworkers [99] in the ProtoMS software package (version 3.2) [100]. The water structure was analysed for a rectangular box, extending 3 Å in all directions from the ligand, starting from the docked results. The proteins (both 2WI7 and 3FT5) were described with the Amber 14SB force field [64] and the ligands with the general Amber force field [80]. The structures were minimised using AMBER 14 [54] (100 steps minimisation via steepest descent) and then solvated with TIP4P water up to a radius of 10 Å around the protein. All the simulations were performed at 298 K, with a 10 Å cutoff for the non-bonded interactions.

Apart from standard Monte Carlo moves, such as translation and rotation, which apply to the whole system, attempts were also made to insert or delete a water molecule within the box region. The probability is controlled by the chemical potential of an ideal-gas reservoir to which the region around the ligand is being coupled. A virtual titration was performed, simulating the system at different chemical potentials (measured by the Adams value [101]). The optimal number of water molecules around the ligand was determined from the titration curve based on the simulation for which the average number of water molecules corresponds to the binding free energy minimum [99]. The simulation with this value of the chemical potential was analysed to obtain water clusters and these were used as starting positions in FES calculations.

For all systems, GCMC simulations were run for 40 evenly spaced Adams values between −20 and +19. The systems were first equilibrated with 10 million Monte Carlo moves. The first 5 million moves were dedicated to inserting, deleting, and moving water molecules within the box region. In the following 5 million moves, translations and rotations of the protein, the ligand, and the rest of the solvent were introduced for every second move, while the other moves were still dedicated to the water molecules within the box. After the equilibration, we performed 200 million moves of production, where the sampling continued in the same manner. Snapshots were recorded every 0.5 million moves of the production.

### Quality measures and uncertainty estimates

The uncertainties of the free-energy estimates were obtained by nonparametric bootstrap sampling (using 100 samples) of the work values in the MBAR calculations using the pyMBAR software [88]. The other approaches (docking, MM/GBSA, and QM/MM) are based on single structures and therefore do not provide any statistical estimate of the uncertainties. The quality of the binding-affinity estimates compared to experimental data [23] was quantified using the mean absolute deviation (MAD), the squared Pearson's correlation coefficient ($R^2$), and the Kendall's rank correlation coefficient ($\tau$). The uncertainties of the quality metrics were obtained by a parametric bootstrap (500 samples) using the uncertainties in both the calculated and experimental estimates. The experimental binding affinities were estimated from the measured $IC_{50}$ values [23] according to $\Delta G^{\circ}_{\mathrm{bind}} = RT \ln(IC_{50}/C^{\circ})$, where $R$ is the ideal gas constant, $T$ is the temperature, 300 K, and $C^{\circ}$ is the standard-state concentration, 1 M. Ligand **61** was reported as a non-binder, i.e. having $IC_{50} > 50$ μM [23] and it was assigned a binding affinity of −24.6 kJ/mol (corresponding to $IC_{50} = 50$ μM). No uncertainties for the experimental affinities were provided by the organisers. Therefore, we instead assumed a typical uncertainty of 1.7 kJ/mol for the experimental affinities [102] when calculating the uncertainties of the quality measures.

To estimate the convergence of the various perturbations, six different overlap measures were employed [10]. We calculated the Bhattacharyya coefficient for the energy distribution overlap ($\Omega$) [103], the Wu & Kofke overlap measures of the energy probability distributions ($K_{\mathrm{AB}}$) and their bias metrics ($\Pi$) [104, 105], the weight of the maximum term in the exponential average ($w_{\mathrm{max}}$) [22], the difference of the forward and backward exponential average estimate ($\Delta\Delta G_{\mathrm{EA}}$), and the difference between the BAR and TI estimates) [10]. $\Omega$ goes from 0, no overlap to 1, perfect overlap [103], and we consider values higher than 0.7 acceptable [10]. $K_{\mathrm{AB}}$ goes from 0—no overlap, via 1—full overlap, to 2—the first distribution is completely inside the second distribution [104, 105], and again values larger than 0.7 are accepted. A negative $\Pi$ indicates poor overlap and values below 0.5 are alarming [104, 105]. $1/w_{\mathrm{max}}$ indicates how many snapshots contribute significantly to the EA estimate and $w_{\mathrm{max}}$ values larger than 0.3 indicate poor convergence [10]. $\Delta\Delta G_{\mathrm{EA}}$ is the hysteresis in the forward and backward EA estimates, whereas $\Delta\Delta G_{\mathrm{TI}}$ indicates the difference between the BAR and TI estimates. In both cases, differences larger than 4 kJ/mol indicate poor convergence [10]. We examined these overlap measures for each of the 26 individual perturbations (13 λ values for simulations with or without the protein). If two of the measures indicated poor overlap (or if $\Pi$ was negative), additional simulations with intermediate λ values were run.

## Results and discussion

In the present work, we studied three congeneric series of HSP90 inhibitors, shown in Fig. 1, within the D3R 2015 grand challenge blind competition [23]. Sets 1 and 2 are

small aminopyrimidine derivatives consisting of five and four molecules, respectively, both containing a 1,3-difluorobenzene group. Set 3 is comprised of ten benzimidazolone derivatives with a 1,3-dihydroxybenzene moiety as the common scaffold. We have estimated absolute binding affinities with molecular docking, MM/GBSA, and QM/MM calculations and relative binding free energies with the FES method. In the following, we will describe the binding modes and affinities obtained with the various methods in separate sections.

### Prediction of binding modes by docking

Initial attempts using a standard docking approach, in which the receptor structure was kept rigid, did not yield satisfactory results, in that only a few ligands docked into the binding pocket. A closer inspection showed that the selected reference crystal structures contain ligands that are smaller than the studied inhibitors, although they contain the proper structural scaffolds. Therefore, steric clashes with either protein residues or surrounding water molecules occurred during the docking of most ligands. To account for protein flexibility, we instead employed the induced-fit docking (IFD) protocol [50, 51], which iteratively performs docking calculations and optimises the protein–ligand complexes through MM minimisations, effectively modelling protein structural changes upon ligand binding. This gave reasonable structures for all complexes.

All ligands bound approximately in the same position and orientation as their corresponding reference structure (Fig. 3), displaying favourable interactions with Asp93 and Gly97 in complex hydrogen-bond networks that involve several conserved water molecules. A summary of the protein–ligand interactions is given in Table 2. It shows that all ligands established a strong hydrogen bond with the Asp93 sidechain (H–O distances of $1.96 \pm 0.09$ Å). Moreover, most of the ligands displayed additional water-bridged hydrogen bonds with Asp93 and Gly97 via one crystal-water molecule (denoted Wat1). Most complexes also showed a stacked interaction between one of the benzene rings and the sidechain of Asn51, with a distance of $\sim 4$ Å between the $N^{\varepsilon 2}$ atom of Asn51 and the centre of the benzene ring [106, 107].

Set 1 ligands also exhibited hydrogen bonds with another crystal-water molecule (Wat2) that directly interacts with Asn51, as well as with Leu48, Ser52, and Thr184 in a network involving two additional water molecules (Fig. 3a). A weak hydrogen-bond with Tyr139 was also identified, where one of the chlorine atom acts as acceptor. Other minor interactions include weak π-stacking interactions with Phe138 and hydrophobic contacts with Lys58. The latter residue showed major variations in the sidechain conformation in the various structures, because this is the only residue that interacts with the variable part of the ligands. In fact, ligand **80** showed a hydrogen-bond with Lys58 sidechain instead, in which the furan oxygen atom acted as the acceptor. For all the other ligands, the sidechain of Lys58 was bent away from the ligand.

The set 2 ligands displayed interactions only with Asp93 and Wat1. However, the cyanide substituent of compound 100 replaced the role of Wat2 in Set 1 and established a hydrogen-bond with Asn51 (cf. Figure 3b). To make the results comparable, Wat2 was excluded in the calculations for all four ligands. After submission, we also tested docking calculations based on the 2WI7 crystal structure (which has a ligand that is chemically more similar to the set 2 scaffold) and kept Wat2 when docking ligands **101**, **105**, and **106**. The results (also included in Table 2) showed that these three ligands can make strong hydrogen bonds to Wat2. Strong interactions with Wat3 and Wat1 were also observed, whereas the interactions with Asp93 became more variable (Fig. 3c). The water molecules bridged interactions with Leu48, Asn51, Asp93, and Gly97. Moreover, the pyrazole ring nitrogen of ligands **105** and **106** established a second hydrogen bond with Wat1 (Fig. 3c).

In general, set 3 inhibitors exhibited a larger number of interactions, and also shorter distances than in the other two sets. In particular, the presence of hydroxyl and carbonyl groups allowed the formation of additional short direct hydrogen bonds with Gly97 and Thr184, where one of the hydroxyl substituents appears to have displaced Wat2 (not present in the reference crystal, 3OW6) in favour of direct hydrogen bonds with Asn51, and allowed for reaching water Wat3, establishing further hydrogen bonds with Leu48, Ser52, Ile91, and Asp93. Major movements were observed for the Lys112 and Phe138 sidechains (Fig. 3d), which were shifted towards the ligands to form cation–π and π-stacking interactions, respectively. The geometry of the cation–π interaction with Lys112 showed a great variability, indicating that this interaction may be important for regulation of the activity. For ligand **61**, only the R conformation was found to bind to the protein in a reasonable mode.

### Binding affinities estimated by docking and MM/GBSA

We have estimated the binding affinities for the three sets of ligands with three scoring functions (all employing the same final IFD structures in Fig. 3): GlideScore (GScore), $E_{model}$ and IFDScore (which is the GScore plus a portion of the Prime MM energy from the refinement calculation). In addition, all docked complexes were scored with MM/GBSA calculations, after minimisation of the docked structures. The calculated binding affinities are shown in Table 3. The performance of the tested scores was

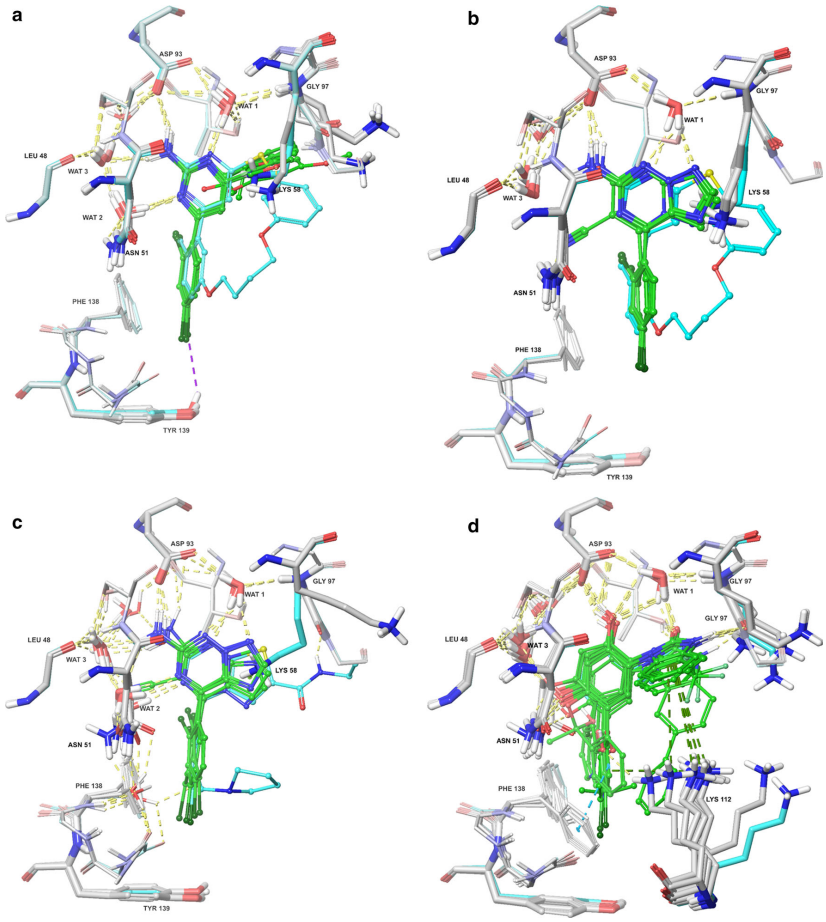



**Fig. 3** Binding modes for the three series of HSP90 inhibitor from the docking calculations: **a** set 1, **b** original docking for set 2, based on the 3VHA crystal structure (submitted), **c** set 2 in the 2WI7 crystal structure, keeping all water molecules, and **d** set 3. Carbon atoms of the residues are shown in *light grey tubes*, showing some movements as result of the induced-fit docking protocol. Carbon atoms of the ligands are shown as *green tubes*. Water molecules that interact with the ligands are displayed in *thick tube* representation and labelled as WAT. Reference crystal structures (3VHA, 2WI7, and 3OW6 [36, 37, 39]) are coloured in *cyan* for comparison (both ligands and protein). Nitrogen and oxygen atoms are *blue* and *red*, respectively. Hydrogen bonds are represented as *yellow dashed lines* (*purple* if the acceptor is a halogen atom). Cation-π and π-stacking interactions are represented as *dark green* and *dark cyan dashed lines*, respectively

evaluated by three quality metrics: the correlation coefficient ($R$), Kendall's rank correction coefficient (τ), and the mean absolute deviation after removal of the systematic error (i.e. the mean signed error; MADtr), which are listed at the bottom of Table 3. The correlation between the experimental [23] and calculated binding affinities are shown in Fig. 4.

The results for set 1 were poor, with a negative or vanishing τ for all methods and a negative (MM/GBSA) or very low correlation ($R = 0.0$–$0.2$). However, the MADtr

**Table 2** Hydrogen bonds (first eight lines) and cation–π interactions (last line, Lys122) in the structures obtained with the induced-fit docking

| Residues | Set 1 | | Set 2 | | Set 2[a] | | Set 3 | |
|---|---|---|---|---|---|---|---|---|
| | *n* | *r* | *n* | *r* | *n* | *r* | *n* | *r* |
| Lys58 | 1 | 2.12 | | | | | | |
| Asp93 | 5 | 2.08 ± 0.09 | 4 | 1.86 ± 0.07 | 4 | 1.90 ± 0.50 | 10 | 1.94 ± 0.11 |
| Wat1 | 5 | 1.90 ± 0.07 | 4 | 2.12 ± 0.14 | 4 | 2.13 ± 0.15 | 10 | 1.84 ± 0.09 |
| Wat2 | 5 | 1.94 ± 0.07 | | | 3 | 2.10 ± 0.05 | | |
| Asn51 | | | 1 | 2.27 | | | 8 | 2.17 ± 0.12 |
| Wat3 | | | | | 4 | 2.20 ± 0.05 | 9 | 1.97 ± 0.17 |
| Gly97 | | | | | | | 10 | 2.19 ± 0.14 |
| Thr184 | | | | | | | 10 | 1.83 ± 0.08 |
| Asn51[b] | 5 | 4.08 ± 0.05 | 4 | 4.30 ± 0.17 | 4 | 5.38 ± 0.71 | 5 | 4.13 ± 0.71 |
| Lys112 | | | | | | | 8 | 5.40 ± 0.76 |

For each interaction, the number of structures in which this interaction is found is given (*n*, out of 5, 4, and 10 structures for sets 1–3, respectively) and the average distance in these structures (*r* in Å), together with the standard deviation over the *n* structures. Wat1–Wat3 are crystal-water molecules

[a] A second set of docking calculations for set 2, using the 2WI7 crystal structure and keeping Wat2 for ligands **101**, **105**, and **106** (but not **100**), done after the experimental results were revealed

[b] Interaction in which the plane of the sidechain amide group is nearly parallel to the plane of the aromatic ring. The average distance between the $N^{\varepsilon 2}$ of Asn51 and the centre of the aromatic ring is given

is good for both GScore and IFDScore, 4 kJ/mol, but this is mainly an effect of the fact that the range of the predicted affinities is small, 4–7 kJ/mol, compared to experimental range of 11 kJ/mol (setting all calculated affinities to the same value gives a MADtr of 3 kJ/mol).

For the original calculations on set 2, all four methods also gave very poor results, with strong negative correlations ($R = -0.9$ to $-1.0$), owing to the fact that all methods predicted ligand **100** to bind best, although it experimentally is the weakest ligand. This also gave a large MADtr to all methods (7–18 kJ/mol) and a negative or vanishing τ.

For the calculations based on the 2WI7 crystal structure, in which Wat2 was kept for ligands **101**, **105**, and **106**, the results were more varying (also included in Table 2). The GScore energies showed no correlation with the experimental data, whereas the internal docking score $E_{model}$ produced reasonable correlation ($R = 0.67$) and a correct ligand ranking ($\tau = 1.00$). The IFDScore showed intermediate results ($R = 0.42$ and $\tau = 0.33$). The MM/GBSA results were very poor, with negative *R* and τ. On the other hand, MADtr was best for IFDScore (5 kJ/mol). All methods still predicted ligand **100** to bind with a potency comparable to the other ligands, probably because the employed docking and MM/GBSA rescoring approaches did not consider the cost of displacing Wat2 when ligand **100** binds. In fact, most quality measures improved significantly if ligand **100** was excluded.

For set 3, the results are somewhat better: all methods gave a positive correlation ($R = 0.1$–$0.7$) and a positive τ (0.1–0.4; however, it should be noted that four of the ligands have experimental affinities within 1 kJ/mol, making it questionable to calculate τ for these—it would be better to consider only statistically significant differences, e.g. $\tau_{90}$ [14]). Both *R* and τ were best for MM/GBSA, but MM/GBSA and $E_{model}$ gave poor MADtr (29 and 22 kJ/mol), which reflects that the results for these two methods have a much larger range than the experimental data (124 and 111 compared to 19 kJ/mol). On the other hand, MADtr of GScore and IFDScore is much better, 4 and 5 kJ/mol, but again the ranges are smaller than for the experimental results, 7 and 13 kJ/mol.

Two sets of absolute affinities were submitted, viz. the original GScore and MM/GBSA (submission entries 56afbe93eeaf4 and 56afbea4a8c67, respectively) results in Table 3 (based on the 3VHA structure without Wat2 for set 2).

## QM/MM estimates

Next, we tried to estimate the binding free energies also with a QM/MM approach. As described in the Methods section, we started from the final induced-fit docked structures, to which a sphere of water molecules was added and optimised (together with the hydrogen atoms). Then, a QM system of 280–320 atoms was optimised by QM/MM at the TPSS/def2-SV(P) level of theory (Fig. 2a, b).

**Table 3** Binding affinities ($\Delta G_{bind}$ in kJ/mol) for the three studied HSP90 inhibitor sets calculated with Glide (GScore and $E_{model}$), induced-fit docking protocol (IFDScore), and MM/GBSA. In addition, the experimental data [23] are included (Exp.)

| | Ligand | Exp. | GScore | $E_{model}$ | IFDScore | MM/GBSA |
|---|---|---|---|---|---|---|
| Set 1 | **80** | −32.6 | −42.6 | −326.8 | −2079.7 | −367.9 |
| | **81** | −38.2 | −46.9 | −398.1 | −2086.9 | −379.7 |
| | **82** | −28.2 | −47.0 | −378.7 | −2083.8 | −395.5 |
| | **83** | −27.5 | −45.4 | −375.6 | −2085.9 | −413.1 |
| | **84** | −29.9 | −45.4 | −368.3 | −2081.5 | −405.6 |
| Set 2 | **100** | −24.6 | −41.3 | −333.9 | −2054.0 | −342.8 |
| 3VHA | **101** | −38.3 | −38.7 | −289.5 | −2047.0 | −311.1 |
| | **105** | −39.5 | −37.4 | −300.7 | −2046.5 | −319.5 |
| | **106** | −40.3 | −37.8 | −308.4 | −2046.6 | −298.7 |
| Set 2 | **100** | −24.6 | −39.7 | −282.2 | −1973.7 | −357.2 |
| 2WI7 | **101** | −38.3 | −38.4 | −282.9 | −1973.4 | −310.9 |
| | **105** | −39.5 | −40.7 | −307.1 | −1978.7 | −348.0 |
| | **106** | −40.3 | −39.5 | −309.7 | −1974.8 | −338.8 |
| Set 3 | **10** | −30.3 | −51.5 | −444.6 | −1980.7 | −292.2 |
| | **11** | −38.1 | −46.9 | −401.6 | −1977.9 | −307.1 |
| | **15** | −29.5 | −54.1 | −512.3 | −1988.8 | −347.7 |
| | **19** | −29.6 | −47.8 | −444.2 | −1980.4 | −306.3 |
| | **21** | −38.3 | −54.2 | −493.0 | −1989.1 | −395.8 |
| | **23** | −31.5 | −50.3 | −468.3 | −1985.9 | −333.7 |
| | **26** | −43.9 | −53.7 | −458.7 | −1987.1 | −383.3 |
| | **28** | −37.4 | −54.4 | −457.6 | −1986.9 | −316.1 |
| | **34** | −29.7 | −51.4 | −447.2 | −1991.3 | −305.0 |
| | **61**(R) | >−24.6 | −50.2 | −415.5 | −1981.7 | −271.4 |
| MADtr | Set 1 | | 3.9 | 17.6 | 3.8 | 18.2 |
| | Set 2 | | 6.8 | 18.5 | 8.3 | 18.0 |
| | 3WI7 | | 5.6 | 8.7 | 4.8 | 17.5 |
| | Set 3 | | 4.3 | 22.3 | 5.1 | 29.4 |
| *R* | Set 1 | | 0.05 | 0.20 | 0.21 | −0.70 |
| | Set 2 | | −0.97 | −0.86 | −1.00 | −0.91 |
| | 3WI7 | | −0.02 | 0.67 | 0.42 | −0.55 |
| | Set 3 | | 0.32 | 0.11 | 0.16 | 0.70 |
| τ | Set 1 | | −0.20 | 0.00 | −0.20 | −0.60 |
| | Set 2 | | −0.67 | 0.00 | −0.67 | −0.67 |
| | 3WI7 | | 0.00 | 1.00 | 0.33 | −0.33 |
| | Set 3 | | 0.20 | 0.20 | 0.11 | 0.42 |

For set 2, two series of results are given, based on either the 3VHA or 2WI7 crystal structures, the latter including Wat2 for ligands **101**, **105**, and **106**. The lower part of the table contains the quality metrics of the various results: the mean absolute deviation after removal of the systematic error (MADtr), the correlation coefficient (*R*) and Kendall's rank correlation coefficient (τ). Only the best scores among all obtained structures are reported

Finally, a big-QM calculation was performed for a QM system involving all protein residues and water molecules within 7.5 Å of the ligand, 970–1160 atoms, shown in Fig. 2c, d), calculated at the TPSS/def2-SV(P) level of theory in a COSMO continuum solvent. To the big-QM energy, entropy, basis-set, and DFT-D3 dispersion corrections were added, in addition to the relaxation energy and a more accurate COSMO-RS solvation energy of the ligand (Eq. 4).

The QM/MM structures were qualitatively similar to the docked structures, but with some differences in the hydrogen-bond distances, as can be seen by comparing Tables 2 and 4. For set 1, the bonds to Asp93 were shortened, whereas those to Wat1 were elongated. For set 2, the structures of the four ligands were more similar, but the hydrogen-bond interaction with Wat1 was strengthened. For set 3, the hydrogen bonds to Asp93, Wat3,

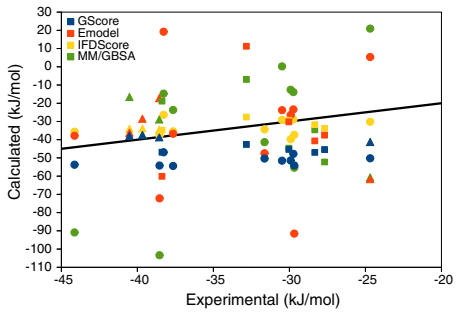



**Fig. 4** Correlation between the experimental [23] and calculated binding affinities. Sets 1–3 are marked with *squares*, *triangles*, and *circles*, respectively. For GScore, the original score is shown, whereas for $E_{model}$, IFDScore, and MM/GBSA, the mean signed error is subtracted (to give a similar scale of all the calculated results). The *line* shows the perfect correlation. Ligand **61** was experimentally found to be a non-binder, i.e. with a $K_i > 50$ μM, which corresponds to $\Delta G_{bind} > -25$ kJ/mol

Gly97, and Thr184 were much shortened, whereas that to Wat1 was elongated.

Table 5 shows the various QM/MM (free) energy components for the 19 ligands and their correlation to the experimental data. It can be seen that the raw QM/MM energies were large and negative (−620 kJ/mol on average). The same applies to the $E^{HL}_{QM1+ptch2}$ energy component (−557 kJ/mol on average), showing that the QM/MM energy is dominated by the QM energy. Neither term showed any convincing correlation to experimental data. The big-QM energies were less negative, especially in the COSMO solvent (−127 kJ/mol on average). However, the correlation to the experimental data was still poor for all three sets of ligands, $R = -0.1$ to 0.3.

The dispersion energy was large and negative, showing a smaller variation than the QM energies (−309 kJ/mol on average). It was compensated by the basis-set correction and the $\Delta G_{therm}$ terms, which both were positive, 177 and 104 kJ/mol on average. Neither term showed any consistent correlation to the experimental data. The relaxation energy of the ligand was 10–61 kJ/mol, largest for the set 3 ligands and smallest for set 2. It showed only a minor variation depending on whether it was calculated with the def2-SV(P) or def2-QZVP basis sets or with or without the COSMO solvation energy (less than 11 kJ/mol). The COSMO-RS solvation energies of the ligand were −48 to −141 kJ/mol, more negative for the set 3 ligands than for the ligands of the other two sets. The COSMO-RS solvation energy was always more negative than the pure COSMO solvation energy, by 23 kJ/mol on average. Neither of the ligand terms showed any consistent correlation to the experimental data.

Adding all the terms according to Eq. 4, we obtained the full QM/MM binding free energy ($\Delta G_{bind}$). From Table 5, it can be seen that it was too negative compared to the experimental data and also with a too large range (−34 to −164 kJ/mol). For sets 2 and 3, it showed a weak correlation with the experimental data ($R = 0.5$ and 0.3, respectively), whereas for set 1, the correlation was negative ($R = -0.7$). For all three sets, MADtr was large, 17–30 kJ/mol. In fact, the results could be improved if the $\Delta G_{therm}$ and $\Delta E_{L,rlx}$ terms were omitted ($\Delta G'_{bind}$ column in Table 5). Then, MADtr was only 6 kJ/mol for set 2 and

**Table 4** Hydrogen bonds (first eight lines) and cation–π interactions (last line, Lys122) in the structures obtained with QM/MM optimisation

| Residues | Set 1 | | Set 2 | | Set 3 | |
|---|---|---|---|---|---|---|
| | *n* | *r* | *n* | *r* | *n* | *r* |
| Lys58 | 1 | 2.29 | | | | |
| Asp93 | 5 | 1.86 ± 0.04 | 4 | 1.83 ± 0.07 | 10 | 1.53 ± 0.05 |
| Wat1 | 5 | 2.08 ± 0.08 | 4 | 1.89 ± 0.05 | 10 | 2.07 ± 0.07 |
| Wat2 | 5 | 1.87 ± 0.05 | | | | |
| Asn51 | | | 1 | 2.52 | 9 | 2.08 ± 0.12 |
| Wat3 | | | | | 9 | 1.62 ± 0.03 |
| Gly97 | | | | | 10 | 1.75 ± 0.03 |
| Thr184 | | | | | 10 | 1.67 ± 0.04 |
| Asn51[a] | 5 | 3.95 ± 0.11 | 4 | 4.22 ± 1.03 | 9 | 3.75 ± 0.21 |
| Lys112 | | | | | 10 | 5.42 ± 0.29 |

For each interaction, the number of structures in which this interaction is found is given (*n*, out of 5, 4, and 10 structures for sets 1–3, respectively) and the average distance in these structures (*r* in Å), together with the standard deviation over the *n* structures. Wat1–Wat3 are crystal-water molecules

[a] Interaction in which the plane of the sidechain amide group is nearly parallel to the plane of the aromatic ring. The average distance between the $N^{\varepsilon 2}$ of Asn51 and the centre of the aromatic ring is given

**Table 5** The various QM/MM (free-) energy terms (kJ/mol): the QM/MM energy ($\Delta E_{QM/MM}$), the $E^{HL}_{QM1+ptch2}$ energy ($\Delta E_{QM+ptch}$), the big–QM energy ($\Delta E_{BQ}$), calculated either with a point-charge (ptch) model of the surroundings or with COSMO solvation, the dispersion energy, the basis-set correction energy (Eq. 2), the $\Delta G_{therm}$ ZPE, entropy, and thermal correction, the ligand relaxation energy ($\Delta E_{L,rlx}$), the ligand solvation energy ($\Delta G_{L,solv}$), calculated either at the COSMO (TPSS/def2–SV(P)) or COSMO–RS (BP/TZVP) levels (the $\Delta\Delta G_{L,solv}$ term in Eq. (4) is the difference of those two energy terms), and the final QM/MM binding free energy from Eq. (4) ($\Delta G_{bind}$) and the same energy, excluding the $\Delta G_{therm}$ and $\Delta E_{L,rlx}$ terms ($\Delta G'_{bind}$). The last nine lines in the table give MADtr, $R$ and $\tau$ compared to the experimental data [23]

| Ligand | $\Delta E_{QM/MM}$ | $\Delta E_{QM+ptch}$ | $\Delta E_{BQ}$ | | $\Delta E_{disp}$ | $\Delta E_{bsc}$ | $\Delta G_{therm}$ | $\Delta E_{L,rlx}$ | $\Delta G_{L,solv}$ | RS | $\Delta G_{bind}$ | $\Delta G'_{bind}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ptch | COSMO | | | | QZP | COSMO | | | |
| **80** | −484.2 | −426.7 | −156.2 | −56.9 | −285.5 | 149.3 | 93.1 | −27.1 | −45.9 | −54.8 | −64.1 | −157.2 |
| **81** | −565.5 | −491.4 | −214.4 | −76.2 | −324.3 | 157.4 | 111.3 | −33.8 | −55.4 | −69.2 | −84.3 | −195.5 |
| **82** | −487.3 | −421.6 | −145.2 | −45.9 | −316.1 | 145.3 | 81.4 | −21.5 | −42.3 | −51.5 | −104.6 | −186.0 |
| **83** | −550.4 | −470.1 | −218.3 | −82.9 | −337.8 | 160.5 | 99.2 | −32.2 | −51.3 | −60.5 | −119.6 | −218.8 |
| **84** | −544.9 | −471.9 | −158.4 | −45.0 | −340.5 | 152.6 | 80.5 | −32.7 | −57.7 | −65.8 | −111.6 | −192.1 |
| **100** | −487.8 | −425.3 | −181.7 | −65.7 | −249.1 | 148.0 | 100.1 | −16.2 | −65.2 | −81.9 | −33.9 | −133.9 |
| **101** | −475.4 | −419.0 | −164.5 | −67.6 | −266.1 | 158.8 | 61.1 | −10.6 | −39.3 | −48.1 | −94.4 | −155.5 |
| **105** | −433.5 | −379.7 | −157.8 | −54.3 | −238.3 | 133.1 | 86.4 | −10.2 | −49.6 | −63.2 | −49.3 | −135.7 |
| **106** | −438.7 | −384.7 | −170.7 | −70.4 | −231.7 | 130.5 | 93.7 | −10.7 | −51.5 | −66.1 | −52.7 | −146.3 |
| **10** | −760.1 | −697.2 | −447.5 | −242.3 | −307.2 | 196.8 | 123.8 | −54.7 | −103.8 | −141.1 | −136.8 | −260.6 |
| **11** | −707.8 | −645.2 | −336.0 | −153.8 | −324.3 | 194.7 | 112.4 | −54.8 | −89.7 | −122.8 | −83.0 | −195.4 |
| **15** | −851.1 | −705.8 | −388.7 | −182.1 | −353.4 | 256.5 | 136.8 | −48.2 | −101.5 | −127.4 | −68.1 | −204.9 |
| **19** | −711.6 | −640.0 | −386.5 | −215.5 | −284.3 | 191.0 | 106.6 | −31.6 | −91.4 | −123.6 | −138.4 | −244.9 |
| **21** | −776.1 | −684.7 | −349.5 | −178.5 | −359.3 | 217.3 | 150.4 | −35.9 | −92.7 | −120.9 | −105.9 | −256.3 |
| **23** | −748.1 | −676.3 | −389.8 | −185.3 | −338.8 | 186.2 | 85.8 | −47.0 | −98.8 | −140.3 | −163.7 | −249.5 |
| **26** | −726.8 | −658.0 | −349.4 | −176.8 | −341.0 | 203.6 | 116.0 | −34.9 | −89.9 | −123.9 | −129.3 | −245.3 |
| **28** | −750.0 | −685.2 | −379.5 | −190.6 | −325.4 | 203.3 | 107.6 | −43.9 | −95.2 | −125.4 | −131.0 | −238.6 |
| **34** | −749.5 | −687.7 | −424.5 | −235.0 | −291.0 | 196.4 | 104.5 | −52.2 | −100.9 | −141.0 | −132.7 | −237.2 |
| **61** | −687.3 | −622.1 | −282.7 | −93.5 | −352.0 | 187.6 | 123.8 | −60.8 | −90.3 | −121.0 | −42.6 | −166.5 |
| MADtr | 31.9 | 25.0 | 29.0 | 13.3 | 16.7 | 6.0 | 11.5 | 4.5 | 5.2 | 5.1 | 21.4 | 14.5 |
| | 27.0 | 24.2 | 12.1 | 7.0 | 15.5 | 9.4 | 10.7 | 7.7 | 12.5 | 14.1 | 17.1 | 6.1 |
| | 31.6 | 23.7 | 37.6 | 30.1 | 21.0 | 16.7 | 13.9 | 11.1 | 7.8 | 9.6 | 30.0 | 23.0 |
| *R* | 0.33 | 0.44 | 0.33 | 0.29 | −0.26 | −0.21 | −0.70 | 0.41 | 0.36 | 0.57 | −0.67 | −0.27 |
| | −0.78 | −0.73 | −0.83 | −0.11 | −0.22 | 0.37 | 0.49 | −0.99 | −0.81 | −0.76 | 0.53 | 0.55 |
| | −0.01 | 0.09 | −0.11 | 0.05 | 0.21 | −0.08 | −0.09 | −0.53 | −0.40 | −0.32 | 0.27 | 0.38 |
| $\tau$ | | | | | | | | | | | −0.60 | −0.20 |
| | | | | | | | | | | | 0.33 | 0.33 |
| | | | | | | | | | | | 0.07 | 0.33 |

14–23 kJ/mol for the other two sets. It is often observed with the similar MM/GBSA approach that the results are improved if the $\Delta G_{therm}$ term is omitted [2]. The reason is probably that the complex and protein structures may relax to different local minima during the MM minimisation. Likewise, MM/GBSA almost invariably exclude the ligand and protein relaxation energies, because they strongly increase the statistical uncertainty of the results [2]. For the rigid octa-acid host–guest system in the SAMPL4 competition, an improvement of the results was obtained if the ligand-relaxation energy was included [16], but with the more flexible ligands in the SAMPL5 competition, the results were deteriorated [108].

Compared to the docking and MM/GBSA results in Table 3, the QM/MM calculations gave much better correlation and $\tau$ for set 2, similar or slightly worse results for set 3, and much worse for set 1 (except for MM/GBSA). MADtr was also better for set 2, whereas it was worse than the GScore and IFDScore for the other two sets. One set of relative QM/MM affinities was submitted (submission entry 56af85ab34dbd), viz. the $\Delta G_{bind}$ results in Table 5, but unfortunately with a sign error in the $\Delta\Delta G_{L,solv}$ term in Eq. (4).

## FES results

Relative binding free energies between pairs of ligands were estimated using alchemical FES calculations and employing the standard thermodynamic cycle with the two ligands either bound to the protein or free in solution [86]. Free-energy differences were calculated with the MBAR, BAR, TI, and EA methods. Most of the calculations in sets 1 and 3 involved reference ligands to make the perturbations smaller.

The average structures of the HSP90–ligand complexes are described in Table 6. For set 1, we find that the ligands bind in a mode that is rather similar to that found in the docking and the QM/MM optimisations (Fig. 5a): all ligands formed a direct hydrogen bonds to Asp93 and the two water molecules Wat1 and Wat2, as well as the stacking interaction between the aromatic ring of the ligand and the sidechain of Asn51. However, in variance to the docked and QM/MM structures, all ligands in the FES structures showed also a hydrogen bond to Wat3.

Ligands from set 2 bind differently in the FES simulations started from the crystal structures 2WI7 and 3FT5. Structures obtained with the 2WI7 structure were quite similar to the docked and QM/MM structures (Fig. 5b), in which each ligand directly interacted with Asp93 and formed a hydrogen bond network involving Wat1, Asp93, Thr184, and Gly97. No water molecule replaced the deleted Wat2 molecule. In the 3FT5 structures, the ligands still showed a direct hydrogen bond to Asp93, but the ligands were rotated so that the hydrogen-bond network was moved towards Asn51 and involved Wat2, Wat3, and a third water molecule, shown in Fig. 5c. Ligand **100** of the 3ft5 subset also formed two direct hydrogen bonds with Thr184 and Gly97 (Fig. 5d).

**Fig. 5** Binding modes in the FES calculations. **a** ligand **80** (set 1; all the other ligands in this set bind in a similar mode), **b** set 2 ligands, based on the 2WI7 crystal structure, **c** ligands **101**, **105**, and **106** (set 2) with three water molecules in different colours (the one in *magenta* corresponds to Wat2 and that in *orange* corresponds to Wat3), **d** ligand **100** (set 2), based on the 3FT5 crystal structure, and **e** ligand **10** (set 3; all the other ligands in this set bind in a similar mode). Hydrogen bonds are indicated by *green dotted lines*

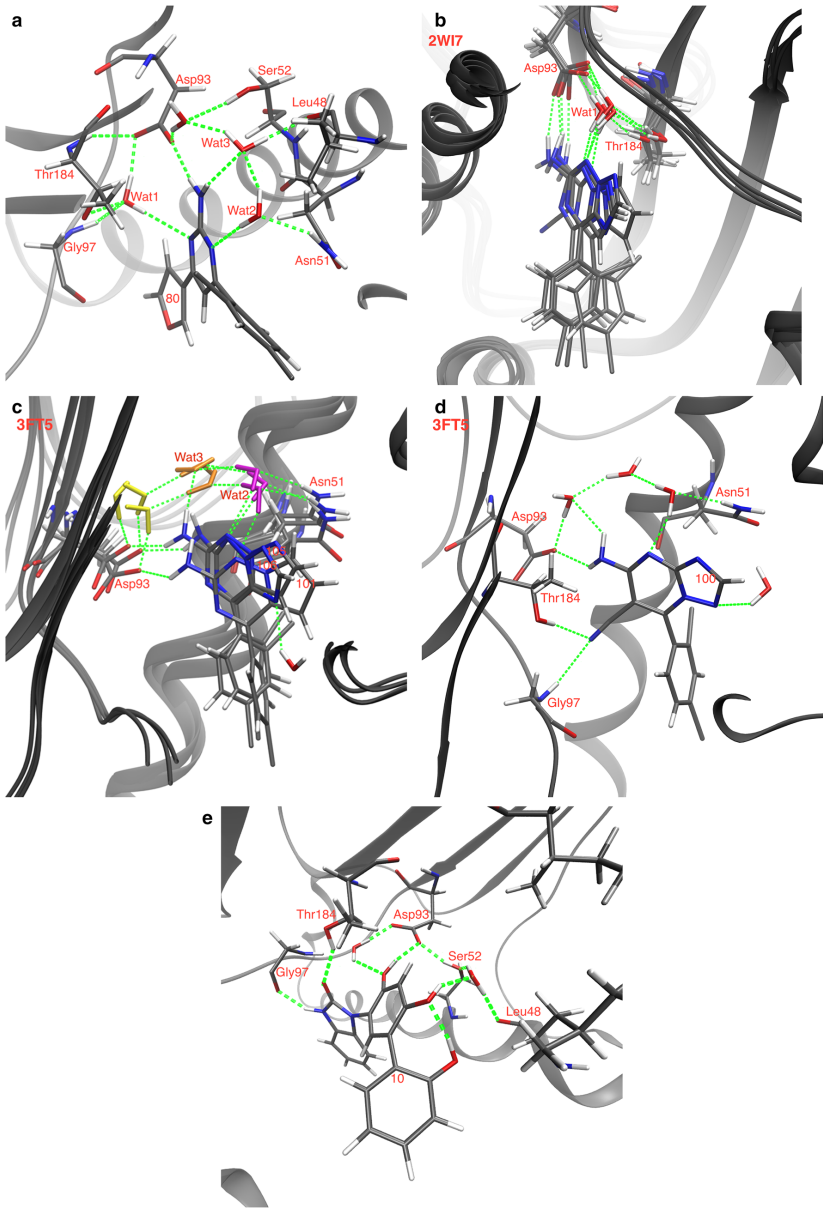
After submission of the results, we performed GCMC calculations to study the water structure around the ligands of set 2. These calculations are described in the Supplementary material. The resulting clustered water molecules around the various ligands are shown in Fig. 6. It can be seen that for the 2WI7 structure, the cyano group in ligand **100** replaced two water molecules that were present for the other three ligands (Wat2 and Wat3). For the 3FT5 structure, only one water molecule (Wat1) was displaced by the cyano group in ligand **100**. Therefore, we performed an additional set of FES calculations (using both the 2WI7 and 3FT5 structures), in which all water molecules were included in the perturbations. For ligand **100** in the 2WI7 structure, Wat2 moved away from the ligand and ended up in bulk solvent, whereas for the other ligands, Wat2 stayed in the original position. Wat3 remained in the starting position in all calculations with the 2WI7 structure (i.e. also for ligand **100**). For the calculations in the 3FT5 structure, Wat1 did not interact directly with any of the ligands (the distance was $\sim$2.7 Å). For ligand **100**, Wat3

**Table 6** Hydrogen bonds in the structures obtained in the FES calculations (the most stable conformation of the ligand for Sets 1 and 3)

| Residues | Set 1 | | Set 2 (2WI7) | | Set 2 (3FT5) | | Set 2 (2WI7 + Wat2) | | Set 2 (3FT5 + Wat1) | | Set 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | *r* | *n* | *r* | *n* | *r* | *n* | *r* | *n* | *r* | *n* | *r* |
| Asp93 | 5 | 2.01 ± 0.07 | 4 | 1.90 ± 0.03 | 4 | 2.15 ± 0.10 | 4 | 1.87 ± 0.03 | 3 | 1.98 ± 0.01 | 10 | 1.69 ± 0.04 |
| Wat1 | 5 | 2.57 ± 0.15 | 4 | 2.16 ± 0.07 | | | 4 | 2.12 ± 0.07 | | | 10 | 2.39 ± 0.09 |
| Wat2 | 5 | 2.28 ± 0.09 | | | 4 | 2.17 ± 0.19 | 3 | 2.16 ± 0.03 | 3 | 2.12 ± 0.11 | | |
| Asn51 | | | | | | | | | | | 1 | 2.11 |
| Wat3 | 5 | 2.22 ± 0.04 | 4 | 2.33 ± 0.05 | 1 | 2.50 | 4 | 2.17 ± 0.05 | 1 | 2.43 | 10 | 1.95 ± 0.27 |
| Gly97 | | | | | 1 | 2.50 | | | | | 10 | 2.05 ± 0.05 |
| Thr184 | | | | | 1 | 2.46 | | | | | 10 | 1.88 ± 0.09 |
| Asn51[a] | 5 | 3.94 ± 0.06 | | | | | | | | | | |

For each interaction, the number of structures in which this interaction is found is given (*n*, out of 5, 4, and 10 structures for sets 1–3, respectively) and the average distance for the various ligands over average in the $\lambda = 0$ or 1 simulations (*r* in Å), together with the standard deviation over the *n* ligands. Wat1–Wat3 are crystal-water molecules. No cation–π interactions with Lys122 were found for any ligand

[a] Interaction in which the plane of the sidechain amide group is nearly parallel to the plane of the aromatic ring. The average distance between the $N^{\varepsilon 2}$ of Asn51 and the centre of the aromatic ring is given


a
Asp93
Ser52
Leu48
Wat3
Thr184
Wat1
Wat2
Gly97
Asn51
80
b
2WI7
Asp93
Wat1
Thr184
c
3FT5
Wat3
Asn51
Wat2
Asp93
d
3FT5
Asn51
Asp93
Thr184
Gly97
e
Thr184
Asp93
Ser52
Gly97
Leu48
10

came in and bridged the interaction with Asp93. Thereby, it interacted very weakly with the protein.

All the ligands from set 3 bound to the protein in a similar way, with rather small variations between the different ligands. Each ligand forms direct hydrogen bonds with Gly97, Thr184, and Asp93, and also an additional water-bridged interaction with the latter residue. Each ligand also binds to Ser52 and Leu48 via a water molecule (Fig. 5e). These binding modes are quite similar to the ones observed in the docking and the QM/MM results for the set 3. However, we do not find any interaction with Ile91, and Lys112 is far away from the ligand.

For set 1, the perturbations involved mainly the substituents of one of the three ring systems, involving the perturbation of one (or in one case two) hydrogen atoms to methyl, methoxy, or ethoxy groups. In one case, the benzene ring was instead perturbed to a furan ring (ref → **80**). In another case, a methyl group is perturbed to an acetate group (**81** → **82**). Set 2 involves perturbations of C and N atoms in a fused six and five-ring system. In one case (**100**), a cyano group is also added. Set 3 is more diverse, although all ligands share a benzimidazolone group joined to a resorcinol group. By the use of three reference ligands, the size of the perturbations was in many cases reduced to the conversion of hydrogen atoms to hydroxyl, chloride, methoxy, $CF_3$, and isopropyl groups, or to the conversion of a carbon atom in the benzene ring to a nitrogen atom (pyridine). However, in one case a hydrogen atom is converted to a benzene ring (**19** → ref1), in one case the benzene ring is converted to quinoline (**23** → ref2), and in one case, the benzene and resorcinol rings are joined by a pyran ring (**61** → ref2).

The raw binding affinities calculated with FES are given in Table 7. It can be seen that the precision of the FES results was reasonable: the standard errors of the MBAR estimates were 0.2–0.9 kJ/mol, indicating good convergence of the perturbations. Results obtained with the BAR, TI, and EA methods are shown in Table S4 in the Supplementary material. The BAR and TI results agreed with the MBAR results with MADs of 0.6 and 0.8 kJ/mol, respectively, which indicates a somewhat worse convergence. In particular, the **21** → ref3 and **26** → ref2 perturbations gave alarming differences of 4 and 5 kJ/mol, respectively. The convergence of all perturbations was examined by considering a set of six overlap measures, as described in the Methods section. All 26 individual simulations for each perturbation were checked for poor overlap and additional simulations were run with intermediate λ values if two of the overlap measures indicated poor overlap or if Π (which is considered to be the most reliable overlap measure, with the best correlation to the other measures [10]) was negative. Consequently, the presented results should be numerically reliable.

As mentioned in the Methods section, many of the ligands in sets 1 and 3 can bind with two conformations, differing by an 180° rotation of the perturbed ring. In the FES calculations, both conformations were tested, starting from the symmetric reference molecules. The best conformation was then selected as the one that gave the most favourable binding energy, compared to the reference

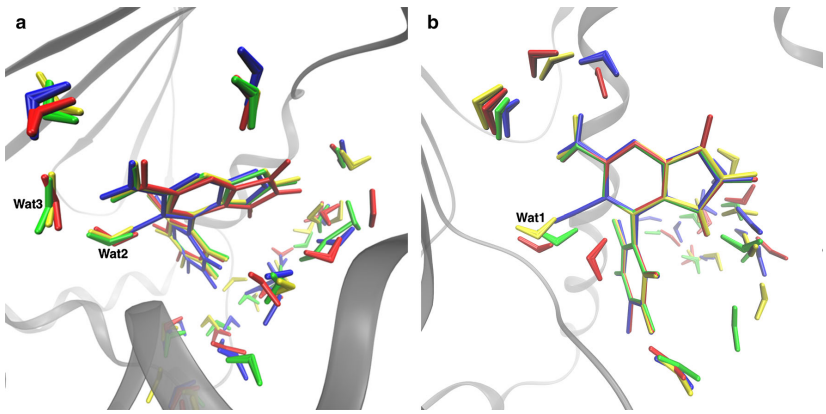



**Fig. 6** Water clusters obtained by GCMC method for the **a** 2WI7 and **b** 3FT5 structures with set 2 ligands. In both figures, ligands and the corresponding water molecules are presented in different colours: ligand **100**—*blue*, ligand **101**—*red*, ligand **105**—*yellow*, and ligand **106**—*green*

**Table 7** Calculated relative binding free-energies and standard errors (obtained with MBAR in kJ/mol) for the studied perturbations

| Transformation<br>Set 1 | Exp. | Results 1<br>Conf. 1 | Results 2<br>Conf. 2 | Results 3 |
|---|---|---|---|---|
| ref → **80** | | 1.8 ± 0.5 | **−3.6 ± 0.5** | |
| **81** → **82** | 10.0 | **−13.2 ± 0.5** | −16.4 ± 0.5 | |
| **82** → ref | | 13.3 ± 0.3 | **16.4 ± 0.3** | |
| **83** → ref | | **3.6 ± 0.5** | −3.5 ± 0.5 | |
| **84** → ref | | **8.3 ± 0.5** | **8.3** ± 0.6 | |
| Set 2 without Wat1/2 | | 2WI7 | 3FT5 | 2WI7/3FT5 |
| **101** → **100** | 13.9 | −12.2 ± 0.5 | 2.7 ± 0.5 | −12.8 ± 0.5 |
| **101** → **105** | −0.1 | −7.5 ± 0.2 | 2.7 ± 0.2 | −8.4 ± 0.2 |
| **101** → **106** | 2.0 | −7.3 ± 0.3 | 3.8 ± 0.3 | −8.7 ± 0.3 |
| Set 2 with Wat1/2 | | 2WI7 | 3FT5 | |
| **101** → **100** | 13.9 | 11.2 ± 0.9 | 18.0 ± 0.9 | |
| **101** → **105** | −0.1 | −6.2 ± 0.4 | 3.5 ± 0.4 | |
| **101** → **106** | 2.0 | −3.7 ± 0.5 | 5.5 ± 0.5 | |
| Set 3 | | Conf. 1 | Conf. 2 | |
| **10** → ref2 | | −4.9 ± 0.4 | **−0.6 ± 0.4** | |
| **11** → ref2 | | **2.3 ± 0.2** | | |
| **15** → ref3 | | **4.8 ± 0.5** | −4.1 ± 0.6 | |
| **19** → ref1 | | **2.9 ± 0.6** | | |
| **21** → ref3 | | **7.3 ± 0.4** | −2.1 ± 0.4 | |
| **23** → ref2 | | **−6.7 ± 0.5** | −13.1 ± 0.6 | |
| **26** → ref2 | | **3.7 ± 0.4** | −12.0 ± 0.4 | |
| **28** → ref2 | | **1.3 ± 0.4** | −1.9 ± 0.4 | |
| **34** → ref2 | | **−0.2 ± 0.7** | −3.7 ± 0.7 | |
| **61S** → ref2 | | **−4.8 ± 0.8** | | |
| **61R** → ref2 | | −19.8 ± 0.4 | | |
| ref 2 → ref1 | | **−11.5 ± 0.6** | | |
| ref 3 → ref2 | | **4.5 ± 0.4** | | |

Experimental data [23] for the relative energies are also given for the transformations that do not involve any reference ligands

molecule (shown in bold face in Table 7). The average dihedral angles observed during the FES simulations and in the docked structures are shown in Table S3. In most structures, the ring systems were not coplanar.

Ligand **61** has two stereoisomers, depending on the orientation of the hydroxyl and methyl groups. We tested both and found the S form to bind more favourably than the R form. This is in striking contrast to the docking calculations, which indicated that only the R form bound to the protein. Experimentally, ligand **61** (racemic mixture) was found to be a non-binder.

For set 2, no reference ligands were employed and therefore, we can directly compare the results of the three studied perturbations with experimental relative affinities. From the results in Table 7, it can be seen that the two results employing the pose in the 2WI7 crystal structure, but using either the 2WI7 or the 3FT5 crystal structures gave similar relative affinities. Therefore, only one of these results is compared with experiments in Table 8. It can be seen that the results were poor with a strongly negative correlation ($R = -0.8$), an incorrect sign for two of the perturbations ($\tau_r = -0.3$, although the sign of one of the experimental relative affinities in not statistically significant), and a MAD of 14 kJ/mol. However, the results based on the 3FT5 crystal structure were much better with a positive correlation ($R = 0.6$), a correct sign of two of the perturbations (those that have statistically significant experimental differences) and a MAD of 5 kJ/mol. The results of the docking and MM/GBSA calculations (for the same *relative* affinities, also shown in Table 8) were much

**Table 8** Performance of the various methods to calculate relative binding free energies (MAD and maximum error, Max, in kJ/mol) compared to experimental results [23]

| | GScore | MM/GBSA | QM/MM | FES | |
|---|---|---|---|---|---|
| *Set 1* | | | | | |
| MAD | 5.8–6.1 | 20.8–26.3 | 17.6–29.1 | 10.9–15.9 | |
| *R* | −0.58 to 0.03 | −0.69 to −0.60 | −0.42 to −0.01 | −0.80 to −0.54 | |
| τ | −1.00 to −0.40 | −0.43 to 0.00 | −0.14 to 0.50 | −1.00 to −0.71 | |
| Max | 10.2 | 32.2–50.4 | 33.3–66.8 | 23.3 | |
| | | | | 2WI7 | 3FT5 |
| *Set 2 without Wat1/2* | | | | | |
| MAD | 7.7 | 27.8 | 12.9 | 14.2 ± 1.0 | 5.3 ± 0.8 |
| *R* | −0.57 | −0.81 | 0.43 | −0.81 ± 0.07 | 0.59 ± 0.10 |
| τ | −1.00 | −1.00 | −0.33 | −0.33 ± 0.33 | 0.33 ± 0.48 |
| Max | 16.4 | 45.6 | 19.9 | 26.0 ± 1.8 | 11.2 ± 1.8 |
| | | | | 2WI7 | 3FT5 |
| *Set 2 with Wat1/2* | | | | | |
| MAD | 6.1 | 41.0 | | 4.8 ± 1.3 | 3.7 ± 1.3 |
| *R* | 0.49 | −0.58 | | 1.00 ± 0.04 | 1.00 ± 0.04 |
| τ | 0.33 | 0.33 | | 0.33 ± 0.43 | 0.33 ± 0.43 |
| Max | 15.2 | 60.2 | | 6.1 ± 1.8 | 4.1 ± 1.8 |
| *Set 3* | | | | | |
| MAD | 4.7–10.4 | 29.6–56.0 | 23.6–47.1 | 8.7–14.6 | |
| *R* | −0.45 to 0.70 | 0.18 to 0.92 | −0.32 to 0.57 | −0.47 to −0.20 | |
| τ | −0.56 to 0.33 | 0.33 to 0.78 | −0.33 to 0.56 | −0.78 to 0.11 | |
| Max | 8.8–16.9 | 55.4–95.6 | 59.1–88.4 | 17.9–27.9 | |

For set 1, the reported values are the range obtained when doing three comparisons: four relative affinities using ligand **82** as the reference, all seven relative affinities that can be obtained by combining two perturbations, or all ten possible relative affinities of the five ligands. For set 2, we present the results of the three perturbations studied by FES, reporting bootstrapped uncertainties, using the observed standard error for FES. Values in brackets for GScore and MM/GBSA were obtained using the 2WI7 crystal structure. For set 3, we present the range obtained by using either ligands **10**, **11**, **23**, **26**, **28**, or **34** as the reference

worse with $R = -0.6$ and $-0.8$, $\tau_r = -1.0$, and MAD = 8 and 28 kJ/mol, respectively. QM/MM results were of intermediate quality with $R = 0.4$ and MAD = 13 kJ/mol.

Keeping the Wat2 crystal water molecule in the FES calculations improved the results for both crystal structures, giving a perfect correlation ($R = 1.0$) and low MADs (5 kJ/mol for 2WI7 and 4 kJ/mol for 3FT5). In particular, both sets of calculations predicted that ligand **100** has a much lower binding affinity (~10 kJ/mol) than the other three ligands. However, in both cases, one of the three relative affinities had an incorrect sign ($\tau_r = 0.3$), although for the 3FT5 structure this involved the transformation for which the experimental estimate is not statistically significant. These calculations also gave an ideal slope of 1.0, whereas it was 1.2 for the calculations based on the 2WI7 structure. Both FES calculations gave better results than the docking and MM/GBSA calculations including Wat2 ($R = 0.5$ and MAD = 6 kJ/mol for GScore).

For the other two sets of ligands, no direct comparison with experiments [23] can be performed, because all studied perturbations (except one) involved reference ligands with unknown experimental affinities. This means that the calculated results need to be combined to compare with experiments, increasing the uncertainty and making the comparison dependent on which data are combined. Moreover, when calculating the correlation coefficient, the results also depend on the sign of the transformation (i.e. whether the **81** → **82** or **82** → **81** perturbation is considered, for example). The latter problem was solved by always considering both directions of the perturbation when *R* was calculated.

For set 1, it may seem natural to compare with ligand **82**, because all relative affinities can be obtained from this ligand using one or two perturbations. However, three additional relative affinities can be obtained by combining two perturbations and all ten possible relative affinities can

be obtained from three perturbations. Therefore, we give in Table 8 the results of three different comparisons (as ranges): four relative affinities using ligand **82** as the reference, all seven relative affinities that can be obtained by combining two perturbations, and all ten possible relative affinities. Numerically, the results vary somewhat, but all results were poor: the correlation was negative ($R = -0.8$ to $-0.5$), MAD = 11–16 kJ/mol, and $\tau_r = -1.0$ to $-0.7$, i.e. only one relative affinity had the correct sign, but the signs of four of the measured relative affinities are not statistically significant. In fact, the largest error (23 kJ/mol) is obtained for the **81** → **82** transformation that is directly comparable with experiments.

The docking gave a smaller MAD and MM/GBSA and QM/MM larger MADs than FES (6, 21–26, and 18–29 kJ/mol, respectively), owing to a smaller and larger ranges of the absolute affinities compared to experiments, 4, 45, and 62 kJ/mol, respectively, compared to 11 kJ/mol for the experimental data. All three methods showed no or negative correlations ($R = -0.0$ to $-0.7$). Likewise, $\tau_r$ was mostly negative ($-0.1$ to $-1.0$) or zero, except when using ligand **82** as the reference for QM/MM ($\tau_r = 0.5$).

For set 3, the situation is even more complicated: all studied transformations involve at least one of the three reference molecules. Any of ligands **10**, **11**, **23**, **26**, **28**, **34**, and **61** can be individually compared employing two perturbations, whereas ligands **19**, **15**, and **21** require the combination of three perturbations. Table 8 shows the range of results obtained when using any of the six ligands in the first group as the reference (excluding ligand 61, because it is experimentally a non-binder). It can be seen that the FES results were quite poor with a negative correlation ($R = -0.5$ to $-0.2$), a varying $\tau_r$ ($-0.8$ to $+0.1$), a MAD of 9–15 kJ/mol and maximum errors of 18–28 kJ/mol.

From Table 8, it can also be seen that the docked results for set 3 were somewhat better with a positive correlation ($R = 0.3$–$0.7$), except when ligand **11** was used as the reference ($R = -0.5$). The same applies to $\tau_r$, which was positive (0.1–0.3), except when using ligand **11** as the reference ($\tau_r = -0.6$). MAD was appreciably better 5–10 kJ/mol, but this is mainly because all relative energies were underestimated: the range of the affinities was only 7 kJ/mol, whereas the experimental range was at least 19 kJ/mol, and in FEP the range was 21 kJ/mol. The MM/GBSA calculations vastly overestimated the range (124 kJ/mol) and therefore gave a very poor MAD of 30–56 kJ/mol and a maximum error of up to 124 kJ/mol (9–17 kJ/mol for the docking). On the other hand, the correlation was always positive, reaching an impressive $R = 0.9$ when using ligand **26** as the reference. Likewise, $\tau_r$ was better than for the other methods, 0.3–0.8. QM/MM gave quite poor results with both $R$ and $\tau_r = -0.3$ to 0.6 and MAD = 24–47 kJ/mol.

One set of relative affinities was submitted (submission entry 56af858f31db8). It was based on the data in Table 7 for sets 2 (2WI7 structure) and 3, but the data in Table S5 for set 1 (i.e. obtained without the improper ca–hn–nh–hn dihedral angle, giving spurious structures, as discussed above). The data were submitted with ligands **80**, **100**, and **10** as the reference, which increases the uncertainty and may affect the calculated quality estimates. Unfortunately, we selected to submit the set 2 results based on the 2WI7 structure (mainly because the 2WI7/3FT5 results were similar), although it turned out that the 3FT5 reproduced the experimental measurements much better.

## Conclusions

In this study, we have tried to estimate the binding affinities of three sets of ligands (with five, four and ten ligands in each) for HSP90 in the D3R 2015 grand challenge blind-test competition. We have employed four different theoretical methods of varying sophistication: docking with the induced-fit protocol in Glide, MM/GBSA calculations with single minimised structures performed by Prime, a new QM/MM approach, based big-QM calculations with various energy terms added, and standard FES calculations of relative binding affinities.

Unfortunately, the results were quite disappointing, with poor and often negative correlation and $\tau$ values for most of the methods and ligand sets. For set 2, the problem could be traced to the displacement of one or two water molecules by one of the ligands. If this effect was properly accounted for, FES and some docking scores gave good results. We employed GCMC calculations to deduce which water molecules dissociate with the various ligands.

Owing to the poor overall results, it is hard to compare the four methods employed. However, our results show no clear-cut advantage of using the more rigorous method FES approach, which comes with a much higher computational effort. In general, the docking calculations with GScore and IFDScore gave small MADtr for all three sets, 4–8 kJ/mol. However, this primarily reflects that these scores underestimate the differences between the various ligands. The $E_{model}$ score and MM/GBSA gave much higher MADtr (9–29 kJ/mol) and a strong overestimation of the range of the calculated binding affinities.

Compared to the other submissions in this blind-test competition, our calculations gave in general mediocre or poor results [23]. However, QM/MM was one of the few methods that gave a non-negative $\tau$ and a positive correlation for set 2, and without the unfortunate sign error, the correct QM/MM results would have given the best $R$ and $\tau$ among all submissions. For set 3, our docked results gave the lowest RMSD and MM/GBSA gave the best $\tau$ among

all submissions (in fact, our four submissions gave among the five best τ values for set 3). Still, this mainly reflects the large variation in the performance of the results from both us and the other groups; the other submissions also gave rather disappointing overall results: in particular, none of the submissions gave positive τ values for all three sets.

In the new QM/MM method, we first reoptimised the docked structures with standard QM/MM calculations, using a quite large QM system (280–320 atoms), including all atoms within 3 Å of the ligand. Then, the QM system was enlarged with all atoms within 7.5 Å of the ligand (970–1160 atoms) and a single-point energy was calculated in a COSMO continuum solvent (Fig. 2). To the rigid interaction energies calculated with this model, we added five energy corrections (Eq. 4), similar to what has been used for host–guest systems [16, 33, 34]: first, a correction term for increasing basis set for the smaller QM system to quadruple-zeta quality. Second, a DFT-D3 dispersion correction, including third-order terms. Third, a thermostatistical correction, including the zero-point energy and entropy, calculated at the MM level with a free-rotor approximation for the low-lying vibrations. Fourth, a ligand-relaxation energy term, and finally an improved solvation energy for the ligand, estimated by the COSMO-RS approach. We also tried to include the solvation free energy of the whole protein with PB or GB methods, but could not obtain any consistent results.

Unfortunately, the QM/MM affinities, showed no consistent improvement over the docked results, although most hydrogen bonds were shortened. Instead, the QM/MM energies showed a similar overestimation of the differences in the binding affinities as the MM/GBSA method, giving MADtr of 17–30 kJ/mol. Still, the results could consistently be improved for all three sets if the ligand-relaxation and thermostatistical terms are omitted (e.g. MADtr = 6–23 kJ/mol). It is probably necessary to employ more than a single minimised structure to obtain consistent and reliable results with QM/MM.

Clearly, the FES results were disappointing, with MADs of 4–15 kJ/mol and maximum errors of up to 26 kJ/mol. Previous large-scale tests of relative FES affinities have shown that MADs of 2–6 kJ/mol are typically obtained for well-behaving systems [9–11]. Such results were only obtained for set 2 if all water molecules are included. The much larger errors obtained for the other two sets can have several causes. First, some of the perturbations in this study are larger than in the large-scale tests. However, we have thoroughly monitored the overlap, convergence, and precision of the calculations, and there is not indication that the perturbations are too large or that the sampling is too short. On the other hand, HSP90 has a flexible binding site and the simulations are much too short to sample larger conformational changes in the binding site or the whole protein. Second, it is possible that the MM force field is not accurate enough to model the chemical variation of the ligands. However, the set 1 ligands show a rather restricted variation, involving mainly methyl, methoxy, ethoxy, and acetate groups, for which the general Amber force field is expected to perform well.

Third, for all FES calculations, we have assumed that all ligands bind in the same mode as the starting crystal structure. Some differences have been observed between the FES and docked structures and also between the various starting structures. If the binding mode in the crystal structure is incorrect or if the binding mode changes between the various ligands, FES is expected to give poor results, and this would affect also the other calculations, because docked structures were accepted only if they were similar to the crystal structures. We believe that this is the main reason for the poor results in this investigation. It should also be noted that the variable parts of the ligands do not show much interactions with the protein. This means that there is a risk that the ligands may bind in a different conformation and that some residues in the protein may show a large change in conformation (to form interactions with this part of the ligand), or that the binding is mainly determined by the interaction of this part with solvent. Clearly, all ranking methods heavily depend on accurate structures, but unfortunately, crystal structures are lacking for all ligands in this investigation. This makes the present test somewhat less informative when it comes to the ranking of different methods to predict binding affinities. To obtain improved binding-affinity predictions for such complicated systems, FES methods involving enhanced sampling could be tested, e.g. metadynamics, accelerated MD, or replica-exchange methods [109–114]. However, many of them are most effective if it is known beforehand which groups need better sampling, which not always is the case. They also significantly increase the computational effort.

**Acknowledgments** This investigation has been supported by grants from the Swedish research council (project 2014-5540) and from Knut and Alice Wallenberg Foundation (KAW 2013.0022). G. D. is supported by a scholarship from the China Scholarship Council. F. A-C. and C. M-G. acknowledge support from doctoral fellowships CONICYT-PCHA/Folio 21120213 and 21120214, respectively. The computations were performed on computer resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lunarc at Lund University and HPC2N at Umeå University.

## References


1. Gohlke H, Klebe G (2002) Angew Chem Int Ed 41:2644–2676
2. Genheden S, Ryde U (2015) Expert Opin Drug Discov 10:449–461
3. Åqvist L, Luzhkov VB, Brandsdal BO (2002) Acc Chem Res 35:358–365
4. Mobley DL, Klimovich PV (2012) J Chem Phys 137:230901
5. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S et al (2006) J Med Chem 49:5912–5931
6. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) J Chem Inf Model 49:1455–1474
7. Brown SP, Muchmore SW (2009) J Med Chem 52:3159–3165
8. Sun H, Li Y, Tian S, Xu L, Hou T (2014) Phys Chem Chem Phys 16:16719
9. Christ C, Fox T (2014) J Chem Inf Model 54:108–120
10. Mikulskis P, Genheden S, Ryde U (2014) J Chem Inf Model 54:2794–2806
11. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, Murcko M, Frye L, Farid R, Ling T, Mobley DL, Jorgensen WL, Berner BJ, Friesner RA, Abel R (2015) J Am Chem Soc 137:2695–2703
12. Genheden S, Luchko T, Gusarov S, Kovalenko A, Ryde U (2010) J Phys Chem B 114:8505–8516
13. Genheden S, Nilsson I, Ryde U (2011) J Chem Inf Model 51:947–958
14. Mikulskis P, Genheden S, Rydberg P, Sandberg L, Olsen L, Ryde U (2012) J Comput Aided Mol Des 26:527–541
15. Muddana HS, Fenley AT, Mobley DL, Gilson MK (2014) J Comput Aided Mol Des 28:305–317
16. Mikulskis P, Cioloboc D, Andrejic M, Khare S, Brorsson J, Genheden S, Mata RA, Söderhjelm P, Ryde U (2014) J Comput Aided Mol Des 28:375–400
17. Coleman RG, Sterling T, Weiss DR (2014) J Comput Aided Mol Des 28:201–209
18. Naïm M, Bhat S, Rankin KN, Dennis S, Chowdhury SF, Siddiqi I, Drabik P, Sulea T, Bayly CI, Jakalian A, Purisima EO (2007) J Chem Inf Model 47:122–133
19. Muddana HS, Varnado CD, Bielawski CW, Urbach AR, Isaacs L, Geballe MT, Gilson MK (2012) J Comput Aided Mol Des 26:475–487
20. Mobley DL, Liu S, Lim NM, Wymer KL, Perryman AL, Forli S, Deng N, Su J, Branson K, Olson AJ (2014) J Comput Aided Mol Des 28:327–345
21. Hogues H, Sulea T, Purisima EO (2014) J Comput Aided Mol Des 28:417–427
22. Gallicchio E, Deng N, He P, Perryman AL, Santiago DN, Forli S, Olson AJ, Levy RM (2014) J Comput Aided Mol Des 28:475–490
23. Gathiaka S, Liu S, Chiu M, Yang H, Burley SK, Walters WP, Amaro RE, Gilson MK, Feher VA (2016) D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. J Comput Aided Mol Des. doi:10.1007/s10822-016-9946-8
24. Lindquist S, Craig EA (1988) Annu Rev Genet 22:631–677
25. Isaacs JS, Xu W, Neckers L (2003) Cancer Cell 3:213–217
26. Cullinan SB, Whitesell L (2006) Semin Oncol 33:457–465
27. Solit DB, Rosen N (2006) Curr Top Med Chem 6:1205–1214
28. McDonald E, Workman P, Jones K (2006) Curr Top Med Chem 6:1091–1107
29. Huth JR, Park C, Petros AM, Kunzer AR, Wendt MD, Wang X, Lynch CL, Mack JC, Swift KM, Judge RA, Chen J, Richardson PL, Jin S, Tahir SK, Matayoshi ED, Dorwin SA, Ladror US, Severin JM, Walter KA, Bartley DM, Fesik SW, Elmore SW, Hajduk PJ (2007) Chem Biol Drug Des 70:1–12
30. Brunko M, Tahir SK, Song X, Chen J, Ding H, Huth JR, Judge RA, Madar DJ, Park CH, Park CM, Petros AM, Tse C, Rosenberg SH, Elmore SW (2010) Bioorg Med Chem Lett 20:7503–7506
31. Glide, version 6.7, Schrödinger, LLC, New York, NY, 2015
32. Prime, version 4.0, Schrödinger, LLC, New York, NY, 2015
33. Grimme S (2012) Chem Eur J 18:9955–9964
34. Antony J, Sure R, Grimme S (2015) Chem Commun 51:1764–1774
35. Hu L, Söderhjelm P, Ryde U (2013) J Chem Theory Comput 9:640–649
36. Bruncko M, Tahir SK, Song X et al (2010) Bioorganic Med Chem Lett 20:7503–7506
37. Brough PA, Barril X, Borgognoni J et al (2009) J Med Chem 52:4794–4809
38. Barker JJ, Barker O, Boggio R, Chauhan V, Cheng RK, Corden V, Courtney SM, Edwards N, Falque VM, Fusar F, Gardiner M, Hamelin EM, Hesterkamp T, Ichihara O, Jones RS, Mather O, Mercurio C, Minucci S, Montalbetti CA, Muller A, Patel D, Phillips BG, Varasi M, Whittaker M, Winkler D, Yarnold CJ (2009) Chem Med Chem 4:963–966
39. Suda A, Koyano H, Hayase T et al (2012) Bioorganic Med Chem Lett 22:1136–1141
40. Kang YN, Stuckey JA Structure of Heat Shock Protein 90 Bound to CS302. To be published, PDB structure 4YKR
41. Schrödinger Suite 2015-2 Protein Preparation Wizard; Epik version 3.2, Schrödinger, LLC, New York, NY, 2015; Impact version 6.7, Schrödinger, LLC, New York, NY, 2015; Prime version 4.0, Schrödinger, LLC, New York, NY, 2015
42. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH (2011) J Chem Theory Comput 7:525–537
43. Søndergaard CR, Olsson MHM, Rostkowski M, Jensen JH (2011) J Chem Theory Comput 7:2284–2295
44. Sastry GM, Adzhigirey M, Day T et al (2013) J Comput Aided Mol Des 27:221–234
45. Banks JL, Beard HS, Cao Y, Cho AE, Damm W, Farid R, Felts AK, Halgren TA, Mainz DT, Maple JR, Murphy R, Philipp DM, Repasky MP, Zhang LY, Berne BJ, Friesner RA, Gallicchio E, Levy RM (2005) J Comp Chem 26:1752–1780
46. Maestro, version 10.2, Schrödinger, LLC, New York, NY, 2015
47. LigPrep, version 3.4, Schrödinger, LLC, New York, NY, 2015
48. Epik, version 3.2, Schrödinger, LLC, New York, NY, 2015
49. MacroModel, version 10.8, Schrödinger, LLC, New York, NY, 2015
50. Sherman W, Day T, Jacobson MP et al (2006) J Med Chem 49:534–553
51. Schrödinger Suite 2015-2 Induced Fit Docking protocol; Glide version 6.7, Schrödinger, LLC, New York, NY, 2015; Prime version 4.0, Schrödinger, LLC, New York, NY, 2015
52. Li J, Abel R, Zhu K et al (2011) Proteins Struct Funct Genet 79:2794–2812
53. Mulakala C, Viswanadhan VN (2013) J Mol Graph Model 46:41–51
54. Case DA, Berryman JT, Betz RM, Cerutti DS, Cheatham TE III, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee TS, LeGrand S, Li P, Luchko T, Luo R, Madej B, Merz KM, Monard G, Needham P, Nguyen H, Nguyen HT, Omelyan I, Onufriev A, Roe DR, Roitberg A, Salomon-Ferrer R, Simmerling CL, Smith W, Swails J, Walker RC, Wang J, Wolf RM, Wu X, York DM, Kollman PA (2014) AMBER 14. University of California, San Francisco
55. Ryde U (1996) J Comput Aided Mol Des 10:153–164

56. Ryde U, Olsson MHM (2001) Int J Quantum Chem 81:335–347
57. Reuter N, Dejaegere A, Maigret B, Karplus M (2000) J Phys Chem A 104:1720–1735
58. Hu L, Söderhjelm P, Ryde U (2011) J Chem Theory Comput 7:761–777
59. Svensson M, Humbel S, Froese RDJ, Matsubara T, Sieber S, Morokuma K (1996) J Phys Chem 100:19357–19363
60. TURBOMOLE 7.0, 2015, developed by University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2015, TURBOMOLE GmbH; http://www.turbomole.com
61. Tao J, Perdew JP, Staroverov VN, Scuseria GE (2003) Phys Rev Lett 91:146401
62. Schäfer A, Horn H, Ahlrichs R (1992) J Chem Phys 97:2571–2577
63. Grimme S, Antony J, Ehrlich S, Krieg H (2010) J Chem Phys 132:154104
64. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) J Chem Theory Comput 11:3696–3713
65. Hu L, Eliasson J, Heimdal J, Ryde U (2009) J Phys Chem A 113:11793–11800
66. Klamt A, Schüürmann J (1993) J Chem Soc Perkin Trans 5:799–805
67. Schäfer A, Klamt A, Sattel D, Lohrenz JCW, Eckert F (2000) Phys Chem Chem Phys 2:2187–2193
68. Sierka M, Hogekamp A, Ahlrichs R (2003) J Chem Phys 118:9136–9148
69. Grimme S, Ehrlich S, Goerigk L (2011) J Comput Chem 32:1456–1465
70. dftd3 software http://toc.uni-muenster.de/DFTD3/getd3html
71. Weigend F, Ahlrichs R (2005) Phys Chem Chem Phys 7:3297–3305
72. Jensen F (1999) Introduction to computational chemistry. Wiley, Chichester, pp 298–303
73. Kongsted J, Ryde U (2009) J Comput Aided Mol Des 23:63–71
74. Genheden S, Kuhn O, Mikulskis P, Hoffmann D, Ryde U (2012) J Chem Inf Model 52:2079–2088
75. Kaukonen M, Söderhjelm P, Heimdal J, Ryde U (2008) J Phys Chem B 112:12537–12548
76. Klamt A (1995) J Phys Chem 99:2224–2235
77. Eckert F, Klamt A (2002) AIChE J 48:369–385
78. Eckert F, Klamt A (2010) COSMOtherm, Version C30, Release 1301, COSMOlogic GmbH & Co KG, Leverkusen (Germany)
79. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) J Chem Phys 79:926–935
80. Wang JM, Wolf RM, Caldwell KW, Kollman PA, Case DA (2004) J Comput Chem 25:1157–1174
81. Bayly CI, Cieplak P, Cornell WD, Kollman PA (1993) J Phys Chem 97:10269–10280
82. Besler BH, Merz KM, Kollman PA (1990) J Comput Chem 11:431–439
83. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery Jr, JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian, Inc., Wallingford CT
84. Seminario JM (1996) Int J Quantum Chem 60:1271–1277
85. Nilsson K, Lecerof D, Sigfridsson E, Ryde U (2003) Acta Crystallogr D Biol Crystallogr 59:274–289
86. Tembe BL, McCammon JA (1984) J Comp Chem 8:281–283
87. Shirts MR, Chodera JD (2008) J Chem Phys 129:124105
88. Shirts MR, Chodera JD, Python implementation of the multistate Bennett acceptance ratio (MBAR) method; http://github.com/choderalab/pymbar
89. Bennett CH (1976) J Comput Phys 22:245–268
90. Kirkwood JG (1935) J Chem Phys 3:300–313
91. Zwanzig RW (1954) J Chem Phys 22:1420–1426
92. Steinbrecher T, Mobley DL, Case DA (2007) J Chem Phys 127:214108
93. Steinbrecher T, Joung I, Case DA (2011) J Comp Chem 32:3253–3263
94. Kaus JW, Pierce LT, Walker RC, McCammon JA (2013) J Chem Theory Comput 9:4131–4139
95. Wu X, Brooks BR (2003) Phys Lett 381:512–518
96. Berendsen HJC, Postma JPM, van Gunsteren WF, Dinola A, Haak JR (1984) J Chem Phys 81:3684–3690
97. Darden T, York D, Pedersen L (1993) J Chem Phys 98:10089–10092
98. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) J Comput Phys 23:327–341
99. Ross GA, Bodnarchuk MS, Essex JW (2015) J Am Chem Soc 137:14930–14943
100. Bodnarchuk M, Bradshaw R, Cave-Ayland A, Genheden S, Martinez AC, Michel J, Ross GA, Woods CJ, ProtoMS, School of Chemistry, University of Southampton: Southampton, U.K.; www.protoms.org
101. Adams D (1974) J Mol Phys 28:1241
102. Brown SP, Muchmore SW, Hajduk PJ (2009) Drug Discov Today 14:420–427
103. Bhattacharyya A (1943) Bull Cal Math Soc 35:99–109
104. Wu D, Kofke DA (2005) J Chem Phys 123:054103
105. Pohorille A, Jarzynski A, Chipot A (2010) J Chem Phys B 114:10235–10253
106. Chakrabarti P, Bhattacharyya R (2007) Progr Biophys Mol Biol 95:83–137
107. Duan G, Smith VH Jr, Weaver DF (2000) J Phys Chem A 104:4521–4532
108. Caldararu O, Olsson M, Riplinger C, Neese F, Ryde U (2016) J Comput Aided Mol Des (in press)
109. Laio A, Parrinello M (2002) Proc Natl Acad Sci USA 99:12562–12566
110. Woods CJ, Essex JW, King MA (2003) J Phys Chem B 107:13703–13710
111. Hamelberg D, Mongan J, McCammon JA (2004) J Chem Phys 120:11919–11929
112. Liu P, Kim B, Friesner RA, Berne BJ (2005) Proc Natl Acad Sci USA 102:13749–13754
113. Zheng L, Yang W (2012) J Chem Theory Comput 8:810–823
114. Wang L, Friesner RA, Berne BJ (2011) J Phys Chem B 115:9431–9438

# Paper II

Journal of Computer-Aided Molecular Design (2018) 32:1027–1046
https://doi.org/10.1007/s10822-018-0158-2

# Binding free energies in the SAMPL6 octa-acid host–guest challenge calculated with MM and QM methods

**Octav Caldararu[1] · Martin A. Olsson[1] · Majda Misini Ignjatović[1] · Meiting Wang[1] · Ulf Ryde[1]**

Received: 1 June 2018 / Accepted: 31 August 2018 / Published online: 10 September 2018


**Abstract**
We have estimated free energies for the binding of eight carboxylate ligands to two variants of the octa-acid deep-cavity host in the SAMPL6 blind-test challenge (with or without endo methyl groups on the four upper-rim benzoate groups, OAM and OAH, respectively). We employed free-energy perturbation (FEP) for relative binding energies at the molecular mechanics (MM) and the combined quantum mechanical (QM) and MM (QM/MM) levels, the latter obtained with the reference-potential approach with QM/MM sampling for the MM → QM/MM FEP. The semiempirical QM method PM6-DH+ was employed for the ligand in the latter calculations. Moreover, binding free energies were also estimated from QM/MM optimised structures, combined with COSMO-RS estimates of the solvation energy and thermostatistical corrections from MM frequencies. They were performed at the PM6-DH+ level of theory with the full host and guest molecule in the QM system (and also four water molecules in the geometry optimisations) for 10–20 snapshots from molecular dynamics simulations of the complex. Finally, the structure with the lowest free energy was recalculated using the dispersion-corrected density-functional theory method TPSS-D3, for both the structure and the energy. The two FEP approaches gave similar results (PM6-DH+/MM slightly better for OAM), which were among the five submissions with the best performance in the challenge and gave the best results without any fit to data from the SAMPL5 challenge, with mean absolute deviations (MAD) of 2.4–5.2 kJ/mol and a correlation coefficient ($R^2$) of 0.77–0.93. This is the first time QM/MM approaches give binding free energies that are competitive to those obtained with MM for the octa-acid host. The QM/MM-optimised structures gave somewhat worse performance (MAD = 3–8 kJ/mol and $R^2$ = 0.1–0.9), but the results were improved compared to previous studies of this system with similar methods.

**Keywords** Ligand binding · Free-energy perturbation · Reference-potential with QM/MM sampling · Semiempirical methods · Density functional theory · Entropy

## Introduction

Estimating the affinity between a small molecule and a biomacromolecule is important in many parts of chemistry, especially in drug design [1, 2]. Therefore, numerous computational methods have been developed with this aim [1], ranging from simple scoring approaches for ligand docking [3], via end-point approaches, like linear interaction energy [4] and MM/PBSA (molecular mechanics combined with Poisson–Boltzmann and surface area solvation) [5, 6], to strict approaches based on free-energy perturbation (FEP) [7, 8] with free energies calculated by exponential averaging (EA) [9], thermodynamic integration [10] or the Bennett acceptance ratio (BAR) approach [11].

The latter methods should in principle be limited only by the accuracy of the potential-energy function and the sampling of the phase space, although uncertainties in the nature of the simulated system (e.g. the protonation state of all involved molecules and residues) may also affect the results [7, 8]. To reduce the sampling problem and allow for a better control of the actual chemical state, there has been quite some interest to study simpler systems, in particular the binding of small molecules to organic molecules

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10822-018-0158-2) contains supplementary material, which is available to authorized users.

✉ Ulf Ryde
Ulf.Ryde@teokem.lu.se

[1] Department of Theoretical Chemistry, Lund University, Chemical Centre, P. O. Box 124, 221 00 Lund, Sweden

of intermediate size (a few hundred atoms), i.e. host–guest systems [12, 13].

Most free energy simulations are performed by empirical potentials in the form of molecular mechanics (MM) force fields. However, during the latest decades, there has been an increasing interest in employing quantum mechanical (QM) calculations to obtain more accurate binding free energies [14]. Such calculations can be performed at many levels of approximation. Owing to the much larger computational cost of QM calculations, most such studies are based either on single-point QM calculations on structures obtained by MM sampling or on structures minimised by QM [14–16] or by combined QM and MM calculations (QM/MM) [17, 18]. Only a few studies have involved sampling at the QM/MM level, typically using a semiempirical QM (SQM) method [19–23].

Most computational studies of ligand-binding affinities are performed on systems for which the experimental affinities are known. Of course, this introduces the risk that the results are biased towards the experimental data. Therefore, prospective studies, in which the experimental results are not known when the calculations are performed, provide a more unbiased view of the performance of various methods. In this regard, the statistical assessment of the modelling of proteins and ligands (SAMPL) blind-test competitions have been invaluable to compare the true predictive value of various computational methods. Since SAMPL3, it has involved host–guest systems [24] and since SAMPL4, it has involved the binding of ligands to the octa-acid deep-cavity host (OAH, shown in Fig. 1) [25, 26], developed by the Gibb group [27, 28].

In a series of publications, we have studied the binding of nine cyclic carboxylate guest molecules to OAH with computational methods at both the MM and QM/MM level [15, 22, 23, 29–31]. In the SAMPL4 challenge, we used FEP to calculate the relative affinities of the nine guests at the MM level [15], which gave the best results in the competition [25], with a mean absolute deviation (MAD) of $3.6 \pm 0.2$ kJ/mol and a correlation ($R^2$) to experimental data of $0.84 \pm 0.04$. We also tried to improve the FEP results by performing QM calculations with density functional theory (DFT) on snapshots from the MM simulations, using large QM systems involving ~ 310 atoms (1800 DFT calculations for each ligand). However, the difference between the MM and QM potentials was so large that no converged results could be obtained. Therefore, the results were very poor with an uncertainty of 6–32 kJ/mol and MADs of 17–27 kJ/mol.

However, by using smaller QM systems (less water molecules and with the acidic groups on OAH removed) and SQM calculations with the PM6-DH2X method, we were able to obtain converged results with a precision of 1 kJ/mol for all relative free energies, using 700,000 QM calculations for each ligand [29]. Unfortunately, the results were still worse than the MM-FEP results, with a MAD of $4.9 \pm 0.4$ kJ/mol and a vanishing correlation. These results were obtained without any sampling at the QM/MM level, but in our next study such sampling was performed (with semiempirical PM6-DH+ calculations and only the ligand in the QM system) [22]. This gave even better results with a precision of 0.5–0.9 kJ/mol, a MAD of $4.7 \pm 0.2$ and a $R^2$ correlation of $0.86 \pm 0.04$. Recently, we have shown that similar results can be obtained with approximately a fourth of the computational effort using multiple short QM/MM simulations [23] or by using non-equilibrium simulations and Jarzynski's equality [32–34].

In the SAMPL4 study, we also tried to estimate OAH binding affinities with minimised QM structures, using a variant of the method suggested by Grimme and coworkers [15, 16]. We optimised the structures of the complexes with three different DFT approaches (in vacuum, in a continuum solvent and in a continuum solvent with four explicit water molecules). Then, binding free energies were obtained with a vacuum DFT calculation with large basis set and empirical dispersion corrections, combined with a COSMO-RS estimation of the solvation free energy and with thermostatistical corrections from a frequency calculation at the MM level. This approach gave absolute binding affinities of an intermediate accuracy with MADs of 7–14 kJ/mol and $R^2$ of 0.60–0.78. After removing systematic errors (the mean signed deviation, MSD), the MADs (called MADtr in the following) were 5–9 kJ/mol. Similar results were obtained also by Sure and Grimme on the same system [35]. An attempt to improve the energies by local coupled-cluster calculations gave much worse results with $R^2$, MAD and MADtr of 0.28, 37 and 14 kJ/mol, respectively [15, 30].

In SAMPL5, we employed a similar approach to calculate binding affinities of six more diverse guest molecules (with either a carboxylate or a trimethylammonium group) [31] to OAH and also to its tetra-endo-methyl variant (OAM) [36]. The calculations were improved by keeping the structures as symmetric as possible, reducing the charge and flexibility of the ligand and performing a restricted sampling of the complexes. Disappointingly, the results were worse than for SAMPL4 with MADtr of 11–22 kJ/mol and $R^2$ below 0.30. The reason for this is probably the larger diversity of the ligands but also problems with some of the geometry optimisations (the guest carboxylate groups become too buried inside the host). The results were not improved by employing DLPNO–CCSD(T) calculations [37] (MADtr = 16–20 kJ/mol and $R^2 = 0$–$0.15$). The best results in the SAMPL5 competition were obtained for free-energy simulations at the MM level, dragging the ligand out of the host [26].

In this paper, we study the binding of eight carboxylic ligands to both the OAH and OAM hosts in the SAMPL6 challenge [38, 39] with four different methods: FEP at the MM level, FEP at the PM6-DH+/MM level, as well

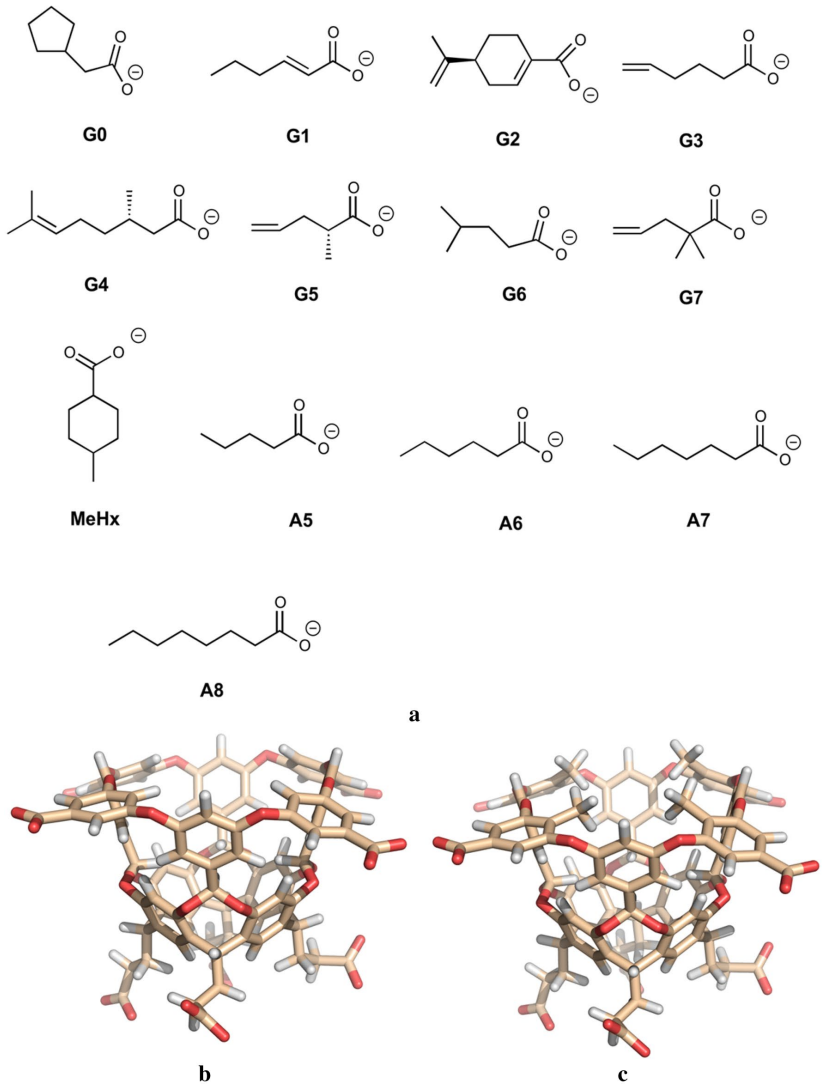



**Fig. 1** **a** Ligands involved in SAMPL6 challenge (G0–G7) and five ligands added to make the perturbations smaller and connected to experimental data (MeHx) and A5–A8. **b** The OAH and **c** OAM host molecules

as optimised structures at the PM6-DH+/MM and DFT/MM levels of theory. For the latter, we used more extensive sampling at the MM level and QM/MM optimised structures with explicit solvent. We also re-examine the SAMPL4 and 5 test cases with the third method to show that it gives improved results. For the first time, we get a

slight improvement of the MM FEP results by employing the PM6-DH+/MM correction.

## Methods

### Setup of the studied systems

We have considered the eight ligands of the SAMPL6 blind-test competition, G0–G7 [38, 39], as well as four aliphatic carboxylates with 5–8 carbon atoms (called A5–A8) and the MeHx ligand from SAMPL4 [40]. They are shown in Fig. 1, together with the OAH and OAM host structures. For some test calculations, we used also the nine cyclic carboxylate ligands from SAMPL4 [40] and the four carboxylate ligands from SAMPL5 [41] (shown in Figures S1 and S2).

The host–guest complexes for the calculations were built from the coordinates for the octa-acid host with the guest molecules from previous blind-prediction challenges [15, 31]. The guest molecules were prepared and modified using the Avogadro software [42] and the geometry of the guest molecules was optimised with the UFF force field [43]. The OAM was constructed by adding four methyl groups at the corresponding hydrogen positions on the upper rim of OAH.

Both the host and the guest molecules were treated with the general AMBER force field (GAFF) [44], whereas the TIP3P model was used for water molecules [45]. Charges for the two host molecules have been reported before [15, 31]. Charges for the ligands were obtained with the same restrained electrostatic potential approach [46]: The molecules were optimised with the SQM AM1 method [47], followed by a single-point calculation at the Hartree–Fock/6-31G* level to obtain the electrostatic potentials, sampled with the Merz–Kollman scheme [48], but at a higher-than-default density (10 layers with 17 points per unit area, giving ~2000 points per atom). These calculations were performed with the Gaussian 09 software [49]. The potentials were then used by antechamber to fit the charges. The charges and atom types of all ligands are given in Table S8 in the supplementary material.

A few parameters missing in the force field were estimated with the Seminario approach [50]: The geometry of the ligands was optimised at TPSS/def2-SV(P) level [51, 52], followed by a frequency calculation using the aoforce module of the Turbomole software [53]. From the resulting Hessian matrix, parameters for the missing dihedrals were extracted with the Hess2FF program [54]. These parameters are given in Table S1 in the supplementary material.

### Molecular dynamics simulation

All molecular dynamics (MD) simulations and FEP calculations were run with the AMBER 16 software suite [55]. Each host–guest complex was solvated in an octahedral box of water molecules extending at least 10 Å from the guest molecules using the tleap module, so that 1504–1513 water molecules were included in the simulations. All nine carboxylic groups on the host and guest molecules were assumed to be deprotonated because the binding affinities were measured at a pH of 11.7 [39]. Thus, the net charge of the host–guest complexes were −9. No counter ions were used in the simulations, as our previous studies have shown that they have a small effect on the calculated free energies [15].

Each complex was subjected to 10,000 steps of conjugate-gradient minimisation, followed by 20 ps constant-volume equilibration and 20 ps constant-pressure equilibration, all performed with heavy non-water atoms restrained towards the starting structure with a force constant of 209 kJ/mol/Å$^2$. Finally, the system was equilibrated for 2 ns without any restraints and with constant pressure, followed by 10 ns of production simulation, during which coordinates were saved every 10 ps. For each host–guest complex, 10 (OAH) or 20 (OAM) independent simulations were run, employing different TIP3P solvation boxes and different starting velocities [56]. Consequently, the total simulation time for each complex was 100 or 200 ns.

All bonds involving hydrogen atoms were constrained to the equilibrium value using the SHAKE algorithm [57], allowing for a time step of 2 fs. The temperature was kept constant at 300 K using Langevin dynamics [58], with a collision frequency of 2 ps$^{-1}$. The pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm [59] with a relaxation time of 1 ps. Long-range electrostatics were handled by particle-mesh Ewald summation [60] with a fourth-order B spline interpolation and a tolerance of $10^{-5}$. The cut-off radius for Lennard−Jones interactions was set to 8 Å.

### FEPs

The guest molecules were manually mapped for the FEP simulations as is shown in Fig. 2, keeping the perturbations as small as possible. To this aim and to connect the relative FEP calculations to experimental data [40, 61], we included also the A5–A8 and MeHx ligands. The FEP simulations were run with the pmemd module of AMBER 16 [37], using the dual-topology scheme with both ligands in the topology file. Each ligand transformation was divided into 13 steps, employing a linear transformation of the force-field potentials with the coupling parameter $\lambda = 0.00, 0.05, 0.10, 0.20, \ldots, 0.80, 0.90, 0.95$ and 1.00. Electrostatic and van der Waals interactions were perturbed concomitantly, using soft-core potentials for both types of interactions [62, 63]. For the simplest perturbations, involving a H → $CH_3$ perturbation (A5 → A6, A6 → A7, A7 → A8, G5 → G7 and G6 → A5), soft-core potentials were used only for the differing atoms. However, for the other seven perturbations,

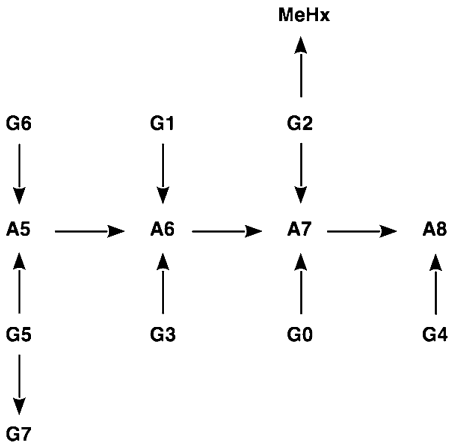



**Fig. 2** Ligand alchemical transformations studied with FEP

soft-core potentials were used for all atoms in the ligands, except those in the carboxylate group, to allow for larger differences in the dynamics of the perturbed groups (atoms without soft-core potentials have identical coordinates in the perturbations).

For each λ value, 100 steps of conjugate-gradient minimisation were performed with the heavy atoms of the host and ligand restrained towards the starting structure with a force constant of 418 kJ/mol/Å$^2$. This was followed by 20 ps constant-volume equilibration with the same restraints and 1 ns constant-pressure equilibration without any restraints. Finally, a 2-ns production simulation was run (still with constant pressure), during which structures and energies were sampled every 2 ps.

Relative binding free energies between two ligands, $L_0$ and $L_1$ ($\Delta\Delta G_{\text{bind}}$), were estimated using a thermodynamic cycle that relates $\Delta\Delta G_{\text{bind}}$ to the free energy of alchemically transforming $L_0$ into $L_1$ when they were either bound to the host, $\Delta G_{\text{bound}}(L_0 \rightarrow L_1)$, or free in solution, $\Delta G_{\text{free}}(L_0 \rightarrow L_1)$ [64, 65]:

$$\begin{aligned}\Delta\Delta G_{\text{bind}} &= \Delta G_{\text{bind}}(L_1) - \Delta G_{\text{bind}}(L_0) \\ &= \Delta G_{\text{bound}}(L_0 \rightarrow L_1) - \Delta G_{\text{free}}(L_0 \rightarrow L_1)\end{aligned} \quad (1)$$

$\Delta G_{\text{bound}}(L_0 \rightarrow L_1)$ and $\Delta G_{\text{free}}(L_0 \rightarrow L_1)$ were estimated by the multi-state Bennett acceptance-ratio (MBAR) method, using the pyMBAR software [66], including only statistically non-correlated energies in the calculations. All FEP calculations were repeated three times using different TIP3P solvation boxes and different starting velocities [56]. Reported free energies are the average over these three calculations, whereas the reported uncertainty is either the standard deviation over these three calculations divided by the square root of three or the square root of the sum of the variances of the three individual estimates divided by three, depending on which of the two values was largest.

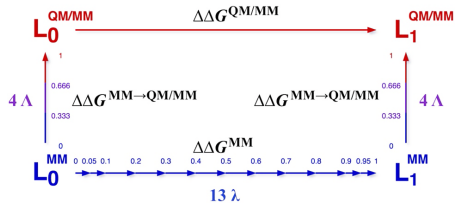



**Fig. 3** Thermodynamic cycle used for the RPQS calculations

## QM/MM FEP calculations

Relative QM/MM binding affinities between two ligands, $L_0$ and $L_1$, were estimated by the reference-potential method with QM/MM sampling (RPQS) [22, 23]. In this approach, the $\Delta\Delta G_{\text{bind}}$ free energies, calculated at the MM level, as described in the previous section, are corrected by a FEP calculation for each ligand in the method space, from the MM potential to the QM/MM potential, as is shown by the thermodynamic cycle in Fig. 3. This was done both for the ligand bound to the host and when free in solution. For each state ($s$ = bound or free), the QM/MM corrected free energy was calculated from

$$\Delta G^{\text{QM/MM}}_{L_0\rightarrow L_1,s} = \Delta G^{\text{MM}}_{L_0\rightarrow L_1,s} - \Delta G^{\text{MM}\rightarrow\text{QM/MM}}_{L_0,s} + \Delta G^{\text{MM}\rightarrow\text{QM/MM}}_{L_1,s}. \quad (2)$$

Finally, the net binding free energies were calculated from

$$\Delta\Delta G^{\text{QM/MM}}_{\text{bind}} = \Delta G^{\text{QM/MM}}_{L_0\rightarrow L_1,\text{bound}} - \Delta G^{\text{QM/MM}}_{L_0\rightarrow L_1,\text{free}} \quad (3)$$

All MM → QM/MM FEP simulations were performed with the AMBER 16 software [55] and for all host–guest systems shown in Fig. 1 except the MeHx ligand. In the QM/MM calculations, only the guest molecule was included in the QM region and it was treated at the SQM PM6-DH+ level of theory [67–69]. The MM → QM/MM free energies were calculated based on the energy function $E(\Lambda) = (1-\Lambda)E_{\text{MM}} + \Lambda E_{\text{QM/MM}}$, where $E_{\text{MM}}$ is the MM energy, $E_{\text{QM/MM}}$ is the QM/MM energy and Λ is a coupling parameter going from 0 to 1. Based on our previous study of OAH with the SAMPL4 ligands [22], we performed calculations at four Λ values: 0.0, 0.333, 0.666, and 1.0. If the overlap with four Λ values was unsatisfactory (see below),

additional Λ values were employed (0.1667, 0.5 or 0.8333; cf. Table S7).

For each Λ value, we performed 100 steps of conjugate-gradient minimisation with the heavy atoms of the host and guest molecules restrained towards the starting structure with a force constant of 418 kJ/mol/Å$^2$. This was followed by 20 ps constant-volume equilibration with the same restraints and 0.5 ns constant-pressure equilibration without any restraints. Finally, a 1 ns production simulation was run, during which structures and energies were sampled every 1 ps. The QM/MM MD simulations were performed in the same manner as described in the "Molecular dynamics simulations" section. Free energies were estimated by MBAR, using the pyMBAR software [66], including only statistically non-correlated energies in the calculations.

## Absolute binding free energies from QM/MM minimised structures

Absolute binding free energies were calculated using the method suggested by Grimme [16, 70], in which the binding free energy is composed of three terms:

$$\Delta G_{\text{tot}} = \Delta E_{\text{QM}} + \Delta G_{\text{solv}} + \Delta G_{\text{therm}} \quad (4)$$

where $\Delta E_{\text{QM}}$ is a single-point vacuum QM energy, which also includes the dispersion energy, $\Delta G_{\text{solv}}$ is the solvation free energy and $\Delta G_{\text{therm}}$ is a thermostatistical correction term. The binding affinity was obtained as the difference in this free energy between the complex, host and guest:

$$\Delta G_{\text{bind}} = \Delta G_{\text{tot}}(\text{complex}) - \Delta G_{\text{tot}}(\text{host}) - \Delta G_{\text{tot}}(\text{guest}) \quad (5)$$

Structures for the free host and guest molecules were taken from the structures of the complexes, without further optimisation (rigid binding free energies; some tests were performed to calculate structure-relaxation energies, but they did not lead to any improvement). The calculations were performed at two levels of QM theory and based on two sets of structures. The two approaches will be called SQM and DFT in the following.

From each of the independent MD simulations of the host–guest complexes at the MM level, the last snapshot was minimised at the PM6-DH+/MM level of theory [67–69] using the AMBER 16 software suite [55]. The quantum system consisted of the host and guest molecules, as well as four water molecules that formed hydrogen bonds with the guest (viz. the two closest water molecules to each of the ligand carboxylate oxygen atoms). It had a net-charge of −9. The solvation box from the MD simulations was kept in all calculations. Conjugate-gradient minimisations were run for 2000 steps without any bond-length restraint for any molecule and with no periodicity (for technical reasons). This gave 10 different host–guest structures for each guest bound to the OAH host and 20 different structures for the OAM host. The resulting structures were used directly for the SQM calculations.

The QM energy for the SQM structures (only isolated host and guest, with all water molecules removed) was calculated as a PM6-DH+ single-point energy using the AMBER sqm program [55]. This method includes dispersion and hydrogen-bond corrections [67–69] and is among the best SQM methods available in the Amber software.

Solvation free energies in water solution were calculated with the conductor-like solvent model (COSMO) [71, 72] real-solvent (COSMO-RS) approach [73, 74] using the COSMOTHERM software [75]. These calculations were based on two single-point BP86 calculations [76, 77] with the TZVP basis set [78], one performed in a vacuum and the other in the COSMO solvent with an infinite dielectric constant. Owing to the extensive negative charge of the hosts, we had to use the undocumented ADEG option to force the program to accept that the solvation energy is very large.

Thermal corrections to the Gibbs free energy at 298 K and 1 atm pressure ($\Delta G_{\text{therm}}$), including zero-point vibrational energy, entropy and enthalpy corrections, were calculated by an ideal-gas rigid-rotor harmonic-oscillator approach [79] from vibrational frequencies calculated at the MM level (i.e. with the GAFF force field and the same charges as in the MD simulations). The frequency calculations were preceded by a minimisation at the same level of theory. To obtain more stable results, low-lying vibrational modes were treated by the free-rotor approximation, using the interpolation model suggested by Grimme with $\omega_0 = 100\ \text{cm}^{-1}$ [16]. The translational entropy was corrected by 7.99 kJ/mol for the change in the standard state from 1 atm to 1 M (used in the experiments [39]). Unfortunately, we discovered after the submission of the results that for most complexes, the ligand dissociated from the host during the MM minimisation before the frequency calculations. Therefore, the thermal corrections were recalculated after the submission, using a restraint to the starting structure during the geometry optimisation.

For the SQM calculations, energies obtained according to Eqs. (4 and 5) were calculated for all 10 or 20 snapshots and the final absolute $\Delta G_{\text{bind}}$ energy was obtained by either taking the minimum value, the average value or the Boltzmann-weighted average value.

In the second (DFT) approach, the structure with the most favourable SQM $\Delta G_{\text{bind}}$ energy was further optimised at the QM/MM level with the host, the ligand and four water molecules in the QM system, treated with the TPSS-D3/def2-SV(P) method [51, 52]. These calculations were performed with the ComQum program [80, 81], which is an interface between AMBER [55] and the QM software Turbomole software [53]. In these calculations, the MM system was kept fixed. The minimisations were run until the

energy change between two iterations was less than 0.003 kJ/mol and the maximum norm of the Cartesian gradients was below $10^{-3}$ a.u. All complexes converged within 150 geometry iterations.

For the optimised structures, $\Delta E_{QM}$ was calculated with the TPSS functional and the def2-QZVP' basis set (the def2-QZVP basis set [52] with the *f*-type functions on hydrogen and the *g*-type functions on the other atoms deleted). The dispersion energy was included using the DFT-D3 approach [82] with Becke–Johnson damping [83] and third-order terms included. All DFT calculations were sped up by expanding the Coulomb interactions in auxiliary basis sets with the resolution-of-identity approximation (RI), using the corresponding auxiliary basis sets [84, 85]. The multipole-accelerated resolution-of-identity J approach was also employed [86]. All DFT calculations were performed using the Turbomole 7.1 or 7.2 software [53]. Finally, absolute $\Delta G_{bind}$ energies were obtained with Eqs. (4 and 5), using the same approach to get $\Delta G_{solv}$ and $\Delta G_{therm}$ as for the SQM structures. However, the final $\Delta G_{bind}$ was based on a single DFT structure.

### Geometric measures

We have used several geometric measures [15] to analyse the structures of the host–guest systems, as described below. Atom names used in the descriptions are shown in Figure S3.

- $r_{Dm}$ measures how deep the guest is inside the host; it is defined as the closest distance between any guest atom and the average of the coordinates of the four HD atoms at the bottom of the host (AD).
- $\alpha_T$ shows the orientation of the ligand inside the host and is defined as the angle between the guest C1–C2 vector (C1 is the carboxylate carbon and C2 is the carbon bound to C1) and the host AD–AB vector, where AB is the averaged coordinates of the four HB atoms on the top of the host.
- $r_{O1}$ and $r_{O2}$ describe how much the guest reaches out of the host. They are the distance between ligand carboxylate O1 or O2 atoms and the average plane defined by the four CC atoms of the host. $\Delta r_O = |r_{O1} - r_{O2}|$ and shows how tilted the carboxylate group is relative to the host.
- $\Delta r_{BB}$ is defined as the difference in distance between two sets of opposite HB host atoms and measures the distortion of the host.
- $r_{C1}$ and $r_{C2}$ describe the orientation of the host carboxylate groups. They are defined as the distance between two opposite CO atoms. $r_{Cav}$ is the average of $r_{C1}$ and $r_{C2}$.

### Error estimates, quality and overlap measures

All reported uncertainties are standard errors of the mean (standard deviations divided by the square root of the number of samples). The uncertainty of the MBAR free energies calculated at each $\lambda$ or $\Lambda$ value was estimated by bootstrapping using the PYMBAR software [66] and the total uncertainty was obtained by error propagation.

The performance of the free-energy estimates was quantified by the mean signed deviation (MSD), the mean absolute deviation (MAD), the MAD after removal of the MSD (MADtr), the root-mean-square deviation (RMSD), the maximum error (Max), the correlation coefficient ($R^2$), the slope of the best correlation line and Kendall's rank correlation coefficient ($\tau$) compared to experimental data [39]. For relative affinities, $\tau$ was calculated only for the transformations that were explicitly studied, not for all combinations that can be formed from these transformations (this is marked by calling it $\tau_r$). Moreover, it was also evaluated considering only differences (both experimental and calculated) that are statistically significant at the 90% level ($\tau_{90}$ and $\tau_{r,90}$ for absolute and relative affinities, respectively) [87]. Note that $R^2$ and the slope for relative affinities depend on the direction of the perturbation (i.e. whether $L_0 \rightarrow L_1$ or $L_1 \rightarrow L_0$ is considered, which is arbitrary). This was solved by considering both directions (both forward and backward) for all perturbations when these two measures were calculated.

The standard deviation of the quality measures was obtained by a simple simulation approach [88]. For each transformation, 1000 Gaussian-distributed random numbers were generated with the mean and standard deviation equal to the MBAR and experimental results [39] for that transformation. Then, the quality measures were calculated for each of these 1000 sets of simulated results and the standard error over the 1000 sets is reported as the uncertainty.

For all $\lambda$ and $\Lambda$ values of all FEP calculations, we have monitored five overlap measures, to ensure that the overlap of the studied distributions is satisfactory, viz. the Bhattacharyya coefficient $\Omega$ [89], the Wu and Kofke overlap measures of the energy probability distributions ($K_{AB}$) [90] and their bias metrics ($\Pi$) [90], the weight of the maximum term in the exponential average ($w_{max}$) [91] and the difference of the forward and backward exponential average estimate ($\Delta\Delta G_{EA}$) [92]. If $\Pi < 0$ or two of the following criteria were not fulfilled: $\Omega > 0.7$, $K_{AB} > 0.7$, $\Pi > 0.5$, $w_{max} < 0.3$, $\Delta\Delta G_{EA} < 4$ kJ/mol, additional $\lambda$ or $\Lambda$ values were included. Overlap measures obtained in the various simulations are listed in Table S7.

## Result and discussion

In this study, we have calculated the free energies for the binding of the eight ligands G0–G7 in Fig. 1 to the normal (OAH) and methylated (OAM) deep-cavity octa-acid hosts (also shown in Fig. 1) within the SAMPL6 blind-test challenge [38, 39]. Thus, the experimental data were not known when the calculations were performed and were revealed only after the predictions were submitted. The experimental affinities were measured by isothermal calorimetry in aqueous 10 mM sodium phosphate buffer at pH 11.7 and 298 K [39]. We employed four different methods and submitted four data sets. The methods are summarised in Fig. 4. First, we performed standard relative FEP calculations at the MM level. Second, we performed QM/MM-FEP calculations using the reference-potential approach with explicit QM/MM sampling (RPQS) [22] at the PM6-DH+ level of theory (only ligand treated by QM). Third, absolute binding free energies were estimated by PM6-DH+/MM optimisations on 10–20 snapshots from MD simulations (host, ligand and 4 water molecules treated by QM), with energies supplemented by continuum solvation and thermostatistical corrections (Eq. 4). Fourth, the most energetically favourable of the latter structures were reoptimised with DFT and energies were calculated with DFT and large basis sets. The results are described below in separate subsections.

### FEP calculations at the MM level

We have calculated the relative binding free energies of the SAMPL6 ligands G0–G7 by FEP calculations at the MM level. As can be seen in Fig. 1, the eight ligands contain a carboxylic group and six to ten carbon atoms. G0 and G2 involve a five- or six-membered ring and all except G0 and G6 have one or two double bonds. Ligands G2, G4 and G5 are chiral and they were used in the isomers shown in Fig. 1 (since the host is achiral, the actual form should not matter for the binding affinities).

We developed a FEP scheme, shown in Fig. 2, in which the eight ligands are connected, keeping the change as small as possible. This was partly accomplished by adding four extra ligands, which are the aliphatic carboxylates with five to eight carbon atoms, A5–A8. Thereby, the perturbations are restricted to the introduction of a double bound, the conversion of a H atom to a methyl group, the closure of a ring, or in one case (G2), formation of a cyclohexene ring by the addition of two carbon atoms. The aliphatic ligands were employed also because experimental binding affinities are

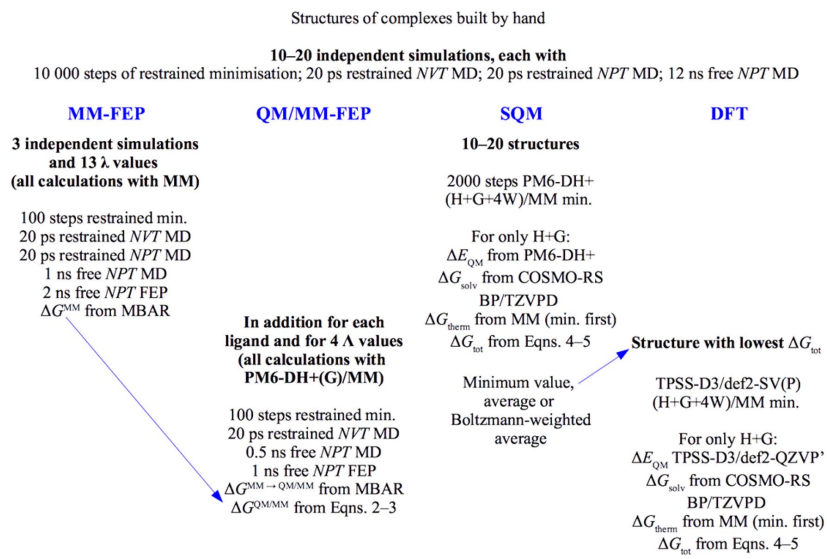



**Fig. 4** Schematic description of the four different approaches employed. H, G and 4W represent the molecules included in the QM system: host, guest and four water molecules

**Table 1** Calculated relative binding free energies (kJ/mol) for the OAH and OAM hosts, obtained with FEP at the MM and PM6-DH+/MM levels for the perturbation scheme in Fig. 2

| | | $\Delta\Delta G_{bind}^{MM}$ | | $\Delta\Delta G_{bind}^{QM/MM}$ | |
|---|---|---|---|---|---|
| | | OAH | OAM | OAH | OAM |
| A5 → A6 | | −14.4 ± 0.1 | −16.1 ± 0.3 | −14.2 ± 0.5 | −17.7 ± 0.5 |
| A6 → A7 | | −4.3 ± 0.2 | −7.7 ± 0.2 | −6.2 ± 0.5 | −8.6 ± 0.5 |
| | Entire ligand | 0.1 ± 0.8 | | −1.8 ± 0.9 | |
| A7 → A8 | | −7.1 ± 0.3 | −8.9 ± 0.1 | −7.2 ± 0.5 | −8.1 ± 0.5 |
| | Entire ligand | −7.9 ± 0.8 | | −8.0 ± 1.0 | |
| G0 → A7 | | −1.8 ± 0.4 | −9.0 ± 0.5 | −2.1 ± 0.6 | −5.8 ± 0.7 |
| G1 → A6 | | −5.8 ± 0.3 | −2.3 ± 0.3 | −11.2 ± 0.5 | −8.0 ± 1.0 |
| G2 → A7 | | 16.8 ± 0.4 | 7.7 ± 0.4 | 6.1 ± 0.5 | 1.3 ± 1.3 |
| G2 → MeHx | | −4.2 ± 1.5 | 4.3 ± 2.6 | | |
| G3 → A6 | | −8.7 ± 0.4 | −10.9 ± 0.4 | −7.0 ± 0.5 | −8.9 ± 0.6 |
| G4 → A8 | | 3.9 ± 2.1 | 3.6 ± 1.9 | 1.2 ± 2.2 | 3.4 ± 1.9 |
| G5 → A5 | | 2.9 ± 0.3 | −1.6 ± 0.4 | 2.1 ± 0.6 | −1.2 ± 0.5 |
| G5 → G7 | | −10.2 ± 0.2 | −6.9 ± 0.3 | −6.4 ± 0.5 | −4.3 ± 0.6 |
| G6 → A5 | | 9.4 ± 0.2 | 9.3 ± 0.2 | 8.4 ± 0.5 | 9.1 ± 0.5 |
| A6@OAM → A6@OAH | | 3.6 ± 0.4 | | 5.2 ± 0.6 | |

available for A6 and A8 to OAH [61], giving us the opportunity to convert the relative energies to absolute affinities. For the same reason, the MeHx ligand from SAMPL4 (shown in Fig. 1a) was also added and connected to G2. To connect the calculations of OAH and OAM, and to obtain absolute affinities for the OAM ligands, we also converted OAH to OAM with and without the A6 ligand bound (adding the four methyl groups at the same time).

The calculated relative affinities are listed in Table 1 ($\Delta\Delta G_{bind}^{MM}$ column; free energies calculated with MBAR). It can be seen that the statistical uncertainty for most $\Delta\Delta G_{bind}$ estimates is low, 0.1–0.5 kJ/mol, owing to the use of three independent FEP calculations. This reflects that the three estimates give similar results, with a variation of up to 1.0 kJ/mol for OAH and 1.6 kJ/mol for OAM. However, for two perturbations with both hosts, G2 → MeHx and G4 → A8, the variation is much larger (9–20 kJ/mol) and therefore the precision is much worse, 1.5–2.6 kJ/mol even if we employed six independent simulations for these perturbations.

Besides the G5 → G7 perturbation, the results in Table 1 are not directly comparable to the experimental data, because they involve the A5–A8 and MeHx ligands that are not involved in the SAMPL6 measurements. We have used two different approaches to solve this problem. For the submitted data, we employed previously published experimental data for A6, A8 and MeHx in OAH [40, 61] to calculate absolute affinities for all ligands. This is a bit risky, because $\Delta\Delta G_{bind}$ measured in different studies (at slightly different conditions) vary somewhat. For example, the experimental free energy of the Hx ligand (Figure S1) binding to OAH, involved in SAMPL4, vary between 21.1 and 23.5 kJ/mol in two publications by the same group [40, 93] and the results for A6, A8 and A10 vary by 1.5–2.8 kJ/mol [61, 93] (we employed the newer data in this article).

Our initial calculations along these lines showed that the calculated data were somewhat problematic: As can be seen in Fig. 2, the A6 and A8 ligands are connected by two perturbations, A6 → A7 → A8. However, the initial result for these perturbations was quite poor, −11.3 ± 0.3 kJ/mol, compared to the difference in the experimental $\Delta\Delta G_{bind}$ for A6 and A8, −4.9 or −6.2 kJ/mol in the two experimental studies [61, 93]. We therefore rerun these two perturbations with the whole ligand included in the perturbed group (instead of only the differing atoms). For A7 → A8, this did not change the results significantly, as can be seen in Table 1 (entry "entire ligand"). However, for A6 → A7, the result changed by 4 kJ/mol, bringing the A6 → A8 estimate closer to experiments, −7.8 ± 1.2 kJ/mol. Unfortunately, we did not have time to rerun all the other perturbations with the whole ligand in the soft-core group, but we used the latter results for the A6 → A7 perturbation and also corrected the corresponding results for OAM with the difference between the two A6 → A7 perturbations for OAH.

Absolute affinities calculated this way are shown in Table 2 (MM columns), together with the reference ligands (from which the experimental data was taken, because there are several possibilities) and the experimental data for SAMPL6 [39] (revealed after submission our results). As can be seen in Fig. 5a, the agreement is rather good with errors of 1.9–9.7 kJ/mol for the 16 predictions. However, the MAD is rather high, 5.6 ± 0.3 and 6.2 ± 0.3 kJ/mol for the two hosts. For most of the ligands, the predicted affinities are less negative than the experimental ones—the MSD is 5.0 ± 0.3 and 2.0 ± 0.4 kJ/mol for the two hosts. Yet, for G4 in both hosts and G2 in OAM, the opposite is true (note that these two

**Table 2** Calculated absolute binding free energies (kJ/mol) for the SAMPL6 ligands in the OAH and OAM hosts obtained with FEP at the MM and PM6-DH+/MM levels

| | OAH | | | | OAM | | | |
|---|---|---|---|---|---|---|---|---|
| | Ref | Exp | MM | QM/MM | Ref | Exp | MM | QM/MM |
| A5 | A6 | | −7.4 ± 0.1 | −7.7 ± 0.5 | A6 | | −9.3 ± 0.5 | −9.3 ± 0.7 |
| A6 | | −21.8 ± 0.1 | | | A6 | | −25.4 ± 0.4 | −27.0 ± 0.6 |
| A7 | A8 | | −21.0 ± 0.3 | −20.9 ± 0.5 | A6 | | −28.8 ± 1.0 | −31.3 ± 1.1 |
| A8 | | −28.0 ± 0.1 | | | A6 | | −37.7 ± 1.0 | −39.4 ± 1.1 |
| G0 | A8 | −23.8 ± 0.1 | −19.2 ± 0.5 | −18.8 ± 0.7 | A6 | −25.4 ± 0.1 | −19.8 ± 1.1 | −25.5 ± 1.2 |
| G1 | A6 | −19.5 ± 0.1 | −16.1 ± 0.1 | −10.6 ± 0.5 | A6 | −25.0 ± 0.2 | −23.1 ± 0.5 | −19.0 ± 1.1 |
| G2 | A8[a] | −35.1 ± 0.1 | −27.6 ± 0.8 | −27.0 ± 0.6 | A6 | −28.5 ± 0.1 | −36.5 ± 1.0 | −32.6 ± 1.6 |
| G3 | A6 | −21.7 ± 0.1 | −13.1 ± 0.1 | −14.9 ± 0.5 | A6 | −23.4 ± 0.2 | −14.5 ± 0.6 | −18.1 ± 0.7 |
| G4 | A8 | −29.7 ± 0.1 | −31.9 ± 2.1 | −29.2 ± 2.2 | A6 | −32.6 ± 0.1 | −41.3 ± 2.1 | −42.8 ± 2.2 |
| G5 | A6 | −19.2 ± 0.1 | −10.3 ± 0.3 | −9.8 ± 0.6 | A6 | −17.4 ± 0.1 | −7.7 ± 0.5 | −8.1 ± 0.7 |
| G6 | A6 | −20.8 ± 0.1 | −16.8 ± 0.3 | −16.1 ± 0.5 | A6 | −22.6 ± 0.1 | −18.6 ± 0.5 | −18.4 ± 0.7 |
| G7 | A6 | −26.0 ± 0.1 | −20.5 ± 0.4 | −16.2 ± 0.6 | A6 | −17.3 ± 0.1 | −14.6 ± 0.6 | −12.4 ± 0.8 |
| MeHx | | −31.8 ± 0.3 | | | A6 | | −32.2 ± 2.8 | |
| MAD | | | 5.6 ± 0.3 | 6.7 ± 0.3 | | | 6.2 ± 0.3 | 5.5 ± 0.4 |
| MADtr | | | 2.6 ± 0.3 | 2.4 ± 0.4 | | | 5.2 ± 0.4 | 5.0 ± 0.5 |
| MSD | | | 5.0 ± 0.3 | 6.7 ± 0.3 | | | 2.0 ± 0.4 | 1.9 ± 0.4 |
| RMSD | | | 6.0 ± 0.2 | 7.3 ± 0.2 | | | 6.8 ± 0.4 | 6.2 ± 0.5 |
| Max | | | 8.9 ± 0.3 | 9.8 ± 0.5 | | | 9.7 ± 0.9 | 10.2 ± 1.6 |
| slope | | | 1.1 ± 0.1 | 1.1 ± 0.1 | | | 2.0 ± 0.1 | 2.1 ± 0.1 |
| $R^2$ | | | 0.77 ± 0.05 | 0.81 ± 0.04 | | | 0.85 ± 0.02 | 0.93 ± 0.02 |
| $\tau$ | | | 0.79 ± 0.02 | 0.79 ± 0.06 | | | 0.71 ± 0.05 | 0.86 ± 0.07 |
| $\tau_{90}$ | | | 0.79 ± 0.02 | 0.84 ± 0.02 | | | 0.84 ± 0.02 | 1.00 ± 0.01 |

The absolute affinities were obtained by using experimental data for A6, A8 or MeHx bound to OAH [40, 61]. The reference employed is specified in the columns Ref. The experimental data are given in the Exp. columns [39]. The last nine rows show quality measures compared to the experimental results

[a]MeHx for MM-FEP

ligands were involved in transformations with a poor precision in Table 1). If the systematic error is removed, the MAD is improved significantly. For OAH, the MADtr is good, 2.6 ± 0.3 kJ/mol, whereas it is worse for OAM, 5.2 ± 0.4 kJ/mol. The reason for this is probably that the experimental data employed to calculate the absolute affinities were all for OAH, so the results for OAM involve more perturbations and therefore the possibility of accumulation of errors.

On the other hand, the correlation between the experimental and calculated results is better for OAM ($R^2 = 0.85 \pm 0.02$) than for OAH (0.77 ± 0.05), although the difference is not fully significant. The same applies also for $\tau_{90}$, which is 0.84 ± 0.02 and 0.79 ± 0.02 for OAM and OAH, respectively.

Alternatively, we instead considered only relative affinities. These were obtained by combining two or three perturbations so that they go between only the G0–G7 ligands. This can be done in a few different ways and one connected and consistent set of seven relative energies are shown in Table 3 (MM columns). It can be seen that the results are quite similar to those of the absolute affinities. The errors vary between 0.4 and 7.3 kJ/mol, except for the G0 → G2 difference in OAM, for which the error is as much as 13.6 kJ/mol (the calculated result overestimates the true difference, but with the correct sign). Consequently, the MAD is larger for OAM (5.1 ± 0.2 kJ/mol) than for OAH (3.1 ± 0.2 kJ/mol). $R^2$ is also better for OAH (0.87 ± 0.02, compared to 0.61 ± 0.04). On the other hand, $\tau_{r,90}$ is perfect for OAM (all statistically significant differences have the correct sign), whereas it is 0.71 for OAH (one difference has the incorrect sign). The single perturbation that involves only SAMPL6 ligands (G5 → G7) gives errors of the same size as the combined perturbations (3–7 kJ/mol), indicating that the results are not biased by poor performance of the added A5–A8 and MeHx ligands. Similar results are also obtained if relative energies involving the G0–G7 ligands, as well as the A6, A8 and MeHx ligands are considered (Table S2).

## FEP calculations at the QM/MM level

Next, we used the RPQS approach to calculate all the relative binding affinities at the QM/MM level. For this, we

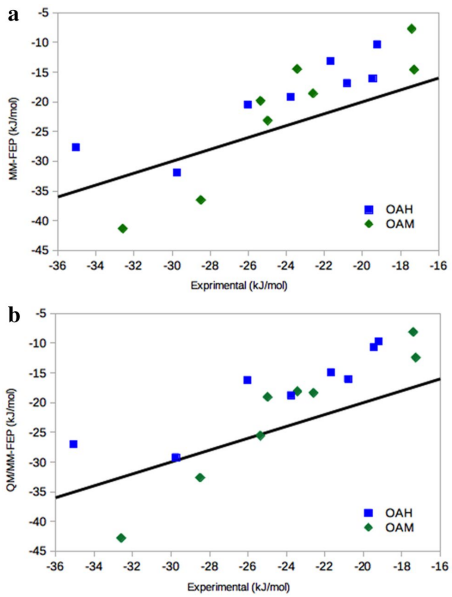



**Fig. 5** Comparison of the experimental and calculated absolute affinities obtained with **a** MM-FEP and **b** QM/MM-FEP methods. The line shows the perfect correlation

performed MM → QM/MM FEP calculations for all G0–G7 and A5–A8 ligands both when bound to the host and free in solution (cf. Fig. 3). The results are shown in Table 4. The individual MM → QM/MM free energies calculated when the ligand is bound to the host $\left(\Delta G_{\mathrm{L,bound}}^{\mathrm{MM\rightarrow QM/MM}}\right)$ or free in solution $\left(\Delta G_{\mathrm{L,free}}^{\mathrm{MM\rightarrow QM/MM}}\right)$, ranged from −507 to −691 kJ/mol, except for G4 (around −260 kJ/mol) and G7 (around −962 kJ/mol). However, for each ligand, the values in the host and in solution were of a similar size, and the resulting MM → QM/MM correction to $\Delta G_{\mathrm{bind}}$ ($\Delta\Delta G_{\mathrm{bind,L}}^{\mathrm{MM\rightarrow QM/MM}}$, shown in Table 4) ranged between −8.8 and +5.7 kJ/mol.

The standard errors were between 0.2 and 0.4 kJ/mol, except for G1 and G2 bound to OAM, for which they were 0.9 and 1.2 kJ/mol. For G0–G7 in OAM, we run duplicate calculations and for these two ligands, the results differed by 1.9 and 2.4 kJ/mol, whereas for the other ligands, they agreed within 0.5 kJ/mol. In fact, the large variation came from the $\Delta G_{\mathrm{L,bound}}^{\mathrm{MM\rightarrow QM/MM}}$ term, which varied by 2.6 and 3.2 kJ/mol for these two ligands, but less than 0.2 kJ/mol for the other ligands. For the $\Delta G_{\mathrm{L,free}}^{\mathrm{MM\rightarrow QM/MM}}$ term, for which we have two or three samples of each, the variation was 0.1–0.8 kJ/mol, except for G2 and G6 (1.2 and 2.0 kJ/mol).

We used five overlap measures (described in the Method section and shown in Table S7) to check that the calculated MM → QM corrections are reliable. Based on these, we added intermediate Λ values for some of the ligands, as is shown in the last two columns of Table 4.

Next, the $\Delta\Delta G_{\mathrm{bind,\ L}}^{\mathrm{MM\rightarrow QM/MM}}$ corrections in Table 4 were combined with the results of the FEP calculations at the MM

**Table 3** Calculated relative binding free energies (kJ/mol) for the SAMPL6 ligands in the OAH and OAM hosts obtained with FEP at the MM and PM6-DH+/MM levels

| | | OAH | | | OAM | | |
|---|---|---|---|---|---|---|---|
| | Via | Exp | MM | QM/MM | Exp | MM | QM/MM |
| G0 → G2 | A7 | −11.3 ± 0.2 | −18.6 ± 0.4 | −8.2 ± 0.6 | −3.1 ± 0.1 | −16.7 ± 0.5 | −7.1 ± 1.3 |
| G1 → G3 | A6 | −2.2 ± 0.1 | 2.9 ± 0.2 | −4.2 ± 0.5 | 1.5 ± 0.2 | 8.7 ± 0.4 | 0.9 ± 1.1 |
| G4 → G2 | A8, A7 | −5.3 ± 0.1 | −5.8 ± 2.1 | 2.2 ± 2.2 | 4.1 ± 0.1 | 4.8 ± 1.9 | 10.2 ± 2.2 |
| G5 → G6 | A5 | −1.6 ± 0.1 | −6.5 ± 0.4 | −6.3 ± 0.6 | −5.2 ± 0.2 | −10.9 ± 0.2 | −10.3 ± 0.5 |
| G5 → G7 | | −6.8 ± 0.1 | −10.2 ± 0.3 | −6.4 ± 0.5 | 0.1 ± 0.1 | −6.9 ± 0.3 | −4.3 ± 0.6 |
| G0 → G1 | A7, A6 | 4.3 ± 0.2 | 3.9 ± 0.9 | 10.9 ± 1.0 | 0.4 ± 0.2 | 1.0 ± 0.5 | 10.8 ± 1.1 |
| G5 → G3 | A5, A6 | −2.5 ± 0.1 | −2.8 ± 0.4 | −5.1 ± 0.6 | −6.0 ± 0.2 | −6.7 ± 0.5 | −10.0 ± 0.7 |
| MAD | | | 3.1 ± 0.2 | 3.9 ± 0.4 | | 5.1 ± 0.2 | 4.9 ± 0.4 |
| MSD | | | −1.7 ± 0.4 | 1.2 ± 0.4 | | −2.6 ± 0.3 | −0.2 ± 0.5 |
| RMSD | | | 4.1 ± 0.2 | 4.5 ± 0.6 | | 6.7 ± 0.2 | 5.6 ± 0.5 |
| max | | | 7.3 ± 0.4 | 7.5 ± 1.6 | | 13.6 ± 0.5 | 10.5 ± 1.1 |
| slope | | | 1.4 ± 0.1 | 0.9 ± 0.1 | | 2.0 ± 0.1 | 2.0 ± 0.1 |
| $R^2$ | | | 0.87 ± 0.02 | 0.56 ± 0.08 | | 0.61 ± 0.04 | 0.73 ± 0.04 |
| $\tau_r$ | | | 0.71 ± 0.01 | 0.71 ± 0.11 | | 0.71 ± 0.13 | 0.71 ± 0.16 |
| $\tau_{r,90}$ | | | 0.71 ± 0.01 | 1.00 ± 0.00 | | 1.00 ± 0.09 | 1.00 ± 0.06 |

The relative affinities involving only the SAMPL6 ligands were obtained by using 1–3 perturbations from Table 1 and the intermediate ligands are specified in the second column. The experimental results for the SAMPL6 ligands [39] are given in the Exp. columns

**Table 4** Calculated MM → QM/MM free energies (kJ/mol) for ligands G0–G7 and A5–A8 ($\Delta\Delta G^{MM\rightarrow QM/MM}_{bind,\,L} = \Delta\Delta G^{MM\rightarrow QM/MM}_{L,bound} - \Delta\Delta G^{MM\rightarrow QM/MM}_{L,free}$)

| | $\Delta\Delta G^{MM\rightarrow QM/MM}_{bind,\,L}$ | | #Λ | |
|---|---|---|---|---|
| | OAH | OAM | OAH | OAM |
| G0 | −4.7 ± 0.3 | −8.8 ± 0.2 | 4 | 4 |
| G1 | 2.3 ± 0.3 | 1.0 ± 0.9 | 5 | 5 |
| G2 | 5.7 ± 0.4 | 0.8 ± 1.2 | 5 | 5–6 |
| G3 | −4.8 ± 0.3 | −6.8 ± 0.2 | 4 | 4 |
| G4 | −2.4 ± 0.3 | −4.5 ± 0.3 | 5 | 5–6 |
| G5 | −2.6 ± 0.3 | −3.5 ± 0.2 | 4 | 4 |
| G6 | −2.3 ± 0.3 | −2.9 ± 0.2 | 4 | 4 |
| G7 | 1.2 ± 0.3 | −1.0 ± 0.2 | 5 | 4–5 |
| A5 | −3.3 ± 0.3 | −3.1 ± 0.3 | 4 | 4 |
| A6 | −3.1 ± 0.3 | −4.7 ± 0.3 | 4 | 4 |
| A7 | −5.0 ± 0.3 | −5.6 ± 0.3 | 4 | 4 |
| A8 | −5.1 ± 0.3 | −4.8 ± 0.3 | 4 | 4 |

The last two columns show the number of Λ values used in the calculations

level ($\Delta\Delta G^{MM}_{bind,L_0\rightarrow L_1}$ in Table 1) to get the final PM6-DH+/MM relative binding free energies $\left(\Delta\Delta G^{QM/MM}_{bind,L_0\rightarrow L_1}\right)$. These results are also included in Table 1. It can be seen that most MM → QM corrections are rather small, 0.1–3.7 kJ/mol (average 2.2 kJ/mol). However, for the G1 → A6 and G2 → A7 perturbations they are − 5.4 to − 10.7 kJ/mol.

These relative energies were then recalculated to absolute affinities in the same way as for the FEP results at the MM level. These results are shown in Table 2 (QM/MM columns) and in Fig. 4b. They differ from the MM results by 2 kJ/mol on average with a maximum difference of 5.7 kJ/mol. For OAH, the results are consistently less negative than the experimental results, by 5.0–9.5 kJ/mol for all ligands except G4 (0.6 kJ/mol; MSD = 6.7 ± 0.3 kJ/mol). Therefore, the MAD is rather high, 6.7 ± 0.3 kJ/mol, but the MADtr is excellent, 2.4 ± 0.4 kJ/mol. For OAM, the deviation is less systematic and more varying with a MSD of 1.9 ± 0.4 kJ/mol, MAD = 5.5 ± 0.4 kJ/mol, MADtr = 5.0 ± 0.5 kJ/mol and a maximum error of 10.2 kJ/mol for G4. However, the correlation is better for OAM ($R^2 = 0.93 \pm 0.02$, compared to 0.81 ± 0.04 for OAH) and $\tau_{90}$ is perfect for OAM, but 0.84 ± 0.02 for OAH. Compared to the MM-FEP results, the performance for OAH is similar (MAD, MSD and Max are worse, but MADtr, $R^2$ and $\tau_{90}$ are better). However, for OAM, the QM/MM-FEP results are clearly better for all quality measures, except for the maximum error.

We also made the corresponding analysis for the relative energies in Table 3 (QM/MM columns). The results, are similar to those obtained for the absolute energies: The MAD is lower for OAH than for OAM (3.9 ± 0.4 compared to 4.9 ± 0.4 kJ/mol). However, the correlation coefficient ($R^2$) is better for OAM, 0.73 ± 0.04, compared to 0.56 ± 0.08. $\tau_{r,90}$ is perfect for both hosts. Compared to the MM-FEP results, the two methods have a similar performance for OAH (MAD, $R^2$ and Max are better for MM-FEP, MSD and $\tau_{r,90}$ is better for QM/MM-FEP), but QM/MM-FEP is better (or equal) for OAM for all quality measures.

## Absolute binding affinities from minimised semi-empirical structures

Next, we tried to calculate absolute binding affinities for all the SAMPL6 host–guest complexes with QM-optimised structures, using a variation of an approach developed by Grimme [16, 70]. In the SAMPL5 study [31], we noticed that vacuum optimisations led to structures that had the guest carboxylate groups too much buried inside the host, forming hydrogen bonds with the host, rather than with water. This could only partly be remedied by running the optimisations with an implicit solvent method, such as COSMO [71, 72] or by including four explicit water molecules in the calculations. Therefore, in this study, we decided to base the calculations on snapshots from long MD simulations of the complex, employing QM/MM optimised structures with explicit water molecules in the MM system and including the host, guest and four water molecules (that form hydrogen bonds with the carboxylate group of the guest) in the QM system. We performed 100 or 200 ns MD simulations for each host–guest complex and extracted 10 or 20 snapshots from these.

To make the calculations rapid, allowing for calibration also on the previous SAMPL4 and SAMPL5 structures, we chose to employ the semiempirical dispersion- and hydrogen-bond-corrected PM6-DH+ method for the QM calculations. This reduced the computational effort to 3–5 h (single-core) for the QM/MM minimisations, compared to 2–4 weeks for the previous DFT optimisations. This could in principle be further sped up by using parallel calculations or by keeping the MM system fixed or restrained during the minimisation. After the minimisation, single-point PM6-DH+ energies were calculated for the isolated host–guest complex and these energies were combined with COSMO-RS solvation energies and thermostatistical corrections from a MM frequency calculation, according to Eq. 4. The PM6-DH+ energy and MM frequency calculations took only some tens of seconds to complete, leaving the COSMO-RS solvation energy calculations as the computational bottleneck, as these can take as much as 1 day to converge (besides the initial MD simulations, which took about 5 h per 10 ns on one GPU).

We started by testing the protocol on the nine cyclic carboxylates binding to OAH in the SAMPL4 competition [25]

**Table 5** $\Delta G_{bind}$ (kJ/mol) calculated for the SAMPL4 ligands with SQM approach and three different ways to combine the 10 energies from different snapshots, plain average, the minimum energy (Min) or the Boltzmann average (Boltz). The second column shows the experimental results [40]

| | Exp. | Average | Min | Boltz |
|---|---|---|---|---|
| Bz | −15.6 ± 0.2 | 10.9 ± 3.9 | −0.3 | 0.6 |
| MeBz | −24.5 ± 0.5 | 7.1 ± 1.9 | −2.4 | −0.9 |
| EtBz | −26.2 ± 0.1 | −0.1 ± 2.8 | −7.6 | −6.5 |
| pClBz | −28.1 ± 0.1 | −9.1 ± 2.4 | −21.2 | −19.9 |
| mClBz | −22.0 ± 0.2 | −6.4 ± 3.6 | −17.3 | −15.6 |
| Hx | −23.5 ± 0.3 | −3.2 ± 3.5 | −19.4 | −18.5 |
| MeHx | −31.8 ± 0.3 | −0.2 ± 1.6 | −7.4 | −5.6 |
| Pen | −15.6 ± 0.2 | 1.9 ± 2.2 | −9.2 | −8.0 |
| Hep | −27.7 ± 0.1 | 2.4 ± 3.3 | −7.7 | −6.6 |
| A6 | −21.8 ± 0.1 | 4.1 ± 3.3 | −10.7 | −9.8 |
| A8 | −28.0 ± 0.1 | −11.4 ± 2.8 | −22.3 | −20.7 |
| A10 | −31.5 ± 0.1 | −5.3 ± 2.3 | −14.9 | −13.2 |
| MAD | | 23.9 ± 0.8 | 13.0 | 14.3 |
| MADtr | | 5.1 ± 0.7 | 6.5 | 6.5 |
| MSD | | 23.9 ± 0.8 | 13.0 | 14.3 |
| RMSD | | 24.6 ± 0.8 | 14.8 | 16.0 |
| Max | | 31.6 ± 1.6 | 24.4 | 26.2 |
| Slope | | 0.7 ± 0.2 | 0.5 | 0.4 |
| $R^2$ | | 0.29 ± 0.10 | 0.12 | 0.10 |
| τ | | 0.36 ± 0.09 | 0.21 | 0.21 |
| $\tau_{90}$ | | 0.58 ± 0.04 | 0.23 | 0.23 |

**Table 6** $\Delta G_{bind}$ (kJ/mol) calculated for the SAMPL5 ligands with SQM approach and three different ways to combine the 10 energies from different snapshots, plain average, the minimum energy (Min) or the Boltzmann average (Boltz). The third column shows the experimental values

| Ligand | Host | Exp | Average | Min | Boltz |
|---|---|---|---|---|---|
| S5–G1 | OAH | −21.09 ± 0.04 | 5.5 ± 2.8 | −7.0 | −5.5 |
| S5–G2 | | −17.78 ± 0.04 | 10.1 ± 2.6 | −6.2 | −5.7 |
| S5–G4 | | −39.20 ± 0.01 | −4.7 ± 3.9 | −21.4 | −20.4 |
| S5–G6 | | −22.31 ± 0.02 | 1.9 ± 2.5 | −10.2 | −9.6 |
| S5–G1 | OAM | −21.92 ± 0.21 | 1.7 ± 2.0 | −9.2 | −7.9 |
| S5–G2 | | −21.09 ± 0.13 | 5.6 ± 6.0 | −13.4 | −12.5 |
| S5–G4 | | −9.96 ± 0.08 | −2.2 ± 2.5 | −11.9 | −10.1 |
| S5–G6 | | −18.91 ± 0.08 | 4.6 ± 1.8 | −5.4 | −4.7 |
| MAD | | | 24.4 ± 1.1 | 11.4 | 12.0 |
| MADtr | | | 4.6 ± 0.8 | 4.0 | 3.9 |
| MSD | | | 24.4 ± 1.1 | 10.9 | 12.0 |
| RMSD | | | 25.4 ± 1.2 | 12.3 | 13.1 |
| Max | | | 34.5 ± 3.4 | 17.8 | 18.9 |
| Slope | | | 0.2 ± 0.2 | 0.4 | 0.5 |
| $R^2$ | | | 0.17 ± 0.13 | 0.48 | 0.52 |
| τ | | | 0.33 ± 0.15 | 0.41 | 0.33 |
| $\tau_{90}$ | | | 0.33 ± 0.08 | 0.41 | 0.33 |

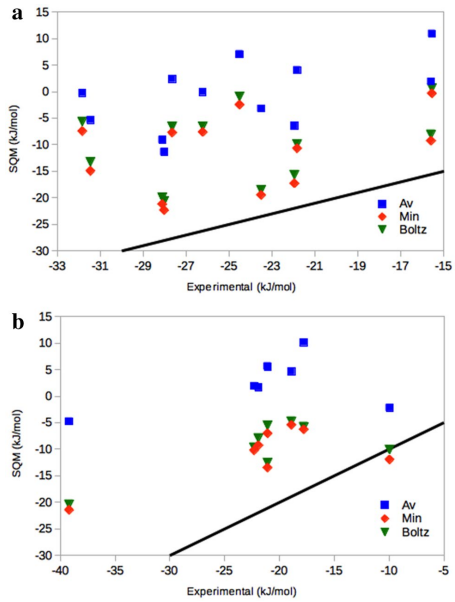



**Fig. 6** Comparison of the experimental and calculated absolute affinities obtained with the SQM method and three different ways to combine the 10 energies from different snapshots, plain average (Av), the minimum energy (Min) or the Boltzmann average (Boltz) for the **a** SAMPL4 and **b** SAMPL5 ligands. The line shows the perfect correlation

(Figure S1), the four carboxylic ligands binding to OAH and OAM in SAMPL5 [26] (S5–G1, S5–G2, S5–G4 and S5–G6, shown in Figure S2; we omitted the two positively charged ligands as all SAMPL6 ligands have a single negative charge), as well as A6, A8 and A10 binding to OAH [61].

As described above, the binding energies were obtained from 10 to 20 snapshots from the MD simulations. Therefore, we need to decide how these binding energies should be combined to single final estimate. To this end, we compared three different approaches: the averaged energy, the minimum energy and the Boltzmann-weighted averaged energy. The results are presented in Tables 5 and 6 and are shown in Fig. 6. It can be seen that all three methods give too weak binding (the calculated $\Delta G_{bind}$ is less negative than the experimental one). Of course, this underestimation is largest for the averaged energies (MSD = 24 kJ/mol for both sets) and smallest for the minimum energies (11–13 kJ/mol). Besides this systematic error, the minimum and Boltzmann-averaged energies give similar results (the

former is slightly better for SAMPL4, whereas the opposite is true for SAMPL5). Moreover, the averaged energies actually give the lowest MADtr and the best $R^2$ and $\tau_{90}$ results for SAMPL4. Theoretically, Boltzmann averaging is the preferred approach and it gave the best results in SAMPL5 [36], so we therefore used this approach for the submitted energies. The better performance of the plain averages for SAMPL4 may indicate that the sampling was incomplete. The averaged energies have the advantage of giving an uncertainty. It is quite high for all ligands, 2–6 kJ/mol, again showing that much more snapshots are needed to reach reliable results.

We have previously studied the same systems with minimised QM structures, but using more expensive DFT-D3 methods. For SAMPL4, the new results are of a similar quality (after removing the systematic error): The MADtr is 5.1–6.5 kJ/mol, which is similar or better than the previous DFT-D3 results, 4.6–8.6 kJ/mol. On the other hand, $R^2$ is lower, 0.10–0.29, compared to 0.60–0.78, and $\tau_{90}$ is also worse (0.23–0.58, compared to 0.71–0.77). However, for SAMPL5, all the new results are much better than the old DFT-D3 results: MADtr = 3.9–4.6 kJ/mol, compared to 11–21 kJ/mol, $R^2 = 0.17$–$0.52$, compared to 0–0.30 (and in many cases negative correlation), and $\tau_{90} = 0.33$–$0.41$, compared to $-0.33$ to 0.33. Still, it should be remembered that we did not include in this study the two ligands with trimethylammonium groups, which gave problems in the previous study. Yet, we believe that the present approach involves an important advantage: The geometries were optimised with PM6-DH+/MM in water, including four water molecules in the quantum system, which resulted in a lower repulsion of the negative carboxylate groups and a more realistic binding pose of the guests (with the guests always above the upper rim of the hosts and the carboxylate groups pointing upwards, forming hydrogen bonds with water molecules), shown in Figure S4 and described in Table S3. However, the vibrational frequencies still seem to be a problem, giving too positive binding affinities. In fact, the results without the frequency term gave a better MADtr for SAMPL4, but not for SAMPL5.

The results for the SAMPL6 ligands with the SQM approach using Boltzmann-weighted energies are collected in Table 7 and shown in Fig. 7a. It can be seen that they are quite similar to the results obtained for the SAMPL4 and SAMPL5 tests, with a systematic underestimation of the binding by MSD = 12–14 kJ/mol. For OAH, the MADtr is quite high 7.8 kJ/mol. The correlation is poor ($R^2 = 0.07$), as is $\tau_{90} = 0.07$. However, for the OAM host, all results are better: MADtr = 2.9 kJ/mol is excellent, and $R^2$ and $\tau_{90}$ are also improved, 0.42 and 0.43, respectively.

Interestingly, all relative results would have been improved if we had selected to submit the results of the pure average instead. For OAH, they give a MADtr of $5.1 \pm 1.4$ kJ/mol, a correlation of $0.16 \pm 0.21$ and a perfect $\tau_{90}$ of $1.00 \pm 0.07$. For OAM, MADtr = $2.3 \pm 0.5$ kJ/mol, $R^2 = 0.88 \pm 0.08$ and $\tau_{90} = 1.00 \pm 0.05$.

In Table S4, the various energy components are shown for the SQM calculations for the snapshot with the best

**Table 7** $\Delta G_{bind}$ (kJ/mol) calculated for the SAMPL6 ligands with SQM and DFT approaches and three different ways to combine the 10 energies from different snapshots, plain average, the minimum energy (Min) or the Boltzmann average (Boltz)

| | OAH | | | | OAM | | | |
|---|---|---|---|---|---|---|---|---|
| | SQM | | | DFT | SQM | | | DFT |
| | Av | Min | Boltz | | Av | Min | Boltz | |
| G0 | −1.2 ± 2.7 | −18.9 | −18.5 | −26.2 | 4.9 ± 1.5 | −8.4 | −6.7 | 5.8 |
| G1 | 1.5 ± 3.7 | −15.9 | −14.8 | −22.0 | 2.5 ± 2.2 | −13.7 | −11.8 | −6.8 |
| G2 | 3.7 ± 8.6 | −7.6 | −6.5 | −38.0 | 0.5 ± 2.0 | −13.8 | −11.5 | −3.8 |
| G3 | 6.9 ± 3.0 | −7.7 | −6.6 | −26.6 | 1.9 ± 1.3 | −10.8 | −9.1 | 5.3 |
| G4 | −12.2 ± 4.7 | −35.3 | −35.1 | −55.2 | −10.6 ± 1.4 | −21.7 | −19.3 | −29.8 |
| G5 | 5.2 ± 3.2 | −8.9 | −8.0 | −26.5 | 9.8 ± 3.1 | −13.5 | −12.4 | 5.5 |
| G6 | 7.9 ± 3.7 | −6.1 | −4.6 | −4.8 | 4.8 ± 1.3 | −5.2 | −3.4 | 6.5 |
| G7 | 3.8 ± 3.3 | −9.1 | −7.7 | −15.5 | 15.5 ± 2.4 | −4.0 | −2.6 | 16.3 |
| MAD | 26.4 ± 1.5 | 12.2 | 13.1 | 9.0 | 27.7 ± 0.7 | 12.6 | 14.4 | 23.9 |
| MADtr | 5.1 ± 1.4 | 7.5 | 7.8 | 7.8 | 2.3 ± 0.5 | 3.0 | 2.9 | 7.0 |
| MSD | 26.4 ± 1.5 | 10.8 | 11.7 | −2.4 | 27.7 ± 0.7 | 12.6 | 14.4 | 23.9 |
| RMSD | 27.1 ± 1.8 | 14.3 | 15.2 | 11.8 | 27.8 ± 0.7 | 13.2 | 15.0 | 25.6 |
| Max | 38.8 ± 6.8 | 27.5 | 28.6 | 25.5 | 32.8 ± 1.9 | 17.4 | 19.2 | 33.6 |
| Slope | 0.5 ± 0.4 | 0.4 | 0.5 | 1.7 | 1.4 ± 0.2 | 0.8 | 0.7 | 2.3 |
| $R^2$ | 0.16 ± 0.21 | 0.07 | 0.07 | 0.38 | 0.88 ± 0.08 | 0.48 | 0.42 | 0.73 |
| $\tau$ | 0.36 ± 0.22 | 0.14 | 0.07 | 0.29 | 0.71 ± 0.11 | 0.64 | 0.43 | 0.64 |
| $\tau_{90}$ | 1.00 ± 0.07 | 0.14 | 0.07 | 0.29 | 1.00 ± 0.05 | 0.63 | 0.41 | 0.63 |

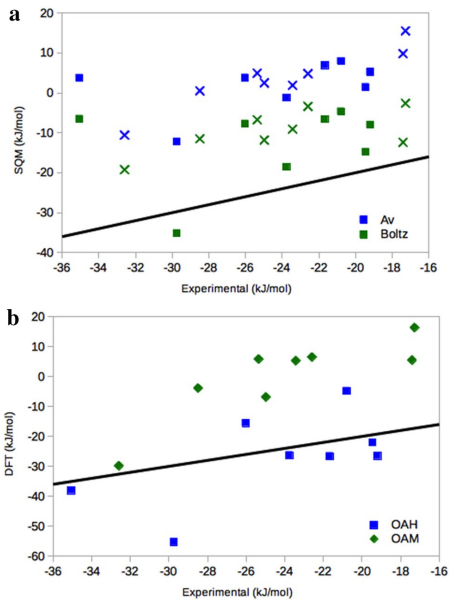



**Fig. 7** Comparison of the experimental and calculated absolute affinities obtained with the **a** SQM and **b** DFT methods, the former with two different ways to combine the 10–20 energies from different snapshots, plain average (Av) or the Boltzmann average (Boltz) for the SAMPL6 ligands. The line shows the perfect correlation. In **a** OAH energies are shown with squares and OAM energies with crosses

binding energy. It can be seen that the thermostatistical term shows a small variation, 46–58 kJ/mol for OAH and 55–78 kJ/mol for OAM. It shows a weak anti-correlation with the experimental binding energies, $R^2 = 0.55$, for OAH and 0.28 for OAM. The QM term is always large and positive, somewhat lower for OAH than for OAM, 798–942 and 758–865 kJ/mol. It shows a poor correlation with the experimental data for OAH ($R^2 = 0.16$), but appreciably better for OAM (0.66). It is more than compensated by the solvation energy, which is again is larger in magnitude for OAH than for OAM, − 875 to − 995 and − 843 to − 932 for OAM. It shows a similar (anti-)correlation to the experimental data as the QM term, 0.12 for OAH, but 0.68 for OAM. The sum of the latter two terms shows an improved correlation to the experimental data for OAH (0.28) but a worse correlation for OAM (0.24). Adding the thermostatistical correction deteriorates the correlation for OAH, but improves it for OAM.

Figure S5 shows the variation of the individual SQM $\Delta G_{\text{bind}}$ results for the eight ligands in the two hosts. It can be seen that it is 20–35 kJ/mol for most ligands. However, G4 and G6 in OAH, as well as G5 and G7 in OAM show a larger variation 40–64 kJ/mol. There is little correlation between the variation and the strength of the binding or the type of host.

## Absolute binding affinities from minimised DFT structures

Finally, we tried to improve the absolute binding affinities by using DFT calculations both in the geometry optimisations and in the energy estimates. Thus, we selected the minimum-energy snapshot according to SQM calculations and performed DFT/MM optimisation with the surrounding water included as a fixed MM system. We then calculated energies of the resulting structures in a way similar to that used for SQM (Eq. 4), still using thermostatistical corrections from MM vibrational frequencies and COSMO-RS solvation energies, but with TPSS-d3/def2-QZVP' energies instead of the PM6-DH+ energies.

The DFT results are also included in Table 7 and they are shown Fig. 7b. They are less positive than the SQM results. In fact, the MSD for OAH is actually slightly negative, − 2.4 kJ/mol, whereas it is 24 kJ/mol for OAM. The solvation energies are of a similar magnitude in the two sets of calculations, whereas the energies are somewhat more positive for the PM6-DH+ calculations on OAH (by 13 kJ/mol on average), but less positive for OAM (by 9 kJ/mol on average). The thermostatistical correction is of a similar magnitude. The MADtr is 7.0–7.8 kJ/mol for the two hosts, $R^2 = 0.38$–$0.73$ and $\tau_{90} = 0.29$–$0.63$ (better for OAM than for OAH), i.e. mostly within the range of the SQM methods.

As mentioned in the "Methods" section, the ligand dissociated in most of the original MM minimisations (to calculate the frequencies for the thermostatistical corrections). This was not discovered until after the submissions. For the DFT calculations, the ligand did not dissociate, but by mistake, the calculations were performed with zeroed partial charges on all atoms of the host and the ligand. The submitted SQM and DFT results are provided in Table S4. The original SQM submission gave a much lower systematic error (MSD), but a similar performance in terms of the relative quality measures (MADtr, $R^2$ and $\tau$).

Finally, we compare structures of the complexes obtained in the MD simulations and after minimisation with either SQM or DFT, employing a number of geometric measures, which are described in the "Methods" section. The results are collected in Table S6 and it can be seen that the guests always bind inside the host, with the carboxylate group 0.7–4.6 Å above the upper rim, forming hydrogen bonds with the water molecules and not with the host. The average

distances between the carboxylate atoms of the guests and the upper rim of the hosts, $r_{O1}$ and $r_{O2}$ are always slightly smaller for the SQM than the MD structures, but only by 0.1–0.7 Å, whereas the results for the DFT structures are more varying (probably because they are based on a single structure). Likewise, the ligand is always less deeply buried in the host in the SQM calculations than in the MD simulations, by up to 0.6 Å and the benzoate groups are less tilted ($r_{Cav}$ is 0.1–0.3 Å smaller). In most cases, the tilt angle ($\alpha_T$) of the ligand was also somewhat smaller (4° on average) with SQM than in MD. All these differences probably reflect differences in the potential-energy method and the fact that the SQM structures are minimised and not from a MD simulation rather than a systematic error of the SQM structures, observed in our previous approaches [15, 31]. Figure S6 compares the DFT and SQM structures, showing that they are very similar.

### Comparison with other submissions

There were 43 submissions for the SAMPL6 octa-acid challenge from eight research groups. Of course, the results depend on whether the two hosts are considered separately or together and how the various measures are combined and weighted. Here, we discuss the results based on six quality measures (MAD, RMSD, MSD, $R^2$, τ and slope), provided by the organisers for the combined OAH and OAM results and give a final ranking based on the sum of the ranks for these six measures. Irrespectively of how the ranking is done, three pairs of calculations always come out among the best. Two submissions from the Merz group, gave the lowest MAD and RMSD (3.2 and 4.0 kJ/mol, respectively). They also gave quite good results for the other measures, e.g. $R^2 = 0.60–0.85$ and $\tau = 0.37–0.74$. The best calculation employed potential-of-mean-force umbrella-sampling simulations (i.e. dragging the ligand out of the host) and scaled the results based on the corresponding results obtained for the SAMPL5 ligands (without the scaling, the results were appreciably worse, ranking around position 14). The other calculation used the movable-type approach, fitted to the former result. In fact, this group submitted 27 data sets, which ranked from the best to the third worst.

A FEP study of absolute free energies by the Michel group, employing GAFF with AM1-BCC charges, TIP3P water and with or without counter ions also gave good results, but only after a linear fit employing the results from the SAMPL5 competition. They obtained MAD and RMSD of 5.3–5.6 and 6.7–7.4 kJ/mol, respectively, $R^2 = 0.78–0.79$ and $\tau = 0.70$, making them the fourth and sixth best methods. Without the linear fit, the results ranked 19–35.

Our MM-FEP and QM/MM-FEP gave similar results with MAD = 5.6–6.1 kJ/mol and RMSD = 6.3–6.8 kJ/mol, $R^2 = 0.66–0.71$ and $\tau = 0.62–0.77$. In fact, τ for MM-FEP was best among all submissions. Based on the sum of the ranks for all six quality measures, MM-FEP gave the third best results and QM-FEP the fifth among all 43 submissions. In particular, they were clearly the best submissions using only the raw SAMPL6 data, without any fit to the SAMPL5 data.

The DFT and SQM results ranked slightly below the middle, positions 28 and 29, with MAD = 9–11 kJ/mol and RMSD = 11–13 kJ/mol, $R^2 = 0.1–0.5$ and $\tau = 0.3–0.4$. However, the performance may have improved if relative quality measures were considered, like MADtr, and they would also have improved if we had submitted the average results, instead of the Boltzmann-weighted results. Still, it is quite satisfying that for the first time, a QM approach, QM/MM-FEP, come within the best six submissions. Moreover, both SQM and DFT gave decent results, better than many of the MM FEP results, e.g. a submission employing FEP with the polarisable AMOEBA force field [94] and the absolute FEP calculations without the linear fit. In particular, they are appreciably better than the other purely QM submission, employing B3PW91 calculations with complete basis sets and a SMD continuum solvent [95], which performed poorly.

## Conclusions

We have studied the binding of eight ligands to two variants of the octa-acid deep-cavity host in the SAMPL6 blind-test competition [38, 39]. We have employed four different approaches (cf. Fig. 4), three of which are based on QM methods. First, we performed standard relative FEP calculations at the MM level with free energies calculated with MBAR and employing the GAFF+TIP3P force fields and RESP charges. Second, we used the reference-potential approach with explicit QM/MM sampling to obtain relative FEP free energies at the PM6-DH+/MM level of theory for the ligand. Third, we employed the same SQM method to obtain QM/MM optimised structures with the ligand, the host and four water molecules in the QM system, for which free energies were calculated by combining the PM6-DH+ interaction energies with COSMO-RS solvation free energies and thermostatistical corrections calculated at the MM level. We employed 10–20 structures taken from a MD simulation of the host–guest complexes. Finally, we reoptimised the best structures from the previous approach with the TPSS-D3/MM method and calculated QM energies with a large basis set, which were then combined with COSMO-RS and thermostatistical corrections.

The MM- and QM/MM-FEP methods gave excellent results for OAH, with MADtr of 2.4–2.6 kJ/mol and $R^2$ of 0.77–0.81. For OAM, the MADtr was somewhat larger, 5.0–5.2 kJ/mol, but the $R^2$ was better, 0.85–0.93. For the former, the two approaches gave similar results, whereas for OAM, QM/MM-FEP was clearly better. These results were

among the five best submissions to SAMPL6 and they were actually the best ones using no fit to data from SAMPL5.

The results obtained with QM/MM optimised structures were somewhat worse, especially for OAH; MADtr = 2.3–5.1 kJ/mol and $R^2 = 0.16–0.88$. Unfortunately, we selected to submit SQM results based on Boltzmann-averaged, rather than plain averaged energies, which gave somewhat worse results, MADtr = 2.9–7.8 kJ/mol and $R^2 = 0.07–0.42$. However, these methods gave similar results as our previous calculations with DFT-optimised structures in SAMPL4 and much better results for SAMPL5. Compared to the other submissions, these results were mediocre, but still comparable to many approaches employing MM-FEP methods. In particular, they gave better performance than other submissions employing QM-optimised structures.

The present results are quite satisfying because for the first time we are able to improve MM-FEP results for the octa-acid host with QM/MM methods and these results are among the best five submissions. These results were obtained with the simple and cheap SQM PM6-DH+ method, demonstrating that appropriate sampling and properly converging the MM → QM/MM FEP is of greater importance than using more rigorous QM methods. However, with a functional QM/MM-FEP approach, our next challenge will be to extend it to more accurate QM methods and larger QM systems.

For the QM-minimised structures, we have shown that the results are improved by employing QM/MM-optimised structures, rather than QM structures optimised in vacuum or in a continuum solvent. This also made the calculations significantly faster. However, there are still several problems to solve with this approach. In particular, there seems to be a problem with absolute free energies, probably related to the entropy term, which vary by 10–40 kJ/mol, depending on what method is used for the geometries and the frequencies. In particular, we observe that the simple PM6-DH+ method gives lower MADtr than the inherently more accurate TPSS-D3 approach. It would be more satisfactory if the same method is used for the geometries and the frequencies. Moreover, much more sampling seems to be needed before the results are stable and reliable. With 10–20 snapshots, plain averages gave better results (but with large uncertainties, 1–5 kJ/mol). Finally, improved methods to estimate the strain energies of the host and the guest in the complexes are needed.

Still it is very satisfying that QM-based methods finally start to have some impact on calculated binding affinities for host–guest systems.

**Acknowledgements** This investigation has been supported by grants from the Swedish research council (project 2014-5540), China Scholarship Council, the Royal Physiographic Society in Lund and from Knut and Alice Wallenberg Foundation (KAW 2013.0022). The computations were performed on computer resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lunarc at Lund University and HPC2N at Umeå University.



## References

1. Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chemie-Int Ed 41:2644–2676
2. Jorgensen WL (2009) Efficient drug lead discovery and optimization. Acc Chem Res 42:724–733
3. Kontoyianni M, Madhav P, Seibel ES (2008) Theoretical and practical considerations in virtual screening: a beaten field? Curr Med Chem 15:107–116. https://doi.org/10.2174/092986708783330566
4. Åqvist J, Luzhkov VB, Brandsdal BO (2002) Ligand binding affinities from MD simulations. Acc Chem Res 35:358–365
5. Kollman PA, Massova I, Reyes CM et al (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33:889–897
6. Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10:449–461. https://doi.org/10.1517/17460441.2015.1032936
7. Wereszczynski J, McCammon JA (2012) Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. Q Rev Biophys 45:1–25. https://doi.org/10.1017/S0033583511000096
8. Hansen N, Van Gunsteren WF (2014) Practical aspects of free-energy calculations: a review. J Chem Theory Comput 10:2632–2647. https://doi.org/10.1021/ct500161f
9. Zwanzig RW (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. J Chem Phys 22:1420–1426. https://doi.org/10.1063/1.1740409
10. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. J Chem Phys 3:300
11. Bennett CH (1976) Efficient estimation of free energy differences from monte carlo data. J Comput Phys 22:245–268
12. Jensen JH (2015) Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods. Phys Chem Chem Phys 17:12441–12451. https://doi.org/10.1039/c5cp00628g
13. Moghaddam S, Inoue Y, Gilson MK (2009) Host–guest complexes with protein–ligand-like affinities: computational analysis and design host—guest complexes with protein—ligand-like affinities. J Am Chem Soc 131(11):4012–4021. https://doi.org/10.1021/ja808175m
14. Ryde U, Söderhjelm P (2016) Ligand-binding affinity estimates supported by quantum-mechanical methods. Chem Rev 116:5520–5566. https://doi.org/10.1021/acs.chemrev.5b00630
15. Mikulskis P, Cioloboc D, Andrejić M et al (2014) Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host-guest binding energies. J Comput Aided Mol Des 28:375–400. https://doi.org/10.1007/s10822-014-9739-x
16. Grimme S (2012) Supramolecular binding thermodynamics by dispersion-corrected density functional theory. Chem-A Eur J 18:9955–9964. https://doi.org/10.1002/chem.201200497

17. Ryde U (2016) QM/MM calculations on proteins. Methods Enzymol 577:119–158
18. Senn HM, Thiel W (2009) QM/MM methods for biomolecular systems. Angew Chemie-Int Ed 48:1198–1229. https://doi.org/10.1002/anie.200802019
19. Reddy MR, Erion MD (2007) Relative binding affinities of fructose-1,6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method. J Am Chem Soc 129:9296–9297. https://doi.org/10.1021/ja072905j
20. Rathore RS, Reddy RN, Kondapi AK et al (2012) Use of quantum mechanics/molecular mechanics-based FEP method for calculating relative binding affinities of FBPase inhibitors for type-2 diabetes. Theor Chem Acc 131:1096. https://doi.org/10.1007/s00214-012-1096-z
21. Świderek K, Martí S, Moliner V (2012) Theoretical studies of HIV-1 reverse transcriptase inhibition. Phys Chem Chem Phys 14:12614–12624. https://doi.org/10.1039/c2cp40953d
22. Olsson MA, Ryde U (2017) Comparison of QM/MM methods to obtain ligand-binding free energies. J Chem Theory Comput 13:2245–2253. https://doi.org/10.1021/acs.jctc.6b01217
23. Steinmann C, Olsson MA, Ryde U (2018) Relative ligand-binding free energies calculated from multiple short QM/MM MD simulations. J Chem Theory Comput 14:3228–3237. https://doi.org/10.1021/acs.jctc.8b00081
24. Muddana HS, Varnado CD, Bielawski CW et al (2012) Blind prediction of host-guest binding affinities: a new SAMPL3 challenge. J Comput Aided Mol Des 26:475–487. https://doi.org/10.1007/s10822-012-9554-1
25. Muddana HS, Fenley AT, Mobley DL, Gilson MK (2014) The SAMPL4 host-guest blind prediction challenge: an overview. J Comput Aided Mol Des 28:305–317. https://doi.org/10.1007/s10822-014-9735-1
26. Yin J, Henriksen NM, Slochower DR et al (2017) Overview of the SAMPL5 host–guest challenge: are we doing better ? J Comput Aided Mol Des 31:1–19. https://doi.org/10.1007/s10822-016-9974-4
27. Xi H, Gibb LD C (1998) Deep-cavity cavitands: synthesis and solid state structure of host molecules possessing large bowl-shaped cavities. Chem Commun. https://doi.org/10.1039/A803571G
28. Liu S, Gibb BC (2008) High-definition self-assemblies driven by the hydrophobic effect: synthesis and properties of a supramolecular nanocapsule. Chem Commun 7345:3709–3716. https://doi.org/10.1039/b805446k
29. Olsson MA, Söderhjelm P, Ryde U (2016) Converging ligand-binding free energies obtained with free-energy perturbations at the quantum mechanical level. J Comput Chem 37:1589–1600. https://doi.org/10.1002/jcc.24375
30. Andrejić M, Ryde U, Mata RA, Söderhjelm P (2014) Coupled-cluster interaction energies for 200-atom host–guest systems. ChemPhysChem 15:3270–3281. https://doi.org/10.1002/cphc.201402379
31. Caldararu O, Olsson MA, Riplinger C et al (2017) Binding free energies in the SAMPL5 octa-acid host–guest challenge calculated with DFT-D3 and CCSD(T). J Comput Aided Mol Des 31:87–106. https://doi.org/10.1007/s10822-016-9957-5
32. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690
33. Hudson PS, Woodcock HL, Boresch S (2015) Use of nonequilibrium work methods to compute free energy differences between molecular mechanical and quantum mechanical representations of molecular systems. J Phys Chem Lett 6:4850–4856. https://doi.org/10.1021/acs.jpclett.5b02164
34. Wang M, Mei Y, Ryde U (2018) Predicting the relative binding affinity: using nonequilibrium simulation for the MM->QM correction. J Chem Theory Comput (**submitted**)
35. Sure R, Antony J, Grimme S (2014) Blind prediction of binding affinities for charged supramolecular host-guest systems: achievements and shortcomings of DFT-D3. J Phys Chem B 118:3431–3440. https://doi.org/10.1021/jp411616b
36. Gan H, Benjamin CJ, Gibb BC (2011) Nonmonotonic assembly of a deep-cavity cavitand. J Am Chem Soc 133:4770–4773. https://doi.org/10.1021/ja200633d
37. Riplinger C, Neese F (2013) An efficient and near linear scaling pair natural orbital based local coupled cluster method. J Chem Phys 138:1–18. https://doi.org/10.1063/1.4773581
38. Rizzi A, Murkli S, McNeill JN et al (2018) Overview of the SAMPL6 host-guest 2 binding affinity prediction challenge. J Comput Aided Mol Des (**in press**); https://www.biorxiv.org/content/early/20
39. Gibb BC (2018) Experimental data for SAMPL6. J Comput Aided Mol Des (**in press**)
40. Gibb CLD, Gibb BC (2014) Binding of cyclic carboxylates to octa-acid deep-cavity cavitand. J Comput Aided Mol Des 28:319–325. https://doi.org/10.1007/s10822-013-9690-2
41. Sullivan MR, Sokkalingam P, Nguyen G et al (2017) Binding of carboxylate and trimethylammonium salts to octa-acid and TEMOA deep-cavity cavitands. J Comput Aided Mol Des 31:21–28
42. Hanwell MD, Curtis DE, Lonie DC et al (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4:17. https://doi.org/10.1186/1758-2946-4-17
43. Addicoat MA, Vankova N, Akter IF, Heine T (2014) Extension of the universal force field to metal-organic frameworks. J Chem Theory Comput 10:880–891. https://doi.org/10.1021/ct400952t
44. Wang JM, Wolf RM, Caldwell JW et al (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174. https://doi.org/10.1002/jcc.20035
45. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935. https://doi.org/10.1063/1.445869
46. Bayly CI, Cieplak P, Cornell WD, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J Phys Chem 97:10269–10280. https://doi.org/10.1021/j100142a004
47. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) A new general purpose quantum mechanical molecular model. J Am Chem Soc 107:3902–3909
48. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11:431–439. https://doi.org/10.1002/jcc.540110404
49. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09 Revision A. 02
50. Seminario JM (1996) Calculation of intramolecular force fields from second-derivative tensors. Int J Quantum Chem 30:1271–1277
51. Tao J, Perdew JP, Staroverov VN, Scuseria GE (2003) Climbing the density functional ladder: non-empirical meta-generalized gradient approximation designed for molecules and solids. Phys Rev Lett 91:146401. https://doi.org/10.1103/PhysRevLett.91.146401
52. Weigend F, Ahlrichs R (2005) Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. Phys Chem Chem Phys 7:3297–3305. https://doi.org/10.1039/b508541a
53. Furche F, Ahlrichs R, Hättig C et al (2014) Turbomole. Wiley Interdiscip Rev Comput Mol Sci 4:91–100. https://doi.org/10.1002/wcms.1162

54. Nilsson K, Lecerof D, Sigfridsson E, Ryde U (2003) An automatic method to generate force-field parameters for hetero-compounds. Acta Crystallogr D 59:274–289. https://doi.org/10.1107/S0907444902021431
55. Case DA, Cerutti TE, Cheatham TE III et al (2016) AMBER 2016. University of California, San Francisco
56. Genheden S, Ryde U (2011) A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations. J Comput Chem 32:187–195. https://doi.org/10.1002/jcc.21564
57. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341. https://doi.org/10.1016/0021-9991(77)90098-5
58. Wu X, Brooks BR (2003) Self-guided Langevin dynamics simulation method. Chem Phys Lett 381:512–518. https://doi.org/10.1016/j.cplett.2003.10.013
59. Berendsen HJC, Postma JPM, van Gunsteren WF et al (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
60. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N-log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
61. Wang K, Sokkalingam P, Gibb BC (2016) ITC and NMR analysis of the encapsulation of fatty acids within a water-soluble cavitand and its dimeric capsule. Supramol Chem 28:84–90. https://doi.org/10.1080/10610278.2015.1082563
62. Steinbrecher T, Mobley DL, Case DA (2007) Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. J Chem Phys 127:1–13. https://doi.org/10.1063/1.2799191
63. Steinbrecher T, Joung I, Case DA (2011) Soft-core potentials in thermodynamic integration: comparing one-and two-step transformations. J Comput Chem 32:3253–3263. https://doi.org/10.1002/jcc.21909
64. Gilson MK, Given JA, Bush BL, Mccammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J 72:1047–1069
65. Genheden S, Nilsson I, Ryde U (2010) Binding affinities of factor Xa inhibitors estimated by thermodynamic integration and MM/GBSA. J Chem Inf Model 51:947–958. https://doi.org/10.1021/ci100458f
66. Shirts MR, Chodera JD (2008) Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys 129(10 pages):124105. https://doi.org/10.1063/1.2978177
67. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173–1213. https://doi.org/10.1007/s00894-007-0233-4
68. Korth M (2010) Third-generation hydrogen-bonding corrections for semiempirical QM methods and force fields. J Chem Theory Comput 6:3808–3816. https://doi.org/10.1021/ct100408b
69. Jurečka P, Černý J, Hobza P, Salahub DR (2007) Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. J Comput Chem 28:555–569. https://doi.org/10.1002/jcc.20570
70. Antony J, Sure R, Grimme S (2015) Using dispersion-corrected density functional theory to understand supramolecular binding thermodynamics. Chem Commun 51:1764–1774. https://doi.org/10.1039/C4CC06722C
71. Klamt A, Schüürmann G (1993) Cosmo—a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc-Perkin Trans 2:799–805
72. Schäfer A, Klamt A, Sattel D et al (2000) COSMO Implementation in TURBOMOLE: extension of an efficient quantum chemical code towards liquid systems. Phys Chem Chem Phys 2:2187–2193. https://doi.org/10.1039/b000184h
73. Klamt A (1995) Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. J Phys Chem 99:2224–2235. https://doi.org/10.1021/j100007a062
74. Eckert F, Klamt A (2002) Fast solvent screening via quantum chemistry: COSMO-RS approach. AIChE J 48:369–385. https://doi.org/10.1002/aic.690480220
75. Eckert F, Klamt A (2010) COSMOtherm, C3.0 Release 13.01, COSMOlogic GmbH & Co KG. http://www.cosmologic.de
76. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic-behavior. Phys Rev A 38:3098–3100. https://doi.org/10.1103/PhysRevA.38.3098
77. Perdew JP (1986) Density-functional approximation for the correlation energy of the inhomogeneous electron gas. Phys Rev B 33:8822–8824
78. Schäfer A, Horn H, Ahlrichs R (1992) Fully optimized contracted Gaussian basis sets for atoms Li to Kr. J Chem Phys 97:2571–2577. https://doi.org/10.1063/1.463096
79. Jensen F (2017) Introduction to computational chemistry, 3rd edn. Wiley, Chichester
80. Ryde U (1996) The coordination of the catalytic zinc in alcohol dehydrogenase studied by combined quantum-chemical and molecular mechanics calculations. J Comput Aided Mol Des 10:153–164. https://doi.org/10.1007/BF00402823
81. Ryde U, Olsson MHM (2001) Structure, strain, and reorganization energy of blue copper models in the protein. Int J Quantum Chem 81:335–347. https://doi.org/10.1002/1097-461X%282001%2981:5%3C335::AIDQUA1003%3E3.0.CO;2-Q
82. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132(19 pages):154104. https://doi.org/10.1063/1.3382344
83. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32:1456–1465. https://doi.org/10.1002/jcc.21759
84. Eichkorn K, Treutler O, Öhm H et al (1995) Auxiliary basis-sets to approximate coulomb potentials. Chem Phys Lett 240:283–289. https://doi.org/10.1016/0009-2614(95)00621-a
85. Eichkorn K, Weigend F, Treutler O, Ahlrichs R (1997) Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. Theor Chem Acc 97:119–124. https://doi.org/10.1007/s002140050244
86. Sierka M, Hogekamp A, Ahlrichs R (2003) Fast evaluation of the Coulomb potential for electron densities using multipole accelerated resolution of identity approximation. J Chem Phys 118:9136–9148. https://doi.org/10.1063/1.1567253
87. Kaus JW, Pierce LT, Walker RC, Mccammon JA (2013) Improving the efficiency of free energy calculations in the amber molecular dynamics package. J Chem Theory Comput 9:4131–4139
88. Genheden S, Ryde U (2010) How to obtain statistically converged MM/GBSA results. J Comput Chem 31:837–846. https://doi.org/10.1002/jcc.21366
89. Bhattacharyya A (1943) On a measure of divergence between two statistical populations defined by their probability distributions. Bull Calcutta Math Soc 35:99–109
90. Wu D, Kofke DA (2005) Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. J Chem Phys 123:1–10. https://doi.org/10.1063/1.1992483
91. Rod TH, Ryde U (2005) Quantum mechanical free energy barrier for an enzymatic reaction. Phys Rev Lett 94(4 pages):138302. https://doi.org/10.1103/PhysRevLett.94.138302
92. Mikulskis P, Genheden S, Ryde U (2014) A large-scale test of free-energy simulation estimates of protein-Ligand binding affinities.

J Chem Inf Model 54:2794–2806. https://doi.org/10.1021/ci5004027

93. Sun H, Gibb CLD, Gibb BC (2008) Calorimetric analysis of the 1:1 complexes formed between a water-soluble deep-cavity cavitand, and cyclic and acyclic carboxylic acids. Supramol Chem 20:141–147. https://doi.org/10.1080/10610270701744302
94. Ponder JW, Wu C, Pande VS et al (2010) Current status of the AMOEBA polarizable force field. J Phys Chem B 114:2549–2564. https://doi.org/10.1021/jp910674d
95. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113:6378–6396. https://doi.org/10.1021/jp810292n

# Paper IV

Organic & Biomolecular Chemistry

PAPER

View Article Online
View Journal | View Issue



Cite this: *Org. Biomol. Chem.*, 2019, **17**, 1081

# Substituted polyfluoroaryl interactions with an arginine side chain in galectin-3 are governed by steric-, desolvation and electronic conjugation effects†

Rohit Kumar,‡[a] Kristoffer Peterson,‡[b] Majda Misini Ignjatović,‡[c] Hakon Leffler,[d] Ulf Ryde,[c] Ulf J. Nilsson[b] and Derek T. Logan*[a]

In the β-D-galactopyranoside-binding protein galectin-3, synthetic inhibitors substituted at the 3-position of a thiodigalactoside core cause the formation of an aglycone binding pocket through the displacement of an arginine residue (Arg144) from its position in the apoprotein. To examine in detail the role of different molecular interactions in this pocket, we have synthesized a series of nine 3-(4-(2,3,5,6-tetrafluorophenyl)-1,2,3-triazol-1-yl)-thiogalactosides with different *para* substituents and measured their affinities to galectin-3 using a fluorescence polarization assay. High-resolution crystal structures (<1.3 Å) have been determined for five of the ligands in complex with the C-terminal domain of galectin-3. The binding affinities are rationalised with the help of the three-dimensional structures and quantum-mechanical calculations. Three effects seem to be involved: Firstly, the binding pocket is too small for the largest ligands with ethyl and methyl. Secondly, for the other ligands, the affinity seems to be determined mainly by desolvation effects, disfavouring the polar substituents, but this is partly counteracted by the cation–π interaction with Arg144, which stacks on top of the substituted tetrafluorophenyl group in all complexes. The results provide detailed insight into interactions of fluorinated phenyl moieties with arginine-containing protein binding sites and the complex interplay of different energetic components in defining the binding affinity.

Received 19th November 2018,
Accepted 2nd January 2019

DOI: 10.1039/c8ob02888e

rsc.li/obc

## 1. Introduction

Structure-based drug design relies on careful analysis of protein–ligand interactions and the structure and dynamics of ligand and binding sites. Improving binding affinity involves modulating the specific interactions that the ligand makes with the binding site by modifying or substituting chemical moieties in the ligand.[1] Investigating such specific interactions requires information about the protein–ligand complex that is often obtained from crystal structures and affinity data.[1] Structural analysis of protein–ligand complexes identifies potential binding interactions and steric restrictions, providing insight into design of new ligands with enhanced binding affinity. However, the energetic components contributing to the binding affinity are not always self-evident from an inspection of the crystal structure.

The drug target of interest here, galectin-3, belongs to the galectin super-family that has 14 members in humans. All galectins have a conserved carbohydrate recognition domain (CRD) that binds β-D-galactopyranosides, and the binding site is a shallow, hydrophilic pocket formed by β-sheets and loops.[2] Galectins are found everywhere in the cell. They are involved in cell growth, differentiation, cell-cycle regulation.[3] Their role in cancer, immunity and inflammatory conditions is well-documented, making them attractive therapeutic targets.[4–9] Galectins bind galactosides with affinities in the millimolar range. Suitable modifications of galactose at the C3 position to introduce specific groups improves the binding affinity drastically to micromolar and even nanomolar affinity. A wealth of structural data is available for the galectin-3 CRD in complex with different compounds[10–15] and we have recently reported the structures of high-affinity phenyltriazole thiogalactosides in

[a]*Biochemistry and Structural Biology, Centre for Molecular Protein Science, Department of Chemistry, Lund University, Box 124, SE-221 00 Lund, Sweden. E-mail: derek.logan@biochemistry.lu.se*
[b]*Centre for Analysis and Synthesis, Department of Chemistry, Lund University, Box 124, SE-221 00 Lund, Sweden*
[c]*Theoretical Chemistry, Department of Chemistry, Lund University, Box 124, SE-221 00 Lund, Sweden*
[d]*Department of Laboratory Medicine, Section MIG, Lund University BMC-C1228b, Klinikgatan 28, 221 84 Lund, Sweden*
†Electronic supplementary information (ESI) available. See DOI: 10.1039/c8ob02888e
‡These authors contributed equally.

complex with galectin-3.[16] This ample availability of structural and affinity data makes galectin-3 an excellent model protein for studying protein–ligand interactions. The high galectin-3 affinity of thiodigalactosides with mono- to trifluorinated 3-(4-aryl-1,2,3-triazol-1-yl) moieties at C3 has been explained using X-ray crystallography by orthogonal multipolar fluorine–amide interactions with backbone amides and a cation–π interaction with Arg144.[15] Arg144 is raised from its normal position in a water-mediated salt bridge on the surface of galectin-3 by the influence of fluorinated phenyl moieties on synthetic ligands, which creates a small pocket beneath Arg144 that could accommodate a larger substituent than fluorine in the *para* position on the phenyl ring.

Herein we report on a systematic probing of the binding interactions near Arg144 in galectin-3 by varying the *para* substituent on 2,3,5,6-tetrafluorophenyltriazoles through affinity measurements using fluorescence polarisation combined with structural analysis and quantum-mechanical calculations.

## 2. Experimental section

### 2.1 General

All reagents and solvents were dried prior to use according to standard methods. Commercial reagents were used without further purification. 2,3,5,6-tetrafluoro-4-hydroxyphenylacetylene was synthesized following a published procedure[17] for the alkene analogue and it used without purification. Analytical TLC was performed using on silica gel 60 $F_{254}$ (Merck) with detection by UV absorption and/or by charring following immersion in a 7% ethanolic solution of sulfuric acid. Purification of compounds was carried out by column chromatography on silica gel (40–60 μm, 60 Å) and/or preparative HPLC (Agilent 1260 infinity system, column SymmetryPrep-C18, 17 ml $min^{-1}$ $H_2O$–MeCN gradient 10–100% 15 min with 0.1% formic acid). Specific rotations were measured on a PerkinElmer model 341 polarimeter. NMR spectra $^1H$, $^{13}C$, $^{19}F$, 2D COSY, HMQC and HMBC were recorded with a Bruker Avance II 400 MHz spectrometer (400 Hz for $^1H$, 100 Hz for $^{13}C$ and 376 Hz for $^{19}F$) at ambient temperature. Chemical shifts are reported in $\delta$ parts per million (ppm). In the $^{13}C$ NMR spectra no signals were observed for the carbons in the fluorinated phenyl or the C4 triazole carbon, due to signal splitting caused by short- and long-range fluorine couplings. However, in the HMBC spectra the cross peak of the triazole C4 and H5 was observed (exemplified in the ESI† for compound **3**). HRMS was determined by direct infusion on a Waters XEVO-G2 QTOF mass spectrometer using electrospray ionization (ESI). Compounds **2–10** were of >95% purity according to HPLC-analysis (Agilent series 1100 system, column Eclipse XDB-C18, 0.8 ml $min^{-1}$ $H_2O$–MeCN gradient 5–95% 13 min with 0.1% trifluoroacetic acid).

### 2.2 Synthesis of compounds (2–5)

**2.2.1 *p*-Methylphenyl 3-deoxy-3-[4-(2,3,5,6-tetrafluoro-4-hydroxyphenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (2).** To a solution of compound **1** (18 mg, 0.058 mmol), 2,3,5,6-tetrafluoro-4-hydroxyphenylacetylene (16 mg, 0.087 mmol) and CuI (5 mg, 0.029 mmol) in MeCN (3 mL) was diisopropylethylamine (0.03 mL, 0.145 mmol) added. The mixture was stirred for 24 h at 50 °C before quenching with sat. aq. $NH_4Cl$ followed by evaporation of the solvent. The obtained residue was purified with column chromatography ($CH_2Cl_2$ : MeOH 14 : 1–>5 : 1) to give **2** (14 mg, 48%) as an amorphous white solid. $[\alpha]_D^{20}$ 56.6 (*c* 0.93, $CH_3OH$). $^1H$ NMR ($CD_3OD$, 400 MHz): $\delta$ 8.32 (s, 1H, Ph), 7.50 (d, *J* = 8.1 Hz, 2H, Ph), 7.15 (d, *J* = 8.1 Hz, 2H, Ph), 4.93 (obscured by water H-3), 4.78 (d, *J* = 9.5 Hz, 1H, H-1), 4.26 (t, *J* = 10.0 Hz, 1H, H-2), 4.16 (d, *J* = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$). $^{13}C$ NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 133.1, 131.6, 130.7, 125.5, 91.8, 80.9, 69.5, 69.2, 68.0, 62.3, 22. $^{19}F$ NMR ($CD_3OD$, 376 MHz): $\delta$ −145.6 (d, *J* = 16.0 Hz, 2F), −165.0 (d, *J* = 15.7 Hz, 2F). HRMS calculated for $[C_{21}H_{19}F_4N_3O_5SNa]^+$, 524.0879; found: 524.0880.

**2.2.2 *p*-Methylphenyl 3-deoxy-3-[4-(2,3,4,5,6-pentafluorophenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (3).** To a solution of compound **1** (185 mg, 0.59 mmol) and CuI (28 mg, 0.15 mmol) in MeCN (15 mL) was pentafluorophenylacetylene (0.14 mL, 0.89 mmol) and diisopropylethylamine (0.10 mL, 0.59 mmol) added. The mixture was stirred for 4.5 h at 50 °C before quenching with sat. aq. $NH_4Cl$ followed by evaporation of the solvent. The obtained residue was purified with column chromatography (heptane : EtOAc 1 : 1–>1 : 2) to give **3** (295 mg, 99%) as an amorphous white solid. $[\alpha]_D^{20}$ 57.6 (*c* 0.59, $CH_3OH$). $^1H$ NMR ($CD_3OD$, 400 MHz): $\delta$ 8.44 (s, 1H, Ph), 7.50 (d, *J* = 8.1 Hz, 2H, Ph), 7.15 (d, *J* = 8.1 Hz, 2H, Ph), 4.95 (obscured by water H-3), 4.78 (d, *J* = 9.5 Hz, 1H, H-1), 4.26 (t, *J* = 10.0 Hz, 1H, H-2), 4.16 (d, *J* = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$). $^{13}C$ NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 133.2, 131.6, 130.7, 126.3, 91.7, 80.9, 69.5, 69.3, 68.0, 62.3, 21.1. $^{19}F$ NMR ($CD_3OD$, 376 MHz): $\delta$ −142.3 (dd, *J* = 13.7, 7.0 Hz, 2F), −157.8 (t, *J* = 20.0 Hz, 1F), −165.0 (m, 2F). HRMS calculated for $[C_{21}H_{19}F_5N_3O_4S]^+$, 504.1016; found: 504.1019.

**2.2.3 *p*-Methylphenyl 3-deoxy-3-[4-(4-azido-2,3,5,6-tetrafluorophenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (4).** A mixture of compound **3** (25 mg, 0.050 mmol) and $NaN_3$ (5 mg, 0.074 mmol) in dry DMF (5 mL) was stirred at 60 °C for 2 days before water was added followed by extraction with EtOAc. The organic phase was washed with brine, dried, evaporated and the obtained residue was purified with column chromatography (heptane : EtOAc 2 : 3–>1 : 2) to give **4** (23 mg, 87%) as an amorphous white solid. $[\alpha]_D^{20}$ 53.8 (*c* 0.89, $CH_3OH$). $^1H$ NMR ($CD_3OD$, 400 MHz): $\delta$ 8.42 (s, 1H, Ph), 7.50 (d, *J* = 8.1 Hz, 2H, Ph), 7.15 (d, *J* = 8.1 Hz, 2H, Ph), 4.95 (dd, *J* = 10.5, 3.0 Hz, 1H, H-3), 4.78 (d, *J* = 9.5 Hz, 1H, H-1), 4.26 (t, *J* = 10.0 Hz, 1H, H-2), 4.16 (d, *J* = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$). $^{13}C$ NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 133.2, 131.6, 130.7, 126.2, 91.7, 80.9, 69.5, 69.3, 68.0, 62.3, 21.1. $^{19}F$ NMR ($CD_3OD$, 376 MHz): $\delta$ −143.2 (dd, *J* = 20.0, 9.0 Hz, 2F), −154.6 (dd, *J* = 20.0, 9.0 Hz, 2F). HRMS calculated for $[C_{21}H_{19}F_4N_6O_4S]^+$, 527.1125; found: 527.1124.

**2.2.4 *p*-Methylphenyl 3-deoxy-3-[4-(4-amino-2,3,5,6-tetrafluorophenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (5).** To a solution of compound **4** (12 mg, 0.023 mmol) in dry MeOH (2 mL) was 1,3-propanedithiol (0.009 mL, 0.91 mmol) added followed by $Et_3N$ (0.013 mL, 0.091 mmol) and the mixture was stirred at rt for 1.5 h. The volatiles were evaporated and the obtained residue was purified with column chromatography (heptane : EtOAc 1 : 1–>1 : 2) to give **5** (11 mg, 96%) as an amorphous white solid. $[\alpha]_D^{20}$ 58.9 (*c* 0.79, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.25 (s, 1H, Ph), 7.50 (d, $J$ = 8.1 Hz, 2H, Ph), 7.15 (d, $J$ = 8.1 Hz, 2H, Ph), 4.91 (obscured by water H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.25 (t, $J$ = 10.0 Hz, 1H, H-2), 4.16 (d, $J$ = 2.8 Hz, 1H, H-4), 3.83–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 133.1, 131.7, 130.7, 125.0, 91.8, 80.9, 69.5, 69.2, 68.0, 62.3, 21.1. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −146.5 (m, 2F), −164.9 (m, 2F). HRMS calculated for $[C_{21}H_{20}F_4N_4O_4SNa]^+$, 523.1039; found: 523.1034.

### 2.3 General procedure for the preparation of compounds (6–10)

Method A for compounds **6–7**: Compound **3** (25 mg, 0.050 mmol) was dissolved in ROH (3 mL) and NaOR (1 M, 1 mL) and stirred for 2 days at rt before quenching with dowex. The mixture was filtered and following evaporation of the filtrate the residue was purified with column chromatography (heptane : EtOAc 1 : 1–>1 : 2).

Method B for compounds **8–10**: A mixture of compound **3** (20 mg, 0.040 mmol) and $K_2CO_3$ (16.5 mg, 0.12 mmol), amine (*x*, 3 equiv.) and DMF (3 mL) was stirred for (*t*) time at 50 °C. After evaporation of the solvent the residue was purified with column chromatography (heptane : EtOAc 1 : 1–>1 : 2).

**2.3.1 *p*-Methylphenyl 3-deoxy-3-[4-(2,3,5,6-tetrafluoro-4-methoxyphenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (6).** Method A, R = Me, Yield 18.0 mg, 70%. $[\alpha]_D^{20}$ 35.2 (*c* 0.91, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.39 (s, 1H, Ph), 7.50 (d, $J$ = 8.0 Hz, 2H, Ph), 7.15 (d, $J$ = 8.0 Hz, 2H, Ph), 4.94 (dd, $J$ = 10.5, 3.0 Hz, 1H, H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.26 (t, $J$ = 9.6 Hz, 1H, H-2), 4.16 (d, $J$ = 2.8 Hz, 1H, H-4), 4.13 (s, 3H, $OCH_3$), 3.84–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 135.4, 133.2, 131.6, 130.7, 125.9, 91.8, 80.9, 69.5, 69.3, 68.0, 62.9, 62.3, 21.1. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −144.1 (dd, $J$ = 19.3, 7.1 Hz, 2F), −160.3 (dd, $J$ = 19.3, 7.0 Hz, 2F). HRMS calculated for $[C_{22}H_{21}F_4N_3O_5SNa]^+$, 538.1030; found: 538.1035.

**2.3.2 *p*-Methylphenyl 3-deoxy-3-[4-(4-ethoxy-2,3,5,6-tetrafluorophenyl)-1*H*-1,2,3-triazol-1-yl]-1-thio-β-D-galactopyranoside (7).** Method A, R = Et, Yield 13.4 mg, 50%. $[\alpha]_D^{20}$ 33.7 (*c* 0.83, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.39 (s, 1H, Ph), 7.50 (d, $J$ = 8.0 Hz, 2H, Ph), 7.15 (d, $J$ = 8.0 Hz, 2H, Ph), 4.94 (dd, $J$ = 10.5, 3.0 Hz, 1H, H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.37 (q, $J$ = 7.0 Hz, 2H, $CH_2$), 4.26 (t, $J$ = 9.6 Hz, 1H, H-2), 4.16 (d, $J$ = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 2.33 (s, 3H, $CH_3$), 1.43 (t, $J$ = 7.0 Hz, 3H, $CH_3$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 135.4, 133.2, 131.6, 130.7, 125.9, 91.8, 80.9, 72.3, 69.5, 69.3, 68.0, 62.9, 62.3, 21.1, 15.7. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −144.2 (dd, $J$ = 19.3, 7.0 Hz, 2F), −159.5 (dd, $J$ = 19.5, 7.1 Hz, 2F). HRMS calculated for $[C_{23}H_{23}F_4N_3O_5SNa]^+$, 552.1187; found: 552.1190.

**2.3.3 *p*-Methylphenyl 3-deoxy-3-{4-[2,3,5,6-tetrafluoro-4-(methylamino)phenyl]-1*H*-1,2,3-triazol-1-yl}-1-thio-β-D-galactopyranoside (8).** Method B, x = methylamine 33 wt% in EtOH, *t* = 3 days. Yield 13.1 mg, 64%. $[\alpha]_D^{20}$ 55.6 (*c* 0.90, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.26 (s, 1H, Ph), 7.50 (d, $J$ = 8.0 Hz, 2H, Ph), 7.15 (d, $J$ = 8.0 Hz, 2H, Ph), 4.91 (obscured by water H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.26 (t, $J$ = 9.6 Hz, 1H, H-2), 4.16 (d, $J$ = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 3.09 (t, $J$ = 2.7 Hz, 3H, $NCH_3$), 2.33 (s, 3H, $CH_3$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 136.4, 133.1, 131.7, 130.7, 124.9, 91.8, 80.9, 69.5, 69.2, 68.0, 62.3, 32.5, 21.1. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −146.1 (dd, $J$ = 23.3, 10.5 Hz, 2F), −164.1 (d, $J$ = 16.1 Hz, 2F). HRMS calculated for $[C_{22}H_{23}F_4N_4O_4SNa]^+$, 537.1196; found: 537.1199.

**2.3.4 *p*-Methylphenyl 3-deoxy-3-{4-[2,3,5,6-tetrafluoro-4-(dimethylamino)phenyl]-1*H*-1,2,3-triazol-1-yl}-1-thio-β-D-galactopyranoside (9).** Method B, x = dimethylamine 2 M in THF, *t* = 4 days. Yield 6.2 mg, 29%. $[\alpha]_D^{20}$ 48.7 (*c* 0.78, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.34 (s, 1H, Ph), 7.50 (d, $J$ = 8.0 Hz, 2H, Ph), 7.15 (d, $J$ = 8.0 Hz, 2H, Ph), 4.93 (obscured by water H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.26 (t, $J$ = 9.6 Hz, 1H, H-2), 4.16 (d, $J$ = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 3.02 (t, $J$ = 2.2 Hz, 6H, $NCH_3$), 2.33 (s, 3H, $CH_3$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 136.0, 133.2, 131.6, 130.7, 125.5, 91.8, 80.9, 69.5, 69.2, 68.0, 62.3, 43.5, 21.1. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −145.0 (dd, $J$ = 18.6, 7.4 Hz, 2F), −153.8 (d, $J$ = 13.7 Hz, 2F). HRMS calculated for $[C_{23}H_{25}F_4N_4O_4S]^+$, 529.1533; found: 529.1532.

**2.3.5 *p*-Methylphenyl 3-deoxy-3-{4-[2,3,5,6-tetrafluoro-4-(pyrrolidin-1-yl)phenyl]-1*H*-1,2,3-triazol-1-yl}-1-thio-β-D-galactopyranoside (10).** Method B, x = pyrrolidine, *t* = 36 h. Yield 21.9 mg, 99%. $[\alpha]_D^{20}$ 40.3 (*c* 0.67, $CH_3OH$). $^1$H NMR ($CD_3OD$, 400 MHz): $\delta$ 8.27 (s, 1H, Ph), 7.50 (d, $J$ = 8.0 Hz, 2H, Ph), 7.15 (d, $J$ = 8.0 Hz, 2H, Ph), 4.91 (obscured by water H-3), 4.78 (d, $J$ = 9.5 Hz, 1H, H-1), 4.25 (t, $J$ = 9.6 Hz, 1H, H-2), 4.15 (d, $J$ = 2.8 Hz, 1H, H-4), 3.84–3.70 (m, 3H, H-5 and H-6), 3.66 (m, 4H, $CH_2$), 2.33 (s, 3H, $CH_3$), 1.96 (m, 4H, $CH_2$). $^{13}$C NMR ($CD_3OD$, 100 MHz): $\delta$ 138.8, 136.5, 133.1, 131.7, 130.7, 124.9, 91.8, 80.9, 69.5, 69.2, 68.0, 62.3, 52.4, 26.7, 21.1. $^{19}$F NMR ($CD_3OD$, 376 MHz): $\delta$ −145.6 (dd, $J$ = 22.2, 9.3 Hz, 2F), −158.0 (d, $J$ = 15.1 Hz, 2F). HRMS calculated for $[C_{25}H_{27}F_4N_4O_4S]^+$, 555.1689; found: 555.1688.

### 2.4 Competitive fluorescence polarization experiments determining galectin-3 affinities

Human galectin-3 was expressed and purified as earlier described.[18] Fluorescence polarization experiments were performed on a PheraStarFS plate reader with software PHERAstar Mars version 2.10 R3 (BMG, Offenburg, Germany) and fluorescence anisotropy of fluorescein tagged probes measured with excitation at 485 nm and emission at 520 nm. Experiments were performed at 20 °C with galectin-3 at 0.20 μM and the fluorescent probe 3,3′-dideoxy-3-[4-(fluorescein-5-yl-carbonylaminomethyl)-1*H*-1,2,3-triazol-1-yl]-3′-(3,5-

di-methoxybenzamido)-1,1′-sulfanediyl-di-β-D-galactopyranoside[19] ($K_d$ 80 nM) at 0.02 μM as previously described.[10,15,19] Compounds were dissolved in neat DMSO at 20 mM and diluted in PBS to **3–6** different concentrations to be tested in duplicate. $K_d$ averages and SEM were calculated from 4 to 25 single-point measurements from at least two independent experiments showing between 20–80% inhibition.

### 2.5 Crystallization of galectin-3 C-terminal domain with compounds (2–5) and (8)

Solutions of the C-terminal CRD of galectin-3C[2] (19.2 mg ml$^{-1}$ in 10 mM phosphate pH 7.4, 100 mM NaCl, 10 mM β-mercaptoethanol and 2 mM EDTA) were mixed with crystallization solution (20% PEG 4000, 0.1 M Tris/HCl pH 7.5, 0.4 M NaSCN, 7.9 mM β-mercaptoethanol). Crystallization drops of 2 + 2 μL were set up over 0.5 mL reservoir solution. The crystals obtained were soaked with compounds. Compounds **2–5** and **8** were dissolved in DMSO to obtain highly concentrated stocks. These stocks were then diluted with PEG400 (final concentration 30%), as the compounds were highly insoluble in water, then a ligand cocktail was prepared using crystallization reservoir and the ligand stock to obtain a final compound concentration of 10 mM. Crystals were placed in 4 μl of these cocktails and left for 15–20 hours. These soaked crystals were flash-cooled in cryoprotectant solution (15% PEG400, 25.5 w/v % PEG 4000, 250 mM NaSCN, 85 mM Tris/HCl pH 7.5, 2.5 mM ligand concentration).

### 2.6 Data collection and structure solution of galectin-3C in complex with compounds (2–5) and (8)

Data for compounds **3–5** and **8** were collected at 100 K at station I911-3 of the MAX-II synchrotron, Lund, Sweden ($\lambda$ = 1.0000 Å), equipped with a marMosaic 225 mm CCD detector. 300–360 images with 0.5° rotation and 1–3 seconds exposure times were collected for **3–5** and **8**. Data for **2** were collected at ID23-2, ESRF, France on a DECTRIS PILATUS3 2M detector. 600 images were collected with 0.5° rotation and 0.2 seconds exposure time. Data for all structures were integrated using XDS and scaled using XSCALE.[21] The structures were refined using phenix.refine[22] and PDB entry 3ZSL (stripped of water molecules and alternate conformations) as starting model, first by rigid-body refinement. Five percent of the total reflections chosen at random were set aside for cross validation. The models were then subjected to model building and maximum likelihood refinement, gradually increasing the resolution to the highest resolutions with anisotropic $B$ factors. After initial refinement of the protein coordinates in phenix.refine,[22] the coordinates of **2–5** and **8** were fitted to the electron density using Coot.[23] Further model building and manipulations were done in Coot. Restraints were generated using eLBOW[24] from Phenix for **3–7**. The structures were refined until convergence and individual anisotropic atomic displacement parameters for each atom were refined. Water molecules were added to positive difference density peaks more than 4.5 or 5 σ above the mean and also present in the $2m|F_o| - D|F_c|$ map at the 1 σ level. Riding hydrogen atoms were added in the final stages of refinement. Refinement statistics are listed in Table S1.† Molecular images were generated using PyMOL (Schrodinger LLC). Model validation and analysis were performed using MolProbity[25] and PDB_REDO.[26] Coordinates have been deposited in the Protein Data Bank with accession numbers 6I75 for compound **2**, 6I74 for compound **3**, 6I76 for compound **4**, 6I77 for compound **5** and 6I78 for compound **8**. For detailed structure refinement statistics, please refer to Table 1 in the ESI.†

### 2.7 Quantum mechanical calculations

Four sets of quantum-mechanical (QM) calculations at different levels of theory were employed to obtain energies that can help to explain the differences in binding affinity of compounds **2–9** to galectin-3. All QM calculations were performed with the Turbomole 7.2 software.[27,28] In all systems, ligands were modelled as the isolated fluorine-substituted benzene moiety by replacing the remaining part of the ligand with a hydrogen atom.

In the first set, we calculated the energy of rotating the variable group of compounds **2**, **4**, **5**, **6** and **8** out of the plane of the tetrafluorophenyl group by changing the value of a C–C–X–Y dihedral angle from 0° to 90° in increments of 10°, where the first and the second carbon atom belong to the ring, whereas the X and Y atoms belong to the varying group (in case of compound **8**, Y atom is carbon). For each dihedral angle value, optimization of all the other degrees of freedom was performed at the B3LYP-D3/def2-SV(P) level of theory.[29–33]

In the next two sets of calculations, we calculated the interaction energy between ligands and three nearby residues, Ser237–Gly238 and Arg144. We performed separate calculations for each of the two residues. The amide group of Ser237–Gly238 was modelled as $CH_3$–CO–NH–$CH_3$, and the side chain of Arg144 was modelled as $[CH_3–NH–C(NH_2)_2]^+$. The coordinates were taken from the crystal structures. Arg144 has two conformations in complex with compounds **3** and **5**, and for these, we performed separate calculations on both conformations. The calculations were performed at the TPSS-D3/def2-TZVP level of theory.[32,34,35] The interaction energy for each compound–residue system was calculated from three single-point calculations as $\Delta E = E_{complex} - E_{residue} - E_{ligand}$.

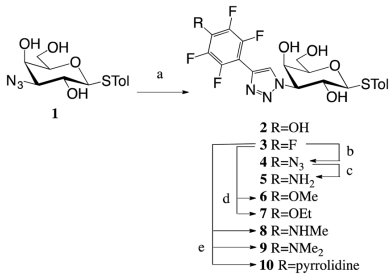
Finally, we calculated the solvation free energies for compounds **2–9**, using the conductor-like screening model for real solvents (COSMO-RS),[36,37] with the dielectric constant for water $\varepsilon_r$ = 80 and optimized radii for all atoms.[38] These calculations were based on two single point BP86 calculations[29,35,39] with the TZVP basis set,[40] as is requested by the method, one in vacuum and one in a continuum solvent with an infinite dielectric constant.[36,37]

## 3. Results and discussion

### 3.1 Synthesis and galectin-3 affinities of 3-(4-aryl-1,2,3-triazol-1-yl)-thiogalactosides

A 1,3-dipolar cycloadditions with alkynes and azide **1**[16] produced penta- and tetrafluoroaryltriazoles **2–3** (Scheme 1).




**Scheme 1** Synthesis of triazoles **2–10**. Reagents and conditions: (a) Alkyne, CuI, DIPEA, MeCN, 50 °C; (b) $NaN_3$, DMF, 60 °C; (c) 1,3-propanedithiol, $Et_3N$, MeOH, rt; (d) NaR, HR, rt; (e) amine, $K_2CO_3$, DMF, 50 °C. Tol = *p*-methylphenyl.

Nucleophilic aromatic substitution of the *p*-fluorine in **3** with alcohols, amines and $NaN_3$ gave tetrafluoroaryltriazoles **4** and **6–10**, while reduction of azide **4** resulted in amine **5**.

The inhibition potencies of thiogalactosides **2–10** were evaluated towards galectin-3 using a previously described competitive fluorescence polarization assay[10,20] and the results are presented in Table 1. The pentafluorophenyl **3** had an affinity of 3.4 μM to galectin-3. Any replacement of the fluorine in the *para* position led to a drop in affinity. Replacing the fluorine with an amine (**5**) or azide (**4**) resulted in a 2–3-fold decrease, while replacement with a hydroxyl (**2**) resulted in a 7-fold decrease. Adding methyl groups (**8–9**) to amine **5** further decreased the affinity 2-fold per methyl group, while adding a methyl group (**6**) to the hydroxide **2** did not affect the affinity. Fluorine replacement with a bulkier ethoxy group (**7**) resulted in an almost 5-fold decrease in affinity compared to methoxide **6**, which is indicative of steric restrictions in the binding pocket. This is further demonstrated by the even bulkier pyrrolidine (**10**) that does not bind galectin-3 at all at the concentrations tested.

**Table 1** $K_d$ (μM) values for aryl triazoles **2–10** and thiodigalactoside as a reference compound, determined by a competitive fluorescence polarization assay

| | *R* | $K_d$ |
|---|---|---|
| **2** | OH | 23 ± 1.7 |
| **3** | F | 3.4 ± 0.21 |
| **4** | $N_3$ | 8.5 ± 1.2 |
| **5** | $NH_2$ | 11 ± 0.6 |
| **6** | OMe | 18 ± 2.1 |
| **7** | OEt | 88 ± 12 |
| **8** | NHMe | 18 ± 0.9 |
| **9** | $NMe_2$ | 40 ± 3.3 |
| **10** | Pyrrolidine | >300[a] |
| | Thiodigalactoside | 49 (ref. 41) |

[a] Does not bind galectin-3 at this concentration.

### 3.2 Structural analysis of thiogalactosides 2–5 and 8 in complex with the galectin-3 CRD (galectin-3C)

In order to further investigate the binding interactions in the pocket below Arg144, high-resolution X-ray structures (all <1.3 Å resolution; see Table S1 in the ESI†) of thiogalactosides **2–5** and **8** in complex with galectin-3C were determined. X-ray structures of thiogalactosides **6–7** and **9–10** could not be obtained owing to solubility issues. The structures revealed a virtually identical binding mode for the galactose unit as earlier reported in many publications, and the triazole extends the *para*-substituted tetrafluorophenyl group into the pocket below Arg144.[15]

The superimposition of all crystal structures (Fig. 1) shows that the ligands reside in the binding pocket in a similar manner to that reported previously. The anomeric *S*-tolyl group of the ligands is disordered, and in this work, focus is on the phenyl group below Arg144 and its *para*-substituents. Fig. 1 also shows that Arg144 adopts two principal conformations, either directly above the phenyl ring or above the *para* substituent, depending on the nature of the phenyl substitutions. Arg144 has split occupancy in the crystal structures of **3** and **5**. This is likely due to a weakened cation–π interaction as a result of the electron-withdrawing fluorines. The *N*-methyl group in **8** is oriented above the phenyl ring plane towards Arg144. As a result, Arg144 resides only above the phenyl ring in this complex, while for both **2** and **4**, Arg144 shows a single conformation directly above the *para* substituent.

Compound **2** has the lowest affinity among the successfully crystallized ligands. The phenolic proton in **2** is, based on the $pK_a$ value of 5.7 for pentafluorophenol,[42] likely to be deprotonated, and the resulting anion could interact favourably with the cation of Arg144. However, as will be discussed below, it will be more disfavoured by desolvation effects than the other ligands. Besides the interaction with Arg144, the

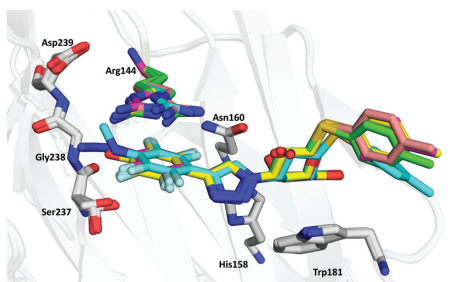



**Fig. 1** Superimposed view of the five crystal structures showing the ligand and neighbouring protein residues. The galactose moiety forms a hydrophobic stacking interaction with Trp181, the triazole linker extends the tetrafluorophenyl group in to the pocket near Arg144, which makes a cation–π interaction with the tetrafluorophenyl group. Ligands **2**, **3**, **4**, **5**, and **8** are shown in yellow, green, purple, magenta and light blue colours, respectively (also for Arg144).

binding site is unable to stabilise the negative charge on the ligand. For example, Ser237 shows two conformations as for the other ligands, and only one of these forms a rather weak hydrogen bond to the hydroxyl group of the ligand (O–O distance of 3.4 Å).

The azide in **4** is located in-plane with the phenyl ring and pointing outwards to solution, because the connecting nitrogen is $sp^2$ hybridized and orienting it inwards to the protein would result in a steric clash with Ile145. The single occupancy of Arg144 in the complex with **4** may be due to a more favourable interaction with the π-system of the azide than with the phenyl ring of **3**. The azide group is within hydrogen bonding distance of three water molecules (Fig. 2c), which may stabilize the ligand, resulting in better affinity than other compounds except **3**. Most of these water molecules are present at very similar positions in all complexes, but they make more interactions with **4** than with the other ligands.

Compound **3** has the highest affinity, showing that fluorine is the best candidate at the *para* position. This fluorine atom forms multipolar orthogonal interactions with a nearby peptide bond (Gly238; Fig. 2f) which increases the affinity. The fluorine atom is at a distance of 3.0 Å from the N atom of the backbone and 3.6 Å from the carbonyl C atom.

### 3.3 Quantum mechanical calculations

To rationalize the affinities and the structural observations, we have made a number of quantum mechanical (QM) calculations with compounds **2**–**9**. First, we calculated the potential-energy surface for rotation of the varying *para* substituent of **2**, **4**–**6**, and **8** out of the plane of the tetrafluorophenyl group (Fig. 3). It can be seen that ligands with OH (**2**) and $N_3$ (**4**) attain their energy minima with the substituent in the plane of the phenyl group. This is in accordance with the crystal structure of the azide **4** in complex with galectin-3. The other three ligands (**5**, **6**, and **8**) attain a shallow minimum around ~20°, which probably reflects a competition between conjugation (favouring an angle of 0°) and the bulk of the methyl groups of **6** and **8**, which prefer a larger dihedral. The figure indicates that ligand **8** with NHMe, for which the angle is 72° in the crystal structure, is strained by ~8 kJ $mol^{-1}$, which might be compensated by polar interactions with the NH groups, although no such interactions are obvious from the crystal structure.

Next, we calculated the interaction energy between the pentafluorophenyl group of **3** and the backbone amide group of Ser237–Gly238 (modelled by $CH_3$–CO–NH–$CH_3$ with coordi-

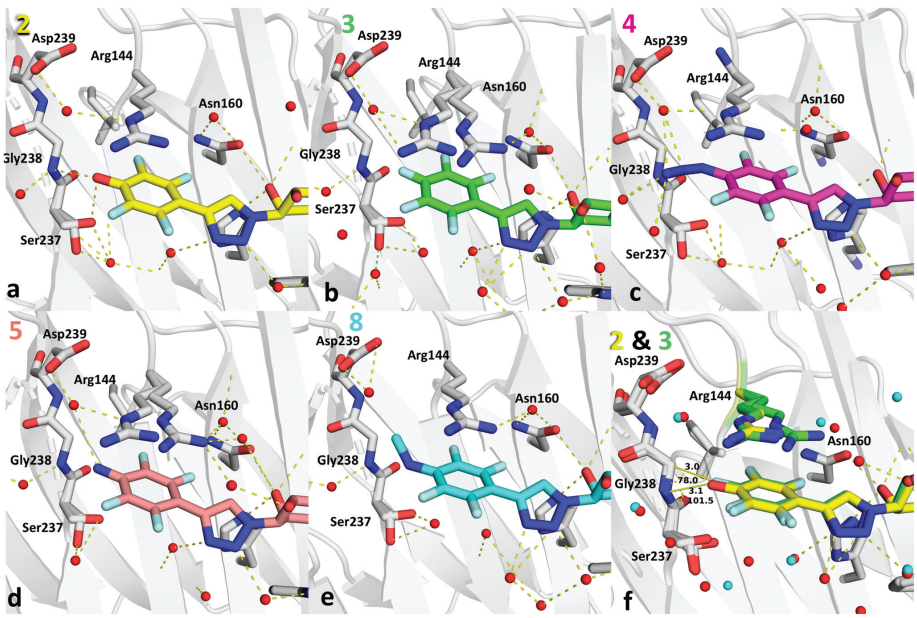



**Fig. 2** Close-up view of the binding pocket in the crystal structures of galectin-3C in complex with phenyltriazoles (a) **2**, (b) **3**, (c) **4**, (d) **5** and (e) **8**. (f) Superimposed view of **2** and **3** showing the important hydrogen bonded water and fluorine–amide interactions. Water in **3** is coloured red and water in **2** is coloured cyan.

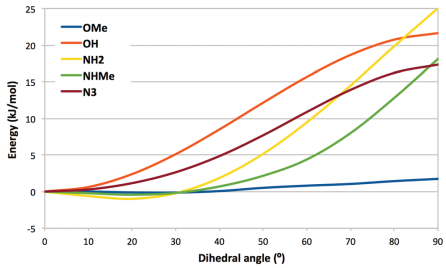



Fig. 3 Potential-energy surface for the C–C–X–Y dihedral angle of the varying group in compounds **2** (OH), **4** ($N_3$), **5** ($NH_2$), **6** (OMe) and **8** (NHMe).

nates taken from the crystal structure). It was 6 kJ $mol^{-1}$, which is of the expected size for an F–amide interaction.[43] On the other hand, ligand **5** gave the same interaction energy (6 kJ $mol^{-1}$) with an amide group.

Third, we calculated the COSMO-RS solvation free energies of the nine *para*-substituted tetrafluorophenyl groups **2–9** (not pyrrolidine **10**). The results are shown in Fig. 4 as a function of the measured binding affinities. It can be seen that for the OEt (**7**), $NMe_2$ (**6**) and OMe (**9**), the estimated solvation free energies are small, −1 to −3 kJ $mol^{-1}$, and there is no correlation with the binding free energies. However, for the F (**3**), $N_3$ (**4**), $NH_2$ (**5**) and OH (**2**) substituents, there is a good negative correlation to the binding affinity ($R$ = −0.95). The ligand with the NHMe (**8**) substituent also falls close to the correlation line, albeit reducing the correlation to −0.83 if included. Taken together, these results indicate that the observed affinities can be explained by two effects. The alkylated substituents, especially OEt (**7**), are too large and steric effects give a low affinity, decreasing further with the number of methyl groups

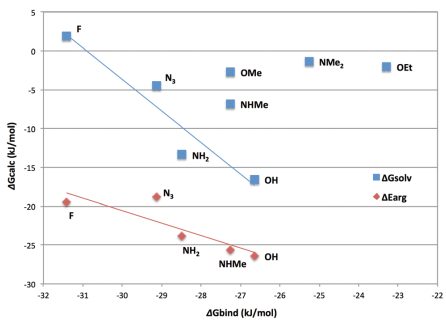



Fig. 4 Solvation free energies (blue squares) and interaction energies between Arg144 and the substituted tetraphenyl group (red diamonds) for compounds **2–9**.

on the substituent. For the other ligands, the affinity is determined by desolvation effects: in the binding site, the ligand is partly buried by the protein and is less solvated than in water solution. This desolvation is modest for the F and $N_3$ ligands (**3** and **4**), which form poor hydrogen bonds and therefore give low solvation energies (>−5 kJ $mol^{-1}$). However, for the $NH_2$ and OH ligands (**5** and **2**), the effect is large and pronounced. The effect would be even larger if the OH ligand **2** is deprotonated (the calculated solvation energy is −240 kJ $mol^{-1}$). For the NHMe ligand **8**, both steric and desolvation effects seem to be significant.

Finally, we also calculated the interaction energy between Arg144 and the substituted tetrafluorophenyl groups, using the geometry from the crystal structure (and two different Arg144 conformations for the F and $NH_2$ ligands **3** and **5**). The results are also included in Fig. 4 (red symbols and line). It can be seen that all ligands give a large cation–$\pi$ interaction energy of 18–27 kJ $mol^{-1}$. All groups give lower interaction energies with Arg144 than an unsubstituted benzene group (−37 kJ $mol^{-1}$). The interaction energies of the two Arg144 conformations for the F and $NH_2$ ligands **3** and **5** differ by 3–6 kJ $mol^{-1}$ (compared to 1 kJ $mol^{-1}$ for benzene, using the two conformations for pentafluorophenyl **3**). The average interaction energies show a good anti-correlation with the ligand-binding affinities ($R$ = −0.87) and a correlation with the solvation energies ($R$ = 0.78), illustrating that all three depend on the polarity of the ligand. Thus, the interaction with Arg144 partly counteracts the desolvation penalty and the difference of these two energies give an excellent anti-correlation to the binding free energies of −0.91, although the slope is rather large at 1.7 (the same as that of the interaction with Arg144, but half as large as that of the solvation free energy).

## 4. Conclusions

A series of 2,3,5,6-tetrafluorophenyl derivatives **2–10** with different *para* substituents were synthesized to analyse in detail the binding interactions within a small pocket beneath Arg144 in galectin-3. The most potent *para* substituent was the fluorine (**3**) that forms a fluorine–amide interaction with the backbone amide of Ser237–Gly238. However, the QM interaction energy between the backbone of Ser237–Gly238 and ligand **3** is not larger than for some of the other ligands, *e.g.* **5**. Instead, the relative affinities seem to be determined by three effects: First, the pocket beneath Arg144 is not large enough for bulkier groups, *e.g.* –$NMe_2$ (**9**) and –OEt (**7**). Second, the solvation energy decreases strongly in the series **2–5–4–3** (OH–$NH_2$–$N_3$–F), implying that the desolvation penalty also decreases in this series, closely following the affinities of these ligands. However, this effect is partly counteracted by the interaction energy of the substituted tetrafluorophenyl group with Arg144, which becomes less favourable in this series (*cf.* Fig. 4). Taken together, given the frequency of employing fluorinated phenyl moieties and substituted derivatives thereof in drug design and drug discovery, the results presented here

provide further in-depth insight into the sometimes conflicting driving forces behind such the interactions of such moieties in arginine-containing protein binding sites.

## Conflicts of interest

UJN and HL are shareholders in Galecto Biotech AB, a company developing galectin inhibitors.

## Acknowledgements

This work was supported by the Swedish Research Council (Grant No. 621-2009-5326, 621-2012-2978 and 2014-5540), the Royal Physiographic Society, Lund, the European Community's Seventh Framework Program (FP7-2007–2013) under grant agreement no. HEALTH-F2-2011-256986 – project acronym PANACREAS, and a project grant awarded by the Knut and Alice Wallenberg Foundation (KAW 2013.0022). We thank staff at the I911-3 beamline of the MAX IV Laboratory and the ID23-2 beamline of the ESRF, for beam time and assistance in data collection. The calculations were performed on computer resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lunarc at Lund University and HPC2N at Umeå University.

## References

1 C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**, 5061–5084.
2 J. Seetharaman, A. Kanigsberg, R. Slaaby, H. Leffler, S. H. Barondes and J. M. Rini, *J. Biol. Chem.*, 1998, **273**, 13047–13052.
3 H. Leffler, S. Carlsson, M. Hedlund, Y. Qian and F. Poirier, *Glycoconjugate J.*, 2002, **19**, 433–440.
4 V. L. Thijssen, R. Heusschen, J. Caers and A. W. Griffioen, *Biochim. Biophys. Acta, Rev. Cancer*, 2015, **1855**, 235–247.
5 A. U. Newlaczyl and L. G. Yu, *Cancer Lett.*, 2011, **313**, 123–128.
6 F. T. Liu, R. J. Patterson and J. L. Wang, *Biochim. Biophys. Acta, Gen. Subj.*, 2002, **1572**, 263–273.
7 S. Di Lella, V. Sundblad, J. P. Cerliani, C. M. Guardia, D. A. Estrin, G. R. Vasta and G. A. Rabinovich, *Biochemistry*, 2011, **50**, 7842–7857.
8 G. A. Rabinovich and D. O. Croci, *Immunity*, 2012, **36**, 322–335.
9 A. C. MacKinnon, M. A. Gibbons, S. L. Farnworth, H. Leffler, U. J. Nilsson, T. Delaine, A. J. Simpson, S. J. Forbes, N. Hirani, J. Gauldie and T. Sethi, *Am. J. Respir. Crit. Care Med.*, 2012, **185**, 537–546.
10 I. Cumpstey, S. Carlsson, H. Leffler and U. J. Nilsson, *Org. Biomol. Chem.*, 2005, **3**, 1922–1932.
11 P. Sörme, P. Arnoux, B. Kahl-Knutsson, H. Leffler, J. M. Rini and U. J. Nilsson, *J. Am. Chem. Soc.*, 2005, **127**, 1737–1743.
12 C. Diehl, O. Engström, T. Delaine, M. Håkansson, S. Genheden, K. Modig, H. Leffler, U. Ryde, U. J. Nilsson and M. Akke, *J. Am. Chem. Soc.*, 2010, **132**, 14577–14589.
13 P. M. Collins, C. T. Öberg, H. Leffler, U. J. Nilsson and H. Blanchard, *Chem. Biol. Drug Des.*, 2012, **79**, 339–346.
14 V. K. Rajput, A. MacKinnon, S. Mandal, P. Collins, H. Blanchard, H. Leffler, T. Sethi, H. Schambye, B. Mukhopadhyay and U. J. Nilsson, *J. Med. Chem.*, 2016, **59**, 8141–8147.
15 T. Delaine, P. Collins, A. MacKinnon, G. Sharma, J. Stegmayr, V. K. Rajput, S. Mandal, I. Cumpstey, A. Larumbe, B. A. Salameh, B. Kahl-Knutsson, H. van Hattum, M. van Scherpenzeel, R. J. Pieters, T. Sethi, H. Schambye, S. Oredsson, H. Leffler, H. Blanchard and U. J. Nilsson, *ChemBioChem*, 2016, **17**, 1759–1770.
16 K. Peterson, R. Kumar, O. Stenström, P. Verma, P. R. Verma, M. Håkansson, B. Kahl-Knutsson, F. Zetterberg, H. Leffler, M. Akke, D. T. Logan and U. J. Nilsson, *J. Med. Chem.*, 2018, DOI: 10.1021/acs.jmedchem.7b01626.
17 I. Dimitrov, K. Jankova and S. Hvilsted, *J. Fluorine Chem.*, 2013, **149**, 30–35.
18 S. M. Massa, D. N. Cooper, H. Leffler and S. H. Barondes, *Biochemistry*, 1993, **32**, 260–267.
19 E. Salomonsson, A. Larumbe, J. Tejler, E. Tullberg, H. Rydberg, A. Sundin, A. Khabut, T. Frejd, Y. D. Lobsanov, J. M. Rini, U. J. Nilsson and H. Leffler, *Biochemistry*, 2010, **49**, 9518–9532.
20 P. Sörme, B. Kahl-Knutsson, M. Huflejt, U. J. Nilsson and H. Leffler, *Anal. Biochem.*, 2004, **334**, 36–47.
21 W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 125–132.
22 P. V. Afonine, W. Ralf, J. J. Headd and C. Thomas, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2012, **68**, 352–367.
23 P. Emsley, B. Lohkamp, W. G. Scott and K. Cowtan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 486–501.
24 N. W. Moriarty, R. W. Grosse-Kunstleve and P. D. Adams, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2009, **65**, 1074–1080.
25 V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 12–21.
26 R. P. Joosten, F. Long, G. N. Murshudov and A. Perrakis, *IUCrJ*, 2014, **1**, 213–220.
27 F. Furche, R. Ahlrichs, C. Hättig, W. Klopper, M. Sierka and F. Weigend, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 91–100.
28 *TURBOMOLE version 7.1*, 2016.
29 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
30 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
31 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
32 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
33 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104–115123.

34 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
35 M. Sitarz, E. Wirth-Dzieciolowska and P. Demant, *Neoplasma*, 2000, **47**, 148–150.
36 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.
37 A. Schäfer, A. Klamt, D. Sattel, J. C. W. Lohrenz and F. Eckert, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2187–2193.
38 A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074–5085.
39 J. P. Perdew, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1986, **33**, 8822–8824.
40 A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571–2577.
41 I. Cumpstey, E. Salomonsson, A. Sundin, H. Leffler and U. J. Nilsson, *Chem. – Eur. J.*, 2008, **14**, 4233–4245.
42 D. A. Kraut, P. A. Sigala, B. Pybus, C. W. Liu, D. Ringe, G. A. Petsko and D. Herschlag, *PLoS Biol.*, 2006, **4**, e99.
43 P. Zhou, J. Zou, F. Tian and Z. Shang, *J. Chem. Inf. Model.*, 2009, **49**, 2344–2355.

# Paper VII

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Cite This: J. Am. Chem. Soc. 2019, 141, 2012−2026
Article
pubs.acs.org/JACS

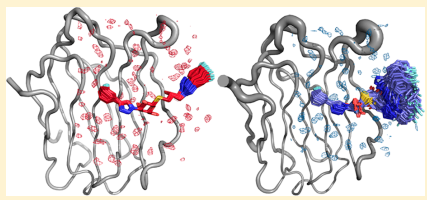
# Interplay between Conformational Entropy and Solvation Entropy in Protein−Ligand Binding

Maria Luisa Verteramo,[†,¶] Olof Stenström,[‡,¶] Majda Misini Ignjatović,[§,¶] Octav Caldararu,[§,¶] Martin A. Olsson,[§,¶] Francesco Manzoni,[∥,¶,▽] Hakon Leffler,[⊥] Esko Oksanen,[#] Derek T. Logan,[∥] Ulf J. Nilsson,[†] Ulf Ryde,[§] and Mikael Akke*[,‡]

[†]Centre for Analysis and Synthesis, Department of Chemistry, Lund University, 221 00 Lund, Sweden
[‡]Biophysical Chemistry, Center for Molecular Protein Science, Department of Chemistry, Lund University, 221 00 Lund, Sweden
[§]Theoretical Chemistry, Department of Chemistry, Lund University, 221 00 Lund, Sweden
[∥]Biochemistry and Structural Biology, Center for Molecular Protein Science, Department of Chemistry, Lund University, 221 00 Lund, Sweden
[⊥]Microbiology, Immunology, and Glycobiology, Department of Laboratory Medicine, Lund University, 221 00 Lund, Sweden
[#]European Spallation Source ESS ERIC, 225 92 Lund, Sweden

Supporting Information

**ABSTRACT:** Understanding the driving forces underlying molecular recognition is of fundamental importance in chemistry and biology. The challenge is to unravel the binding thermodynamics into separate contributions and to interpret these in molecular terms. Entropic contributions to the free energy of binding are particularly difficult to assess in this regard. Here we pinpoint the molecular determinants underlying differences in ligand affinity to the carbohydrate recognition domain of galectin-3, using a combination of isothermal titration calorimetry, X-ray crystallography, NMR relaxation, and molecular dynamics simulations followed by conformational entropy and grid inhomogeneous solvation theory (GIST) analyses. Using a pair of diastereomeric ligands that have essentially identical chemical potential in the unbound state, we reduced the problem of dissecting the thermodynamics to a comparison of the two protein−ligand complexes. While the free energies of binding are nearly equal for the R and S diastereomers, greater differences are observed for the enthalpy and entropy, which consequently exhibit compensatory behavior, $\Delta\Delta H^{\circ}(R - S) = -5 \pm 1$ kJ/mol and $-T\Delta\Delta S^{\circ}(R - S) = 3 \pm 1$ kJ/mol. NMR relaxation experiments and molecular dynamics simulations indicate that the protein in complex with the S-stereoisomer has greater conformational entropy than in the R-complex. GIST calculations reveal additional, but smaller, contributions from solvation entropy, again in favor of the S-complex. Thus, conformational entropy apparently dominates over solvation entropy in dictating the difference in the overall entropy of binding. This case highlights an interplay between conformational entropy and solvation entropy, pointing to both opportunities and challenges in drug design.



## INTRODUCTION

Molecular recognition is fundamental to biology in that it governs signaling within and between cells, with prominent examples provided by the immune system, hormonal control of distant organs in higher organisms, and specificity of enzyme reactions. Modern medicine is to a large extent based on the possibility to interfere with and control molecular recognition by the design of synthetic ligands or effectors that bind to a specific protein in a given signaling pathway. Drug design aims to generate such protein ligands that have high affinity and specificity for the target. Despite the enormous resources contributed by industry and academia over the past several decades, rational structure-based design of ligands by computational approaches remains extremely challenging. One reason is that the free energy of binding is in most cases a small difference between large numbers arising from the different interactions between the protein, ligand, other solutes, and solvent molecules. In addition, the energy terms are strongly dependent on the detailed molecular conformations, due to their sharp dependence on interatomic distances and orientations. Furthermore, entropic contributions can be significant because proteins have many degrees of freedom, are generally flexible, and consequently populate a wide range of conformations. Recent work has indeed highlighted the role of protein conformational entropy in ligand binding,[1−8] as well as the highly heterogeneous response of water molecules around binding sites[9−12] and ligands.[13]

Received: October 15, 2018
Published: January 8, 2019

  
DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

We have identified the carbohydrate recognition domain (CRD) of galectin-3 (denoted galectin-3C) as an interesting system for investigating the role of conformational entropy[4,5] and solvation in ligand binding.[14] Galectin-3 has a relatively solvent-accessible binding site placed in a shallow groove across one of the two $\beta$-sheets, with water molecules forming an integral part of the binding site by bridging between the ligand and protein.[14,15] Galectin-3 is a member of the galectin family of mammalian lectins, defined by the CRD with its conserved sequence motif that confers affinity for $\beta$-galactoside containing glycans.[16,17] Galectins play important roles in cell growth, cell differentiation, cell cycle regulation, signaling, and apoptosis, which target them for pharmaceutical intervention to treat inflammation and cancer,[16−19] with specific examples reported for galectin-3.[20−22]

Here, we report a comparative analysis of galectin-3C in complex with two diastereomeric ligands. The advantage of this approach is that the differences in binding thermodynamics are dominated by the properties of the two ligand−protein complexes, while the unbound diastereomers have nearly identical chemical potential in the unbound state and thus cancel in the comparative analysis. We used a combination of experimental and computational approaches including isothermal titration calorimetry (ITC), competitive fluorescence polarization assay, X-ray crystallography, NMR spectroscopy including $^{15}$N backbone and $^{2}$H side-chain methyl relaxation, and molecular dynamics (MD) simulations followed by conformational entropy and grid inhomogeneous solvation theory (GIST) calculations. Following on our previous work,[4,5] we focus on entropic contributions to the free energy of binding. In the present work, we extend the analysis to include not only conformational entropy of the protein and ligand but also solvent entropy. Our results show that conformational entropy makes a greater contribution than solvent entropy to the difference between ligands in overall entropy of binding, and further highlight an interplay between conformational entropy and solvent entropy in contributing toward ligand binding affinity and specificity.

## MATERIALS AND METHODS

**Ligand Synthesis.** The two diastereomeric compounds (2*R*)- and (2*S*)-2-hydroxy-3-(4-(3-fluorophenyl)-1*H*-1,2,3-triazol-1-yl)-propyl) 2,4,6-tri-*O*-acetyl-3-deoxy-3-(4-(3-fluorophenyl)-1*H*-1,2,3-triazol-1-yl)-1-thio-$\beta$-D-galactopyranoside (denoted ligands R and S, respectively) were synthesized from triisoprolylsilyl 2,4,6-tri-*O*-acetyl-3-azido-3-deoxy-1-thio-$\beta$-D-galactopyranoside[23] and *R*- and *S*-glycidyl nosylate. Reaction conditions, physical data, and purity data are given in the Supporting Information.

**Protein Expression and Purification.** Galectin-3C was expressed and purified by the Lund Protein Production Platform (LP3) at Lund University following published protocols,[4,5] yielding a protein stock solution of 9.2 mg/mL in ME-PBS buffer (10 mM $Na_2HPO_4$, 1.8 mM $KH_2PO_4$, 140 mM NaCl, 2.7 mM KCl, pH 7.3, 2 mM ethylenediaminetetraacetic acid (EDTA), 4 mM $\beta$-mercaptoethanol), and 150 mM lactose. The protein stock solution was stored at 278 K.

**Isothermal Titration Calorimetry.** Galectin-3C samples were prepared by extensive dialysis against 5 mM 4-(2-hydroxyethyl)-1-piperazinethanesulfonic acid (HEPES) buffer to remove all lactose, followed by centrifugation at 14 000 rpm to remove any aggregates. Both ligands were dissolved in stock solutions of dimethyl sulfoxide (DMSO) to prepare stock solutions 20.7 mM and 20.3 mM for R and S, respectively, and sonicated immediately prior to experiments. Isothermal titration calorimetry (ITC) experiments were performed on MicroCal iTC200 and MicroCal PEAQ−ITC instruments (Malvern) at a temperature of 301 K by titrating the protein at a concentration of 0.22 mM into the cell containing the ligand at a concentration of 0.02 mM. The DMSO concentrations in the cell and the syringe were carefully matched to minimize the heat of dilution, and were the same for the two ligands. Five replicate experiments were performed for each complex. Peak integration was done using NITPIC.[24] A single-site binding model was fitted simultaneously to the 5 titrations curves to yield the binding enthalpy ($\Delta H$), fraction of binding-competent protein ($n$), and dissociation constant ($K_d$), using in-house MATLAB routines with Monte Carlo error estimation.[25] The heat released or absorbed during the $i$th injection is given by[26]

$$\Delta Q(i) = Q(i) - Q(i-1) + (V_i/V_0)[Q(i) - Q(i-1)]/2 + Q_{\mathrm{off}}$$

where $V_i$ is the volume of the $i$th injection, $V_0$ is the cell volume, $Q_{\mathrm{off}}$ is an offset parameter that accounts for heats of mixing, and $Q(i)$ is the heat function following the $i$th injection:

$$Q(i) = (\Delta H V_0/2)[\alpha - \sqrt{\alpha^2 - 4nM_tX_i}]$$

where $\alpha = nM_i + X_i + K_d$, and $M_i$ and $X_i$ are the total concentrations of protein and ligand, respectively, in the cell at any given point of the titration. The free energy and entropy of binding were subsequently determined using the relationships $\Delta G^\circ = RT\ln(K_d)$ and $-T\Delta S^\circ = \Delta G^\circ - \Delta H^\circ$.

**Competitive Fluorescence Polarization Experiments.** The binding affinity between galectin-3C and each ligand was determined using competitive fluorescence polarization experiments described previously,[22] using the fluorescent probe 3,3′-dideoxy-3-[4-(fluorescein-5-yl-carbonylaminomethyl)-1*H*-1,2,3-triazol-1-yl]-3′-(3,5-dimethoxybenzamido)- 1,1′-sulfanediyl-di-$\beta$-D-galactopyranoside.[27]

**X-ray Crystallography.** Crystals of lactose-bound galectin-3C were grown with the hanging drop method in NeXtal plates and with the following reservoir condition: 28% (w/v) PEG 4000, Tris-HCl pH 7.5, 0.4 M NaSCN, 15 mM $\beta$-mercaptoethanol. The drop volume was 2 $\mu$L and the protein solution:reservoir ratio was varied between 0.5:1, 1:1, and 2:1. The crystals were then moved to drops containing the same reservoir with the addition of 10 mM of the ligand (R or S), made from a 100 mM stock solution in neat DMSO. Soaking lasted for 7 h for the R diastereomer and 20 h for S. Before data collection, crystals were placed for a couple of seconds in a drop containing 1 volume of 100% PEG400 and 3 volumes of crystallization solution as a cryoprotectant, before cryocooling to 100 K in a stream of gaseous $N_2$. Data were collected at 100 K at beamline I911−3 of the MAX-II synchrotron, Lund, Sweden.[28] All data were integrated using XDS.[29] Diffraction data for R were collected in a single pass, while that for S involved two passes, one at low resolution with lower exposure time followed by one at high resolution, and subsequently scaled and merged with XSCALE.[29]

MTZ files were generated with Aimless.[30] Cross-validation during refinement was based on 10% of the reflections. An initial structure solution was determined through rigid-body refinement in Refmac5[31] using as a starting model the lactose−galectin-3C structure[14] with lactose and water molecules removed and with the resolution limit set to 3.5 Å. The structures of the R and S ligand stereoisomers were built manually using Chimera[32] and geometric restraints for the ligands were obtained through phenix.eLBOW.[33] Restrained refinement was then performed using phenix.refine[34] using data to the diffraction limit. Manual rebuilding, including addition of water molecules, was done using Coot.[35]

**Ensemble Refinement of Crystal Structures.** Ensemble refinement of the X-ray diffraction data was performed using the module phenix.ensemble_refinement in the Phenix software suite.[36] The X-ray crystal structures of the *S*-galectin-3C and *R*-galectin-3C complexes from the previous section were used as starting structures. The crystallographic water molecules were kept and hydrogen atoms and missing atoms in the protein were added using the Leap module from the Amber 14 software.[52] Ligand restraints and coordinates were the same as those used in the original refinement.


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

The collective dynamics of the protein were described using a TLS model with a single group, which included both the protein and the ligand atoms. A model including two TLS groups was also tested—one for the ligand and one for the protein—but it gave worse results ($R_{free}$ values of 0.20 compared to 0.17 for the single TLS model). The percentage of atoms included in the TLS-fitting ($p_{TLS}$) was optimized by testing five different values (0.5, 0.6, 0.7, 0.8 and 0.9) and choosing the one that yielded the lowest $R_{free}$, which was $p_{TLS}$ = 0.7 for both protein−ligand complexes. An ensemble of structures was then generated by running MD simulations, in which the model was restrained by a time-averaged X-ray maximum-likelihood target function. The X-ray weight-coupled temperature bath offset was kept at the default value of 5 K. A 1.25 ps relaxation time of the time-averaged-restraints was used, resulting in 25 ps long MD simulations, with structures stored every 0.05 ps. All structures generated by ensemble refinement were kept, resulting in 500 different structures in each ensemble. Atomic fluctuations were calculated using the cpptraj module of Amber after removal of the water molecules.[37]

**NMR Sample Preparation.** The galectin-3C concentration was 0.32, 0.2, and 0.34 mM for the $^{15}N$, $^{15}N/^{13}C$, and $^{15}N/^{13}C/^{2}H$ samples, respectively. The ligands were dissolved in neat DMSO to a concentration of 8.2 mM for S and 35 mM for R. The protein−ligand complexes were prepared by titrating the ligand into the protein, while monitoring the $^{15}N$ heteronuclear single-quantum correlation (HSQC) spectra. The final DMSO content in the NMR sample was 4.3% for S and 1.2% for R.

**NMR Resonance Assignments and Chemical Shift Analysis.** Backbone chemical shift assignments were based on HNCACB[38] spectra and previous assignments for various galectin-3C complexes.[5] Methyl groups were assigned using CCH-TOCSY and HCCH-TOCSY experiments.[39,40] All spectra were processed using NMRPipe,[41] employing a processing protocol including a solvent filter, square cosine apodization, and zero filling to twice the number of points in all dimensions. All spectra were analyzed using the CCPNmr program suite.[42] Chemical shift differences were evaluated as weighted distances: $([\Delta\delta(^{1}H)]^2 + [0.1\Delta\delta(^{15}N)]^2)^{1/2}$ for backbone amides and $([\Delta\delta(^{1}H)]^2 + [0.25\Delta\delta(^{13}C)]^2)^{1/2}$ for methyls.

**NMR Relaxation Experiments and Data Analysis.** $^{15}N$ $R_1$, $R_2$, and $\{^{1}H\}-^{15}N$ nuclear Overhauser effect (NOE) experiments targeting the backbone amides were performed at magnetic field strengths of 11.7, 14.1, and 21.1 T, and a temperature of 301 K. Spectral widths were 14−16 ppm and 28−30 ppm for $^{1}H$ and $^{15}N$, respectively, covered by 1024 and 128 points. Relaxation decays were recorded with 10 relaxation delays ranging between 0−1 s for $R_1$ acquired at 11.7 and 14.1 T, 0−3 s for $R_1$ acquired at 21.1 T, and 0−0.2 s for $R_2$ (at all fields) with a 1.2 ms delay between refocusing pulses. The NOE was measured using a $^{1}H$ saturation time of 7 s and a recycle delay between experiments of 3 and 7 s for experiments acquired at 11.7 and 14.1 T, respectively, while the reference experiment was acquired using a recycle delay of 10 and 14 s at 11.7 and 14.1 T, respectively. NOE experiments performed at 21.1 T employed a $^{1}H$ saturation time of 6 s and a recycle delay between experiments of 2 s, while the reference experiment was acquired with a recycle delay of 14 s. Peak intensities were evaluated as partial peak volumes calculated over 3 × 5 points in the direct and indirect dimension, respectively. Monoexponential functions were fitted to the $R_1$ and $R_2$ relaxation decays using the CCPNmr program suite and bootstrap error estimation. NOEs were calculated as the ratio of the peak intensities in the saturated and reference experiments, and the standard errors were determined by propagating the errors of intensities estimated from the baseplane noise.

$R_1(D_Z)$, $R(3D_Z^2 - 2)$, $R_2(D_+)$, and $R(D_+D_Z + D_ZD_+)$ $^{2}H$ relaxation experiments[43] targeting the methyl groups were recorded at 11.7 and 14.1 T. Spectral widths were 16 and 20 ppm for $^{1}H$ and $^{13}C$, respectively, covered by 1024 points in the $^{1}H$ dimension at both field strengths, and 70 and 84 points for $^{13}C$ at 11.7 and 14.1 T, respectively. The number of points recorded were 1024 for $^{1}H$ at both static magnetic field strengths. Relaxation decays were sampled by 9 points covering 0−0.1 s for $R_1(D_Z)$ and $R(3D_Z^2 - 2)$, 0−20 ms for $R_2(D_+)$ and $R(D_+D_Z + D_ZD_+)$. The recycle delay was 1.8−2 s. Peak volumes were evaluated using the program suite PINT.[44] Mono-exponential functions were fitted to the relaxation decays using an in-house MATLAB script with Monte Carlo error analysis.[25]

$^{15}N$ CPMG relaxation dispersion experiments were performed at 301 K and static magnetic field strengths of 11.7 and 14.1 T on *S*-galectin-3C and 14.1 T on *R*-galectin-3C, using a single experimental data point per refocusing frequency.[45,46] A series of 19 relaxation dispersion spectra were acquired with CPMG refocusing frequencies ranging from 50 to 800 Hz, and in addition a single reference spectrum was recorded without any CPMG refocusing pulses. The relaxation dispersion data were analyzed using the general equation for two-state exchange.[47−49]

**Model-Free Analysis of NMR Relaxation Data.** Backbone amide model-free parameters were fitted using the program suite relax,[50−52] using a N−H bond length of 1.02 Å and a $^{15}N$ chemical shift anisotropy of −172 ppm. The backbone optimization was restricted to five different models defined by the parameter sets: $\{O^2\}$, $\{O^2, \tau_e\}$, $\{O^2, R_{ex}\}$, $\{O^2, \tau_e, R_{ex}\}$, or $\{O^2_f, O^2_s, \tau_s\}$, where $O^2$, $O^2_f$ and $O^2_s$ denote the order parameter with subscripts *f* and *s* indicating that the order parameter can be resolved into amplitudes of fluctuation taking place on separate time scales (fast and slow), $\tau_e$ and $\tau_s$ denote effective correlation times for the internal motion with subscript *s* indicating that the correlation time is associated with the slower time scale, and $R_{ex}$ denotes exchange contributions to $R_2$; in addition, the correlation time for overall rotational diffusion, $\tau_c$, was also fitted.[53] Side-chain methyl-axis model-free optimization was performed using in-house routines implemented in MATLAB. The $^{2}H$ quadrupolar coupling constant was set to 167 kHz.[54] Three different models were fitted using two $\{O^2, \tau_f\}$, three $\{O^2, \tau_f, \tau_{eff}\}$, or four $\{O^2_f, O^2_s, \tau_f, \tau_{eff}\}$ parameters, where $\tau_f$ is associated with fast motions, $\tau_{eff} = (1/\tau_c + 1/\tau_s)^{-1}$, and $\tau_s$ denotes the correlation time for slow internal motions on par with $\tau_c$.[55] The global correlation time was fixed to the value obtained from the backbone model-free optimization. Model selection was performed using an *F*-test at the level $\alpha = 0.95$ ($p < 0.05$).[56]

**Conformational Entropy Estimates from Order Parameters.** The backbone conformational entropy change, going from state A to B, was estimated from the NMR order parameters using the relationship[1,57]

$$\Delta S_{AB} = R \sum_k \ln[(1 - O^2_{B,k})/(1 - O^2_{A,k})] \quad (1)$$

where $O_{X,k}$ is the order parameter for residue *k* in state *X*, and the sum runs over all residues. In a similar way, the conformational entropy change of the side chain methyl-axis was determined using[57]

$$\Delta S_{AB} = R \sum_m C_m \sum_n (O^2_{B,n} - O^2_{A,n}) \quad (2)$$

where $C_m$ is a function of the residue type. The sums run over all residues *n* of type *m*. $C_m$ = 1.32 for Val and Thr, 3.1 for Ile and Leu, and 2.31 for Met. The entropy for Ala side chains was calculated using eq 1.

The entropy estimated from eqs 1−2 rests on a number of assumptions that have been discussed in the literature.[1,58,59] Most importantly, the approach does not account for contributions from conformational fluctuations with correlation times greater than $\tau_c$, and does not consider the effects of correlated motion. An alternative approach, based on empirical calibration, has been proposed recently.[60] Here, the total conformational entropy is estimated from the average methyl-axis order parameters:

$$\Delta S_{AB} = s_d N_\chi \Delta\langle O^2\rangle_{AB} \quad (3)$$

where $s_d = -(4.8 \pm 0.5) \times 10^{-3}$ kJ/mol/K is an empirically determined constant,[60] $N_\chi$ denotes the number of dihedral angles, and $\Delta\langle O^2\rangle_{AB}$ is the difference between states A and B in their average methyl-axis order parameter. This empirically calibrated estimate of conformational entropy is believed to capture also the effects of correlated motion and motions occurring on time scales greater than $\tau_c$.[60]


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

**Molecular Dynamics Simulations and Analysis.** All MD simulations were run with the Amber 14 software suite.[61] The X-ray crystal structures of the *S*-galectin-3C and *R*-galectin-3C complexes were used as the starting points for MD simulations. The PDB structure 3ZSL was used for the simulations of apo galectin-3C. Separate simulations were run for the two different conformations observed for ligand S. All crystal-water molecules were kept in the simulations. Each galectin-3C complex was solvated in an octahedral box of water molecules extending at least 10 Å from the protein using the tleap module, so that 4965−5593 water molecules were included in the simulations. The simulations were set up in the same way as in our previous studies of galectin-3C.[4,62,63] All Glu and Asp residues were assumed to be negatively charged and all Lys and Arg residues positively charged, whereas the other residues were neutral. The active-site residue His158 was protonated on the ND1 atom, whereas the other three His residues were protonated on the NE2 atom, in accordance with the neutron structure of the lactose-bound state,[15] NMR measurements, and previous extensive test calculations with MD.[64] This resulted in a net charge of +4 for the protein. No counterions were used in the simulations.

The protein was described by the Amber ff14SB force field,[65] water molecules with the TIP4P-Ewald model,[66] whereas the ligands were treated with the general Amber force field.[67] Charges for the ligands were obtained with the restrained electrostatic potential method.[68] The ligands were optimized with the semiempirical AM1 method, followed by a single-point calculation at the Hartree−Fock/6-31G* level to obtain the electrostatic potentials, sampled with the Merz−Kollman scheme.[69] These calculations were performed with the Gaussian 09 software.[70] The potentials were then used by antechamber to calculate the charges. A few missing parameters were obtained with the Seminario approach:[71] The geometry of the ligands was optimized at TPSS/def2-SV(P) level, followed by a frequency calculation using the aoforce module of Turbomole 7.01.[72] From the resulting Hessian matrix, parameters for the missing angles and dihedrals were extracted with the Hess2FF program.[73] These parameters are given in Table S1 in the Supporting Information.

For each complex, 10 000 steps of minimization were used, followed by 20 ps constant-volume equilibration and 20 ps constant-pressure equilibration, all performed with heavy nonwater atoms restrained toward the starting structure with a force constant of 209 kJ/mol/Å$^2$. Finally, the system was equilibrated for 2 ns, followed by 10 ns of production simulation, both performed with constant pressure and without any restraints. For each protein−ligand complex, 10 independent simulations were run, employing different solvation boxes and different starting velocities.[74] Consequently, the total simulation time for each complex was 100 ns. All bonds involving hydrogen atoms were constrained to the equilibrium value using the SHAKE algorithm,[75] allowing for a time step of 2 ps. The temperature was kept constant at 300 K using Langevin dynamics,[76] with a collision frequency of 2 ps$^{-1}$. The pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm[77] with a relaxation time of 1 ps. Long-range electrostatics were handled by particle-mesh Ewald (PME) summation[78] with a fourth-order B spline interpolation and a tolerance of $10^{-5}$. The cutoff radius for Lennard−Jones interactions between atoms of neighboring boxes was set to 8 Å. The snapshots were analyzed with the cpptraj module.[37]

**Conformational Entropy Estimates from MD Simulations.** To validate the MD trajectories by NMR, we calculated order parameters from the MD trajectories. The N−H order parameters were obtained using isotropic reorientational eigenmode dynamic analysis.[79] The covariance matrix of the NH bond vectors was obtained from the trajectories by the cpptraj module[37] in the Amber 14 software.[61]

A total of 10 000 snapshots with a 10 ps sampling frequency were used for entropy and order parameter estimates, employing separate simulations for the complexes, for free galectin-3C and for the solvated ligands. Conformational entropies were calculated from the ensemble of configurations of the protein and ligands by analyzing the dihedral angle fluctuations.[4,63,80,81] The Cartesian coordinates from the trajectories were transformed to internal coordinates and the entropies were then calculated from probability distributions over all possible states of these coordinates using a bin size of $S°$ (i.e., 72 bins per dihedral). Entropies were normalized to that of a free rotor.[4] All entropies are reported as $-T\Delta S$ at 301 K.

Both entropies and order parameters were calculated as averages over 50 simulations of 2 ns each (with 200 snapshots in each, i.e., each of the 10 simulations were divided into five parts of equal length). The 2 ns time window is similar to the rotational correlation time of the protein. This procedure yields more stable entropy estimates by restricting the dependence on rare events.[63] The reported uncertainties are standard errors over these 50 simulations.

To estimate the effect of correlation, entropies were also calculated employing the maximum information spanning tree algorithm[82,83] (MIST), with the pdb2entropy program.[84] Entropies were calculated to the tenth nearest neighbor to account for high-order correlations, whereas entropies calculated to the first nearest neighbor were considered correlation-free.

**Water Structure and Solvation Thermodynamics.** We analyzed the structure and thermodynamics of the solvent around the two ligands (R and S) bound to galectin-3C, using GIST,[85] implemented in the cpptraj module of the Amber 14 software. The method requires snapshots from MD simulations in which the solute is kept restrained. Therefore, we first performed clustering of the trajectories from the unrestrained simulations described above, using the hierarchical agglomerative clustering approach, implemented in the cpptraj module, with average-linkage criteria and the ligand RMS as distance metric. The minimum distance between clusters was set to 3.5 Å. Subsequently we performed 10 independent 10 ns long MD simulations for each identified cluster. In these simulations the protein was kept restrained toward the starting crystal structure, and the ligand was kept restrained toward the conformation which best represents the cluster, both with a force constant of 10 kcal/mol/Å$^2$.

For each cluster, the water−water interaction energy, $E_{w-w}$, and solute−water interaction energy, $E_{s-w}$, as well as translational, $S_{trans}$, and rotational, $S_{rot}$, entropy contributions were calculated for a rectangular grid of dimensions 30 Å × 21 Å × 21 Å, centered on the ligand and extended at least 3 Å on each side of the ligand. The grid was divided into cubic boxes (0.5 Å × 0.5 Å × 0.5 Å), for which the thermodynamic properties were calculated. The sum of these properties over the entire region reveals changes in the hydration thermodynamics of the region for each cluster, relative to the thermodynamics of the bulk water. For each of the two ligands, the solvation free energy, $\Delta G_{solv}$, was calculated as a sum over solvation free energies for each cluster, $\Delta G_{solv}(i)$, multiplied by the probability of finding the ligand in conformation $i$, $p(i)$:

$$\Delta G_{solv} = \sum_i \Delta G_{solv}(i) p(i)$$

A separate set of solute-restrained MD simulations was performed in which both the protein and the ligand were restrained toward the crystal structure. To analyze these simulations, we used a 27 Å × 14 Å × 15 Å grid.

## RESULTS AND DISCUSSION

**Ligand Design and Synthesis.** We investigated the driving forces underlying affinity and selectivity in ligand binding by carrying out a comparative analysis involving the binding of two diastereomeric ligands R and S (Figure 1) to galectin-3C. The design of ligands R and S was inspired by the high-affinity ($K_d$ = 2 nM) galectin-3 ligand 1−1′-sulfanediyl-bis-{3-deoxy-3-[4-(3-fluorophenyl)-1*H*-1,2,3-triazol-1-yl]-*β*-D-galactopyranoside}[22,86] (Figure 1A). The high-affinity ligand interacts with galectin-3 via one of the galactose residues (that on the left-hand side in Figure 1A) in the conserved galactose binding site and the fluorophenyltriazolyl moieties interacts via face-to-face stacking with arginine side chains and one fluorine−amide orthogonal multipolar interaction.[22] The second galactose moiety interacts with only a single hydrogen


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

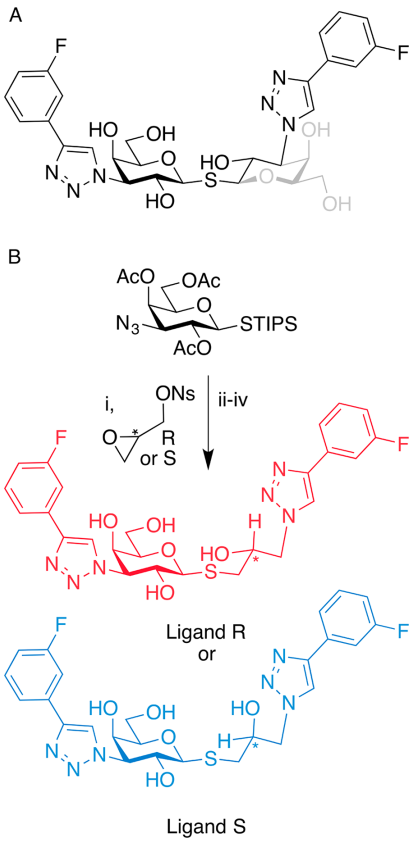



**Figure 1.** Chemical structures and synthesis of ligands. (A) Chemical structure of the parent, high-affinity ligand 1−1′-sulfanediyl-bis-{3-deoxy-3-[4-(3-fluorophenyl)-1*H*-1,2,3-triazol-1-yl]-*β*-D-galactopyranoside}. The noninteracting atoms of one galactopyranose are depicted in gray. (B) Synthesis and structures of the two ligands R (red) and S (blue). The stereocenter is located at the propyl C2 (marked by an asterisk). Reagents and conditions: (i) TBAF·3$H_2O$, dry THF. (ii) $NaN_3$, $NH_4Cl$, dioxane/$H_2O$ 1:1. (iii) 1-Ethynyl-3-fluorobenzene, CuI, $Et_3N$, DMF. (iv) MeONa, MeOH.

bond (Figure 1A; the noninteracting parts are depicted in gray) to the protein, leading us to hypothesize that this galactopyranose ring could be mimicked by a 2-hydroxypropyl chain, which would open up for the synthesis of two diastereomeric ligands R and S (Figure 1B).

Synthesis of the ligands R and S relied on fine-tuning the reactivity between a 1-sulfhydryl-galactopyranose nucleophile and a doubly electrophilic glycidyl derivative: In situ fluoride-mediated activation of the masked nucleophilic triisopropylsilyl thiogalactoside and (*R*)- and (*S*)-glycidyl nosylate, respectively, proceeded stereoselectively in high yields, while other galactose nucleophiles (-SAc, -SH, thiouronium salts, and thioxanthate) and glycidyl electrophiles (glycidyl tosylate, *tert*-butyl dimethyl silyl glycidyl, and epi-chlorohydrin) gave lower yields and stereochemical scrambling due to nucleophilic attack occurring on both C1 and C3 of the glycidyl derivatives, or due to epoxide opening followed by intramolecular substitution to epoxide reclosing. Regioselective ring-opening of the epoxide with $NaN_3$, Cu(I)-catalyzed cycloadditions with 1-ethynyl-3-fluorobenzene, and finally Zemplen transesterification gave ligands R and S in 99+% purities.

**Overall Binding Thermodynamics.** We characterized the thermodynamics of ligand binding using ITC. We carried out five replicate titrations for each of ligands R and S, and analyzed the binding isotherms by performing a combined fit of the replicate data sets (Figure 2; Figure S1). Table 1 lists the resulting binding thermodynamics. Both ligands have dissociation constants in the low micromolar range, $K_d(R) = (1.0 \pm 0.03) \times 10^{-6}$ M and $K_d(S) = (2.1 \pm 0.1) \times 10^{-6}$ M, and the results correlate well with those obtained in competitive fluorescence polarization experiments, $K_d(R) = (0.43 \pm 0.04) \times 10^{-6}$ M and $K_d(S) = (0.67 \pm 0.5) \times 10^{-6}$ M. As reported previously,[5] $K_d$ values determined by ITC are typically found to be higher by a factor of 2−4 than those measured by fluorescence polarization, but the relative affinities are unchanged within errors. The free energies of binding differ

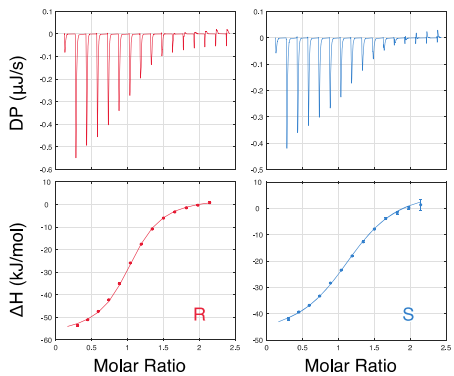



**Figure 2.** ITC experiments of ligand binding to galectin-3C. Example isotherms describing the titration of galectin-3C with ligand R (left-hand side) and ligand S (right-hand side). The top panels show the raw thermograms of differential power plotted versus the ligand:protein molar ratio, while the lower panels show the resulting isotherms. The binding curve results from global fitting of 5 replicate data sets. Error bars are smaller than the size of the symbols, except for the last titration point for ligand S. (Figure S1 shows all 5 isotherms for each ligand).

**Table 1. Overall Binding Thermodynamics from ITC**

| complex | $K_d$ ($10^{-6}$ M) | $\Delta G°_{tot}$ (kJ/mol) | $\Delta H°_{tot}$ (kJ/mol) | $-T\Delta S°_{tot}$ (kJ/mol) |
|---|---|---|---|---|
| *R*-galectin-3C | 1.0 ± 0.03 | −34.6 ± 0.1 | −60.4 ± 0.4 | 25.8 ± 0.4 |
| *S*-galectin-3C | 2.1 ± 0.1 | −32.7 ± 0.1 | −55.7 ± 0.9 | 22.9 ± 0.9 |
| difference (R − S) | | −1.9 ± 0.1 | −5 ± 1 | 3 ± 1 |


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

by only $\Delta\Delta G°(R - S) = -1.9 \pm 0.1$ kJ/mol, but the differences in $\Delta H°$ and $-T\Delta S°$ are greater and consequently opposite in sign, indicating enthalpy−entropy compensation: $\Delta\Delta H°(R - S) = -5 \pm 1$ kJ/mol and $-T\Delta\Delta S°(R - S) = 3 \pm 1$ kJ/mol.

**Crystal Structures Reveal Subtle Differences in Binding Modes.** The crystal structures of the *R*- and *S*-galectin-3C complexes were refined to resolutions of 1.34 and 1.19 Å, respectively (see Table S2 for a summary of refinement statistics). The quality of the electron density data is sufficient to reveal the chirality of the ligands unambiguously (Figure 3 and Figure S2). As shown in Figure 3, the two complexes have closely similar structures, with essentially no difference in the protein backbone conformation. The RMS deviation between the two structures is 0.13 Å for 473 backbone atoms and 0.59 Å when 2054 atoms are compared, including side chains.

Below we will denote the aromatic ring substituents on galactose C3 as the "left hand side" (LHS), while the aromatic rings connected to the propylic chain will be referred as the "right hand side" (RHS); this notation is according to the viewpoint of Figure 3 and all subsequent renditions of the structures. The LHS shows perfect overlap between the two complexes. The 3-fluorophenyl substituent sits in a pocket generated by the displacement of Arg144, with the fluorine atom pointing toward the protein backbone. Key interactions involving the meta-fluorinated phenyl triazole on the LHS have been described previously.[22]

The *B*-factors of the ligand atoms on the LHS are very similar in the two complexes (10−15 Å$^2$), and lower than those of the RHS (20−35 Å$^2$ in R and 20−40 Å$^2$ in S), indicating that the LHS is more ordered. The electron density for Arg144, which stacks with the fluorinated phenyl ring of the LHS, is slightly less well-defined in R than in S. The difference in mobility of Arg144 does not seem to be correlated to the minor differences in water structure (see below).

Although R and S have a different configuration at propyl C2, the conformation of the ligand adjusts to allow the hydroxyl group of the stereocenter to maintain a hydrogen bond with Glu184. The configuration of the R-stereoisomer enables the propyl linker to adopt the same conformation as the corresponding segment in the glucose ring of the parent compound (cf. Figure 1). Thus, the C2 hydroxyl group of R makes an H-bond to Glu184 with its hydrogen atom in a staggered conformation with respect to the aliphatic hydrogen atom on the C2 carbon, as observed in the lactose and glycerol complexes by neutron crystallography.[15] In contrast, the hydroxyl group in S is positioned in an eclipsed conformation with respect to the aliphatic hydrogen, which is expected to be energetically less favorable. This conformational adjustment results in different interactions of the two ligands with the protein at the RHS of the binding site. Furthermore, the RHS of R is modeled with a single conformation, whereas the RHS of S is modeled as two conformations in which the fluorinated ring has two orientations related by an 180° flip. At the RHS, both R and S interact with Arg186, despite the differences in conformation at this end of the ligand. S appears at first glance to have tighter interaction with Arg186 due to a better alignment between the $\pi$ orbitals of the ligand phenyl ring and the face of the arginine guanidinium group. However, the results from ensemble refinement suggest that the S isomer in fact has higher mobility (see below).

Water molecules are well conserved around the binding site. Particularly, we see that waters around the LHS overlap very

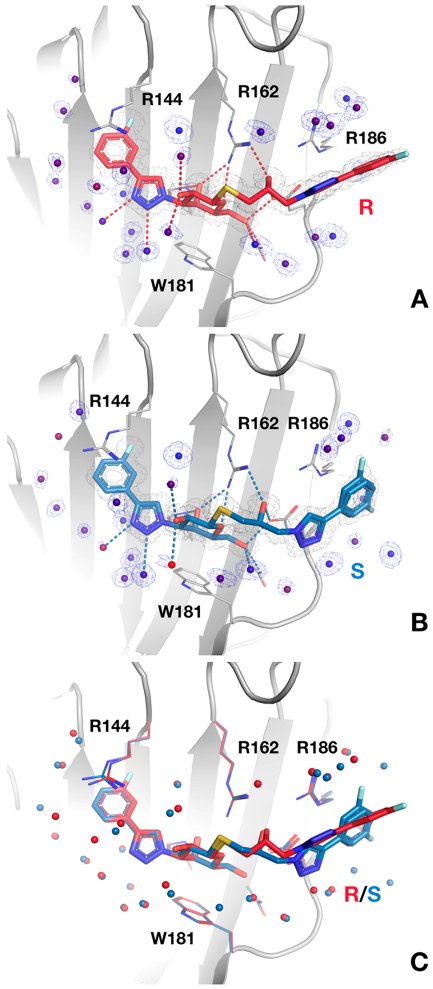



**Figure 3.** X-ray crystal structures of the ligand−galectin-3C complexes. (A) *R*-galectin-3C (PDB ID 6QGF). (B) *S*-galectin-3C (PDB ID 6QGE). (C) Overlay of the two complexes. The protein backbone is shown in ribbon representation (gray), key ligand-coordinating side chains are shown in stick representation, and hydrogen bonds to the ligands are shown as dashed lines. The 2m|$F_o$|−D|$F_c$| electron density map of the ligand and water molecules, contoured at 1.0 $\sigma$, is shown as a gray mesh. Carbon atoms of the R ligand are colored red, while those of the S ligand are blue. Water molecules that are within 5 Å of either ligand are represented as small spheres. In panels A and B, water molecules are colored by *B*-factor, on a spectrum from dark blue at 15 Å$^2$ to bright red at 70 Å$^2$. Water molecules shown without electron density are visible at <1.0 $\sigma$, but are poorly ordered. In panel C, water molecules belonging to *R*-galectin-3C are colored red and those belonging to *S*-galectin-3C are colored blue.


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

well between the two complexes. The minor differences observed could be due to the slightly different resolutions of the two complexes. For the RHS the different conformations of the ligands result in more distinct water structures.

**Chemical Shift Mapping of Ligand Binding.** Chemical shift assignments of *R*-galectin-3C and *S*-galectin-3C were based on a HNCACB experiment and the apo galectin-3C assignments reported previously.[4,5] Minor chemical shift differences are observed for the backbone amides throughout the protein; the RMS chemical shift difference between the ligand-bound and apo forms of galectin-3C in the $^{1}$H and $^{15}$N dimensions are 0.06 and 0.30 ppm for *R*-galectin-3C (Figure S3A) and 0.05 and 0.26 ppm for *S*-galectin-3C (Figure S3B). The methyl chemical shifts show changes similar to those of the backbone, with RMSDs of 0.03 ppm ($^{1}$H) and 0.1 ppm ($^{13}$C) for *S*-galectin-3C, and 0.04 and 0.1 ppm for *R*-galectin-3C. The largest chemical shift changes induced by ligand binding are observed for residues in close proximity to the ligand in the crystal structure (Figure S3C), demonstrating that the binding mode observed in the crystal structure is maintained in solution.

Significant chemical shift differences between the *R*- and *S*-galectin-3C complexes are observed in the binding site (Figure 4). The overall chemical shift RMSD is 0.02 and 0.14 ppm for backbone $^{1}$H and $^{15}$N, respectively, and 0.04 and 0.02 ppm for methyl $^{1}$H and $^{13}$C, respectively. Two methyl groups, Val170$\gamma$1 and Val172$\gamma$1, show a weighted chemical shift difference greater than 0.05 ppm between the two complexes. Furthermore, the $^{1}$H and $^{15}$N chemical shifts of the Arg162 and Arg186 guanidine groups differ between the two complexes. In both cases, the $^{1}$H chemical shift is greater in *R*- than in *S*-galectin-3C, suggesting that the NH$\varepsilon$ atom forms a stronger hydrogen bond or that the population of hydrogen bonded conformations is greater in the R-complex. These four side chains are located closely together and in proximity of the stereocenter of the ligand. Notably, chemical shift differences are also observed in regions of the protein where the average structures are virtually identical between the two complexes, such as the backbone amides of Leu147 and Phe159, which form a pair of NH−CO hydrogen bonds. This observation indicates that subtle differences exist in the conformational ensembles sampled by the two complexes, a topic that we address in more detail below.

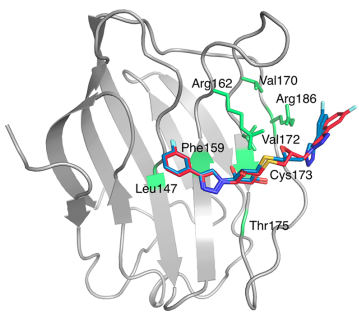



**Figure 4.** Chemical shift differences between the *R*- and *S*-galectin-3C complexes. Residues with weighted chemical shift differences | $\Delta\delta$(R − S)| ≥ 0.05 ppm are highlighted in green on the structure of the *R*-galectin-3C complex with ligand S superimposed. These include the backbone amides of residues Leu147, Phe159, Cys173, Thr175, and Arg186, as well as the methyl groups of Val170 and Val172 and guanidine groups of Arg162 and Arg186, all located in the binding site. Val172 is situated beneath the side chain of Arg162 in the view of the figure.

**Ensemble Refinement of Crystal Structures Highlights Differences in Flexibility.** To investigate the conformational mobility of each complex in the crystal we carried out ensemble refinement of the structure against the X-ray diffraction data.[36] The resulting ensembles yield $R_{free}$ = 0.1709 and $R$ = 0.1358 for *R*-galectin-3C, and $R_{free}$ = 0.1625 and $R$ = 0.1339 for *S*-galectin-3C, values that are comparable to those resulting from traditional refinement (Table S2).

The results indicate that the S diastereomer shows larger fluctuations in the crystal than does R, due to a large variation in the RHS $sp^3$ dihedral angles, as shown in Figure 5. This result agrees with the dual conformation of ligand S observed in the traditionally refined crystal structure, although the conformational variation is much greater in the ensemble representation. In particular, the H-bond between the C2 hydroxyl group and Glu184 is broken in a much larger proportion of the ensemble structures for S than for R. The ensemble refinement also confirms that the R ligand stays in a single conformation, although with some translational movement of the RHS end.

The protein also exhibits variable flexibility. The side chains of Asn160, Arg162, Glu165, and Arg186, which form hydrogen bonds with both ligands, have well-defined positions, whereas larger fluctuations are observed for Arg144, which interacts through $\pi-\pi$ stacking with the ligand phenyl ring at the LHS, and Arg168, which does not interact with the ligand. Arg144 shows slightly greater amplitudes of motion in the R-complex, in keeping with the difference in *B*-factors of the traditionally refined structures. The great variability in the side-chain orientation of Arg144 is also reflected by the NMR data (see below). On the other hand, the ensemble-refined crystal structure of *S*-galectin-3C shows higher fluctuations of several parts of the protein, e.g., the Asn164−Arg168 loop region (neighboring the RHS of the bound ligand), Lys138−Ala142, and Pro113−Val118 (Figure 5, insets).

The resulting ensembles indicate that the *S*-galectin-3C complex shows considerably higher mobility than does *R*-galectin-3C, providing qualitative evidence that protein and ligand conformational entropy is greater in *S*-galectin-3C. We attempted to quantitate the entropy difference from the ensembles, resulting in calculated values that were qualitatively consistent with our other results; however, the estimated standard errors were far greater than the difference between the R- and S-complexes (data not shown).

**Differences in Conformational Fluctuations Measured by NMR.** We carried out a suite of NMR relaxation experiments that probe conformational dynamics on the picosecond to nanosecond time scale to yield the amplitudes of conformational fluctuations in terms of order parameters, denoted $O^2$. We measured $^{15}$N backbone relaxation rates at three static magnetic field strengths and methyl $^{2}$H relaxation rates at two static magnetic field strengths. Out of 138 residues, $^{15}$N relaxation data could be measured for 101 and 100 backbone amides in *R*-galectin-3C and *S*-galectin-3C, respectively. Likewise, out of a total of 85 methyl groups, $^{2}$H relaxation rates could be measured for 65 and 47 methyl groups in *R*-galectin-3C and *S*-galectin-3C, respectively. The


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

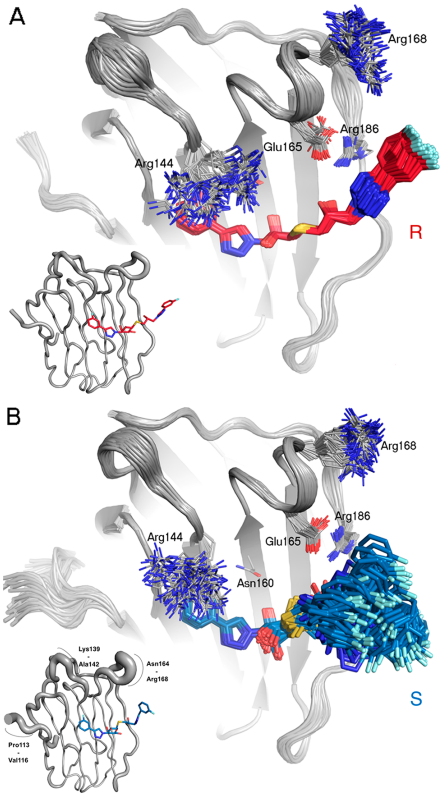



**Figure 5.** Ensemble refined X-ray crystal structures. Overlay of the 100 structures with the lowest $R_{free}$ generated by ensemble refinement for (A) *R*-galectin-3C (red ligand) and (B) *S*-galectin-3C (blue ligand). Insets: The protein backbone is displayed as a tube with a diameter corresponding to the ensemble RMS fluctuations for all atoms of that residue (the ligand is shown in its crystal structure conformation).

missing residues had cross-peaks that were overlapped or too broadened to allow for quantitative analysis.

We characterized the amplitudes of conformational fluctuations using the model-free formalism.[87,88] The best-fit rotational diffusion tensor is anisotropic with a correlation time ($\tau_c$) of 7.5 and 8.1 ns, anisotropy of 1.1 and 1.1, and rhombicity of 0.9 and 1.2 for *R*-galectin-3C and *S*-galectin-3C, respectively. The higher value of $\tau_c$ observed for *S*-galectin-3C is fully explained by the slightly higher concentration of DMSO in this sample, which increases the solvent viscosity.[89,90]

The backbone order parameters are very similar in the two complexes; the mean values and standard deviations are $\langle O^2 \rangle$ = 0.85 ± 0.05 and 0.84 ± 0.05 for *R*-galectin-3C and *S*-galectin-3C, respectively. A significant difference in $O^2$ is observed for residues Tyr118, Ile132, Ile171, Asp178, Arg183, and Leu242, none of which is located directly in the binding site (Figure 6).

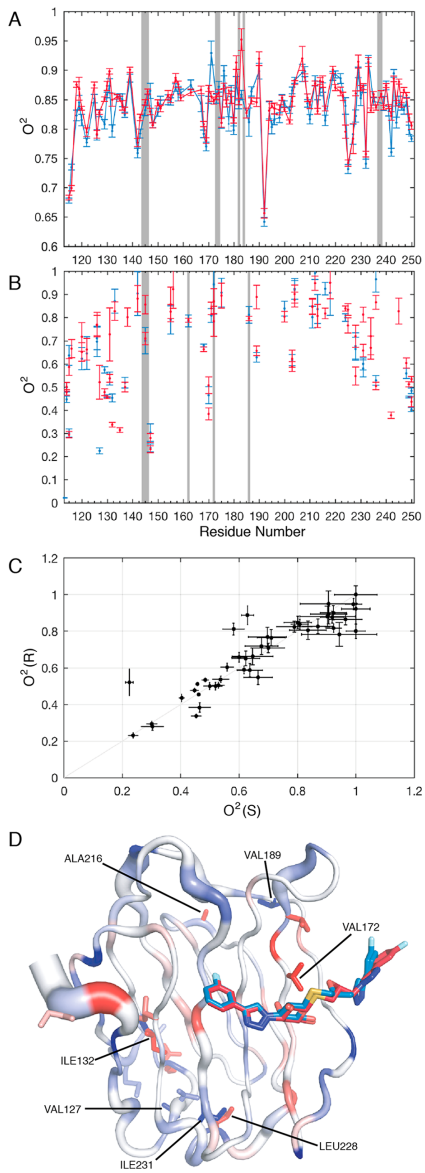



**Figure 6.** NMR order parameters for *R*- and *S*-galectin-3C. (A) Backbone $O^2$ values. (B) Side chain $O^2$ values for arginine $^{15}N\varepsilon$ and methyl axes. Data for *R*- and *S*-galectin-3C are shown in red and blue, respectively. Gray bars indicate residues in contact with the ligand (residues for which any backbone amide atom or methyl atom is within 5 Å of any ligand atom). (C) Scatter plot comparing the



DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

**Figure** 6. continued

methyl-axis $O^2$ values for *R*- and *S*-galectin-3C presented in panel B. The straight line with slope of 1 is drawn to guide the eye. (D) $\Delta O^2$ color coded onto the *R*-galectin-3C structure with ligand S superimposed. Residues with $\Delta O^2(R - S) > 0$ are colored blue, while those with $\Delta O^2(R - S) < 0$ are colored red. The intensity of the color scales with the magnitude of $\Delta O^2$ from red via pink to white ($-0.1 \leq \Delta O^2(R - S) < 0$) and from white via light blue to dark blue ($0 < \Delta O^2(R - S) \leq 0.1$). Residues for which no data are available are colored white. Side chains are shown in stick representation for residues with a difference in side-chain order parameters of $|\Delta O^2(R - S)| > 0.05$, and labeled residues have $|\Delta O^2(R - S)| > 0.1$. Backbone and side-chain $\Delta O^2$ are represented by the color of the tube and sticks, respectively. The width of the tube indicates the average backbone $O^2$ values in the two complexes: a wider tube indicates a lower order parameter and vice versa.

This result indicates that the different stereochemistry of the ligand and the associated differences in protein conformation affect the amplitudes of backbone fluctuations at remote locations. Backbone order parameters are relatively low in the loop regions at the top of the structure in the view of Figure 6. There is also a difference in order between the two complexes with $O^2$ being higher for the R-complex. Both of these observations agree well with the ensemble-refined crystal structures.

The order parameters for the methyl-bearing side chains vary significantly over the protein (Figure 6B,C). However, the differences between the two complexes are overall small, except for residues Val127, Ile132, Val172, Val189, Ala216, Leu228, and Ile231, which show $|\Delta O^2(R - S)| > 0.10$; Figure S4A shows the distribution of $\Delta O^2(R - S)$. Out of these residues, only Val172 is located in the binding site, next to the stereocenter of the bound ligand. The side chain of Val172 shows a greater degree of freedom in the *R*-galectin-3C complex. The mean values and standard errors of the mean for the methyl-axis order parameters are $\langle O^2 \rangle = 0.68 \pm 0.02$ and $0.64 \pm 0.03$ for *R*-galectin-3C and *S*-galectin-3C, respectively, when calculated over all residues, and $\langle O^2 \rangle = 0.66 \pm 0.03$ and $0.65 \pm 0.03$, when calculated over those residues for which data are available for both complexes.

Arginine side chains play a special role in ligand coordination by galectin-3C. Arg144, Arg162, and Arg186 form close interactions with the ligand (cf. Figure 3). However, the side-chain guanidine group of Arg144 is not observed in the NMR spectra, presumably as a consequence of intermediate exchange between alternative positions. This result is in agreement with the ensemble-refined crystal structures, in which Arg144 shows extensive flexibility. The fact that Arg168, which also is highly variable in the structure ensembles, is observed in the NMR spectra indicates that this side chain undergoes dynamic averaging on a faster time scale than does Arg144.

$^{15}N\varepsilon$ side-chain order parameters could be measured for 5 out of 9 arginines. Arg162 and Arg186, which interact with the bound ligands, have $O^2$ values (0.78−0.81) that are higher than the average value of the guanidine groups and only slightly lower than the average value of the backbone. However, there is no significant difference in $O^2$ between the R- and S-complexes for these two side chains. Only Arg129 and Arg224 show minor differences between the two complexes, $|\Delta O^2(R - S)|$ of 0.09 and 0.04, respectively (Figure 6B), and both of these residues are located peripherally to the binding site.

Order parameters derived from relaxation measurements report on conformational entropy due to fluctuations with correlation times shorter than $\tau_c$. To investigate whether there are motions occurring on slower time scales, we performed $^{15}N$ CPMG relaxation dispersion experiments, which sample motions on the 100 $\mu$s to 100 ms time scales.[45] In both the R- and S-bound states, a single residue, Val189, exhibits conformational exchange. The exchange rate is identical, within error: $k_{ex} = 6300 \pm 1300\ s^{-1}$ (R) and $4900 \pm 300\ s^{-1}$ (S), indicating that there are no major differences between the two complexes in the extent of conformational sampling on this time scale.

**Differences in Conformational Fluctuations Determined by MD Simulations.** To complement the information on conformational fluctuations obtained via NMR order parameters for the backbone, guanidine- and methyl-bearing side chains, we performed MD simulations that probe the intramolecular dynamics of all parts of the protein and ligand. Since the crystal structures of the ligand−galectin-3C complexes show two conformations of S, we initiated separate MD simulations for the two conformers. We validated the MD simulations by comparing order parameters calculated from the MD trajectories with those measured by NMR. There is reasonable, but variable, residue by residue agreement between the backbone $O^2$ values determined by NMR and MD for each complex. The RMSD is 0.05 in all 3 comparisons (Figure S4B,C), which is on par with previous results for other proteins.[65,91]

We studied how the conformation of the ligand varied in the MD simulations by following the dihedral angle representing the orientation of the RHS phenyl ring. In each of the three trajectories, the ligand samples a unimodal and equally wide (~50°) distribution of the dihedral, indicating that the rotation barrier is high enough that the ligand does not change conformation on the nanosecond time scale.

**Conformational Entropy Differences Estimated by NMR.** On the basis of the experimental order parameters, we estimated the difference in the conformational entropy between the two complexes, see eqs 1−2. Despite the average values of $O^2$ being highly similar for the two complexes, residue-specific differences lead to a significant difference in backbone conformational entropy between galectin-3C in the R- and S-bound states, $-T\Delta\Delta S_{bb}(R - S) = 17 \pm 5$ kJ/mol. By contrast, the corresponding result for the methyl-axis $O^2$ is not statistically significant: $-T\Delta\Delta S_{sc}(R - S) = -5 \pm 6$ kJ/mol. Taken together, the NMR order parameters yield an estimate of $-T\Delta\Delta S_{bb+sc}(R - S) = 12 \pm 8$ kJ/mol, indicating that galectin-3C in the R-bound state has lower conformational entropy than in the S-bound state (Table 2). That is, the conformational entropy difference between the two complexes has the same sign as, but a greater magnitude than the difference in total entropy, $-T\Delta\Delta S°(R - S)$, obtained by ITC, suggesting that the conformational entropy makes a significant contribution to the overall binding thermodynamics. It should be noted that the NMR-based estimate, $-T\Delta\Delta S_{bb+sc}(R - S)$, covers only a subset of the dihedral angles in the protein. However, it serves as a useful reference for validating the MD simulations, which provide the total conformational entropy of both galectin-3C and the bound ligand.

We also used the empirically calibrated approach,[60] embodied in eq 3, to estimate the change in the total conformational entropy of the protein. The results yield $-T\Delta\Delta S_{conf}(R - S) = 16 \pm 14$ kJ/mol, suggesting, as might be



DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

**Table 2. Conformational Entropy Differences between *R*- and *S*-Galectin-3C**

| method | $-T\Delta\Delta S$ (kJ/mol) |
|---|---|
| NMR backbone + methyls[a] | 12 ± 8 |
| MD backbone + methyls[b] | 8 ± 3 |
| NMR protein[c] | 16 ± 14 |
| MD protein[d] | 11 ± 5 |
| MD protein + ligand[e] | 10 ± 5 |

[a]Includes protein dihedrals of the backbone and methyl-bearing side chains, calculated using eqs 1−2. [b]Includes protein dihedrals of the backbone and methyl-bearing side chains. [c]Includes all protein dihedrals, calculated using eq 3. [d]Includes all protein dihedrals. [e]Includes all protein and ligand dihedrals.

expected, that $-T\Delta\Delta S_{bb+sc}$ underestimates the change in total conformational entropy (Table 2).

**Conformational Entropy Differences Determined by MD.** We calculated the conformational entropy of galectin-3C and the bound ligands in both complexes, using dihedral angle distributions from the MD simulations.[4,63] Table 2 shows the difference in conformational entropy between the two complexes. For both complexes, the dihedral flexibility of galectin-3C decreases upon ligand binding (Table S3). The effect yields a change in conformational entropy, $-T\Delta S_{conf}$, of 86 ± 5 kJ/mol and 74−75 ± 5 kJ/mol for the protein in the R- and S-bound states, respectively (separate MD simulations were initiated from the two conformations of S observed in the crystal structure and these resulted in entropies that agree within 1 kJ/mol). Comparing directly with the NMR-based estimate of conformational entropy associated with the backbone and methyl-bearing side chains, $-T\Delta\Delta S_{bb+sc}$(R − S) = 12 ± 8 kJ/mol, the corresponding value obtained by MD is 8 ± 3 kJ/mol (Table 2).

The decrease in entropy is greatest for Arg186 in both complexes ($-T\Delta S_{conf}$ = 8−9 kJ/mol). This residue forms hydrogen bonds with Glu184, which interacts with both ligands R and S, and shows the second largest decrease in entropy when ligand S binds ($-T\Delta S_{conf}$ = 4 kJ/mol), but a rather small decrease upon binding ligand R ($-T\Delta S_{conf}$ = 1 kJ/mol). Arg144 also gives a rather large negative entropy contribution upon binding either ligand ($-T\Delta S_{conf}$ = 3−4 kJ/mol). Ile171 gives a large contribution ($-T\Delta S_{conf}$ = 4 kJ/mol) when binding S, but smaller when binding R (1 kJ/mol). This difference is also observed in the backbone $O^2$ determined by NMR (Figure 6), whereas there is no significant difference between the two complexes in the methyl-axis $O^2$ values for this residue, whose side chain is oriented away from the binding site. However, the NMR data reveal greater flexibility in the R-complex for the side chains of the neighboring residues Val170 and Val172, which are both oriented toward the binding site. Significantly increased conformational entropy is observed for 3−5 of the residues upon ligand binding, with the largest contribution coming from Asp148 ($-T\Delta S_{conf}$ = −1 kJ/mol).

The total conformational entropy of the protein is greater for *S*-galectin-3C than for the R-complex, $-T\Delta\Delta S_{conf}$(R − S) = 11 ± 5 kJ/mol (taking into account both MD trajectories for *S*-galectin-3C), which is statistically significant at the 95% level. This result agrees well with the estimate obtained from NMR methyl-axis order parameters, $-T\Delta\Delta S_{conf}$(R − S) = 16 ± 14 kJ/mol, which implicitly includes the effects of correlated motions and motions on time scales greater than $\tau_c$. Thus, the general agreement supports the conclusion from MIST calculations that effects from correlated motions are minor, and further suggests that slower motions have no major bearing on $\Delta\Delta S_{conf}$ in keeping with the relaxation dispersion data.

The difference between complexes arises from small contributions from many residues (Figure 7). At the level of individual residues, 22−23% show a statistically significant contribution with the same sign as the total difference, whereas 9−11% show the opposite behavior. Among the latter, the largest contributions (−3 kJ/mol) come from Ile171 and Glu184 (Figure 7).

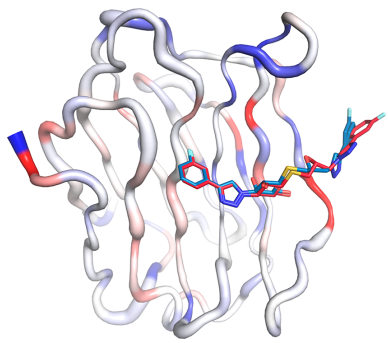


**Figure 7.** Conformational entropy contributions to $-T\Delta\Delta S_{conf}$(R − S), reported per residue. $-T\Delta\Delta S_{conf}$(R − S) is color coded onto the galectin-3C structure with blue hues indicating $-T\Delta\Delta S_{conf}$(R − S) > 0 and red hues indicating $-T\Delta\Delta S_{conf}$(R − S) < 0, with the color intensity ranging from weak (white) for $T\Delta\Delta S_{conf}$ = 0 to intense (maximally blue or red) for $|T\Delta\Delta S_{conf}|$ = 2.8 kJ/mol. The width of the tube indicates the average conformational entropy values per residue in the two complexes: a wider tube indicates higher average conformational entropy and vice versa. The figure is based on the crystal structure of *S*-galectin-3C.

The change in conformational entropy of the ligand upon complex formation is $-T\Delta S$ = 24 ± 1 kJ/mol and 25−26 ± 1 kJ/mol for *R*- and *S*-galectin-3C, respectively. The difference between R and S is not statistically significant, neither in the bound nor in the free states. The indistinguishable conformational entropy of the free ligands is in line with the expectation that they should have nearly identical chemical potential in the free state, based on their diastereomeric relationship.

We used the MIST approach[82,83] to investigate whether correlated motions affect the estimates of conformational entropy. The results show that the effect of correlation on $-T\Delta\Delta S_{conf}$(R − S) is minimal, with 1 kJ/mol difference between the first- (without correlation) and tenth-order (with correlation) approximation. Thus, correlations are highly similar in the two states, in agreement with previous results for other proteins.[92]

Thus, taking into account the results for both ligand and protein, the difference in conformational entropy between the two complexes, $-T\Delta\Delta S_{conf}$(R − S) = 10 ± 5 kJ/mol, is slightly greater than the difference in the net binding entropy, $-T\Delta\Delta S^{\circ}_{tot}$(R − S) = 3 ± 1 kJ/mol, indicating that protein conformational entropy makes a dominant contribution to $\Delta\Delta G^{\circ}_{tot}$(R − S). Note that we have designed this comparative


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

study in such a way that the only other contribution to the entropy of binding should originate from differences in solvation entropy of the two complexes, a topic we turn to next.

**Grid Inhomogeneous Solvation Theory Reveals Key Differences in Solvation between the Two Complexes.** In the standard GIST protocols, sampling of water sites is carried out while keeping the protein and ligand restrained.[85,93] The present case, where the protein and ligand show significant conformational fluctuations in the bound state, presents a challenge to calculations of hydration thermodynamics. We approached the problem by clustering trajectories from the unrestrained MD simulations, which resulted in three clusters for ligand R and four clusters for ligand S (two clusters for each of the two sets of unrestrained MD simulations in the latter case). The subsequent solute-restrained MD simulations, started from each of the clusters for *R*-galectin-3C and *S*-galectin-3C, reveal differences in their hydration thermodynamics (Table 3). Figure 8 affords an overview of water sites,

**Table 3. Solvation Thermodynamics from GIST Calculations**[a]

| complex | *R*-galectin-3C | *S*-galectin-3C | difference (R − S) |
|---|---|---|---|
| $-T\Delta S_{rot}$ | 398.8 ± 0.6 | 397.7 ± 1.3 | 1.2 ± 1.5 |
| $-T\Delta S_{trans}$ | 319.3 ± 0.4 | 317.4 ± 0.3 | 1.9 ± 0.5 |
| $-T\Delta S_{solv}$ | 718.1 ± 0.9 | 715.0 ± 1.4 | 3.1 ± 1.6 |
| $\Delta H_{s-w}$ | −2914.1 ± 2.0 | −2805.4 ± 1.1 | −108.7 ± 2.3 |
| $\Delta H_{w-w}$ | −12813.8 ± 2.0 | −12877.0 ± 2.0 | 63.2 ± 2.9 |
| $\Delta H_{solv}$ | −15727.9 ± 2.3 | −15682.4 ± 2.2 | −45.5 ± 3.2 |
| $\Delta G_{solv}$ | −15009.7 ± 1.8 | −14967.3 ± 2.7 | −42.4 ± 3.3 |

[a]Rotational, $\Delta S_{rot}$, and translational, $\Delta S_{trans}$, entropy as well as the solute−water interaction energy, $\Delta H_{s-w}$, and water−water interaction energy, $\Delta H_{w-w}$, of the studied region, shown relative to bulk water. $\Delta S_{solv} = \Delta S_{rot} + \Delta S_{trans}$, $\Delta H_{solv} = \Delta H_{s-w} + \Delta H_{w-w}$, and $\Delta G_{solv} = \Delta H_{solv} - T\Delta S_{solv}$. All terms are in kJ/mol. Reported uncertainties are the standard errors over the ten independent MD simulations.

i.e., regions with higher density than bulk water, surrounding the bound ligands. Overall, the distributions of highly populated water sites are similar in the two complexes (compare Figures 8A and 8B) and agree well with the crystal structures. However, the close-up view in Figure 8C reveal subtle differences in water positions, especially in the RHS region, where the two structures differ the most. Water molecules in the crystal structures with a low *B* factor overlap well with the highly populated water sites from the GIST analysis, whereas the overlap is poorer for water molecules with a higher *B* factor (Figure S5). Those GIST water densities also have a less spherical shape, indicating a larger mobility of the water structure.

There is a large difference in solvation enthalpy, which is compensated by protein−protein and protein−solvent enthalpies (outside the grid) that are large and hard to estimate accurately, whereas the difference in protein−ligand interaction energies between the R- and S-complexes is modest. Thus, we conclude that the higher binding affinity for the R diastereomer includes a contribution from favorable hydration enthalpy that is dominated by solute−water interactions around the binding site.

Focusing next on solvent entropy, we note that the difference between the two complexes amounts to only $-T\Delta\Delta S_{solv}(R - S) = 3 \pm 2$ kJ/mol. Although barely

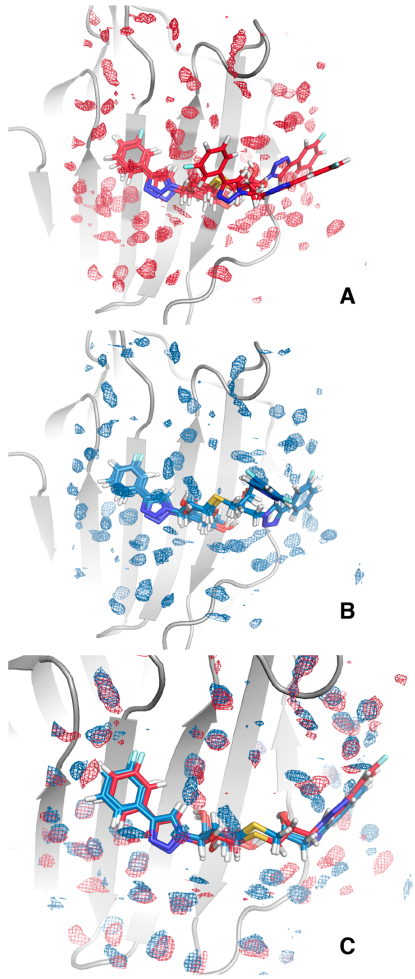



**Figure 8.** Differences in solvation around the binding site. Regions with high density of water relative to bulk water (six times the bulk water density) are represented as (A) red mesh for *R*-galectin-3C and (B) blue mesh for *S*-galectin-3C. (C) Close-up view of the binding site with the R- and S-complexes superimposed. For clarity, only the highest-occupancy clusters are shown for R and S (both conformations in the latter case) in panel C. See the text for details.

significant, the entropic contribution from solvation appears to add constructively to the conformational entropy ($-T\Delta\Delta S_{conf}(R - S) = 10 \pm 5$ kJ/mol). Arguably, this result is intuitive, as greater disorder in the protein and ligand conformations might be expected to translate to the surrounding water molecules. However, the opposite behavior has also been observed in MD simulations of other systems.[11]



DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

The net contribution from conformational and solvent entropy, $-T\Delta\Delta S_{conf+solv}(R - S) = 13 \pm 5$ kJ/mol, is greater than the overall entropy difference determined by ITC, $-T\Delta\Delta S_{tot}(R - S) = 3 \pm 1$ kJ/mol (Figure 9), but the difference is not significant at the 95% confidence level. Taken together, the present results indicate that conformational entropy dominates over solvation entropy in determining the difference in binding entropy between the two ligand−galectin-3C complexes. It remains an open question to what extent these results are general, but we surmise that the relative contributions from conformational entropy and solvent entropy are highly system dependent.

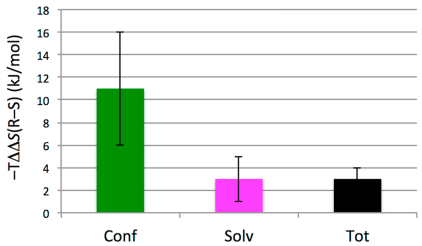



**Figure 9.** Entropy contributions to the differential binding of ligands R and S to galectin-3C. The bars indicate contributions from conformational entropy, $-T\Delta\Delta S_{conf}$ (green); solvation, $-T\Delta\Delta S_{solv}$ (magenta); and total entropy of binding determined by ITC, $-T\Delta\Delta S_{tot}$ (black). Error bars indicate the standard error (one standard deviation).

Galectin-3 has a relatively exposed and solvent-accessible binding site, which engages numerous water molecules, a feature that certainly contributes greatly to the present results. It would be of great interest to carry out future research to investigate other proteins with different types of binding sites, e.g., those that are less solvent accessible.[12]

## CONCLUDING REMARKS

We have carried out a comparative analysis of ligand binding to galectin-3C using two diastereomeric ligands and a range of experimental techniques combined with computational methods. This approach has the important advantage that any differences in the thermodynamics of the two binding processes can be related to the bound state, while the contributions from the free states are expected to cancel—as borne out by the present results. Thus, on the basis of this experimental design, we were able to dissect the thermodynamics underlying the difference in ligand affinity.

The two ligands exhibit closely similar free energies of binding, as might be expected for diastereomers. However, the pair exhibits enthalpy−entropy compensation, so that the two complexes still manifest meaningful differences in both binding enthalpy and entropy that we investigated to pinpoint the driving forces underlying the thermodynamic signatures of binding. Our results demonstrate that the enthalpy−entropy compensation involves interplay between the protein and solvent degrees of freedom. GIST analyses of MD trajectories indicate that the difference in enthalpy includes a sizable contribution from solute−water interactions in favor of the *R*-galectin-3C complex. This contribution is counteracted by a difference in conformational entropy of the protein and a minor entropic component from the solvent that both favor the *S*-galectin-3C complex. Thus, conformational entropy dominates over solvation entropy in determining the difference in binding entropy between the two stereoisomers.

The sum of the conformational and solvation entropies, determined by NMR, MD simulations, and GIST calculations has the same sign as, but is greater than, the total entropy of binding, determined by ITC. Thus, the individual estimates of conformational and solvent entropy correctly identify which protein−ligand complex is favored, but the remaining deviation of $\Delta\Delta S_{conf+solv}$ from $\Delta\Delta S_{tot}$ suggests room for further methodological refinements.

The combination of high-resolution crystal structures, analyzed by ensemble refinement, NMR relaxation data, and MD simulations enable us to examine the structural origin of the thermodynamic differences outlined above. Differences in the interactions involving the hydroxyl group at the stereocenter of the diastereomers apparently lead to conformational strain and more pronounced conformational fluctuations in the S-stereoisomer at the RHS of the binding site, which couple with increased fluctuations of the surrounding protein. These results reinforce the notion that structure-based ligand design, when guided solely by static X-ray structures, addresses only one part of the picture and might be misleading.

In a broader perspective, improved knowledge about the sensitive interdependence of solvent entropy and protein conformational entropy adds to our understanding of molecular recognition. The phenomenon indicates both opportunities and challenges in rational drug design. On the one hand, contributions from solvation entropy to the free energy of binding are well-known, and the present results reiterate the concept of targeting individual water sites to achieve increased binding affinity.[12,94] On the other hand, efforts to design ligands that perturb the solvent structure around the binding site might not achieve the expected result due to changes in conformational entropy of the ligand and protein, as exemplified herein.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.8b11099.

Details of ligand synthesis and purification, including NMR and MS spectra; Tables detailing ligand force field parameters, X-ray crystallography data collection and refinements statistics, conformational entropy differences obtained from MD between the apo state and each of the R- and S-complexes; Figures showing replicate ITC data sets, a close-up view of the electron density at the stereocenter of the bound ligands, chemical shift changes upon ligand binding, and a comparison between GIST water densities and water molecules observed in the crystal structures (PDF)

## AUTHOR INFORMATION

**Corresponding Author**

*mikael.akke@bpc.lu.se

**ORCID**

Hakon Leffler: 0000-0003-4482-8945
Derek T. Logan: 0000-0002-0098-8560
Ulf J. Nilsson: 0000-0001-5815-9522
Ulf Ryde: 0000-0001-7653-8489



DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

Mikael Akke: 0000-0002-2395-825X

### Author Contributions

[¶]MLV, OS, MMI, OC, MAO, and FM contributed equally.

### Notes

The authors declare the following competing financial interest(s): UJN and HL are shareholders in Galecto Biotech AB, a company developing galectin inhibitors.

[▽]Francesco Manzoni passed away on March 12, 2017.

The crystal structures and diffraction data have been deposited in the Protein Data Bank with accession IDs 6QGE and 6QGF. The chemical shift assignments, relaxation rate constants, and order parameters have been deposited in the Biological Magnetic Resonance Bank (BMRB) under accession codes 27721 and 27722.

## ■ ACKNOWLEDGMENTS

We thank Ulrich Weininger and Göran Carlström for assistance with NMR experiments, Barbro Kahl-Knutson for help with FP experiments, Natalia Markova (Malvern) for helpful discussion on ITC measurements and analysis, and Malvern Instruments for access to ITC analysis software. Protein production was carried out by the Lund Protein Production Platform (LP3) at Lund University. High-field (900 MHz) NMR spectroscopy was carried out at the Swedish NMR Center at University of Gothenburg, supported by the Knut and Alice Wallenberg Foundation (NMR for Life). This work was supported by the Knut and Alice Wallenberg Foundation (KAW 2013.022), the Swedish Research Council (project 2014-5540 to UR), and the European Spallation Source (ERIC; UR and DL).

## ■ REFERENCES

(1) Akke, M.; Brüschweiler, R.; Palmer, A. G. NMR Order Parameters and Free Energy: An Analytical Approach and Its Application to Cooperative $Ca^{2+}$ Binding by Calbindin D9k. *J. Am. Chem. Soc.* **1993**, *115*, 9832−9833.

(2) Marlow, M. S.; Dogan, J.; Frederick, K. K.; Valentine, K. G.; Wand, A. J. The Role of Conformational Entropy in Molecular Recognition by Calmodulin. *Nat. Chem. Biol.* **2010**, *6*, 352−358.

(3) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational Entropy in Molecular Recognition by Proteins. *Nature* **2007**, *448*, 325−330.

(4) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational Entropy Changes upon Lactose Binding to the Carbohydrate Recognition Domain of Galectin-3. *J. Biomol. NMR* **2009**, *45*, 157−169.

(5) Diehl, C.; Engström, O.; Delaine, T.; Håkansson, M.; Genheden, S.; Modig, K.; Leffler, H.; Ryde, U.; Nilsson, U. J.; Akke, M. Protein Flexibility and Conformational Entropy in Ligand Design Targeting the Carbohydrate Recognition Domain of Galectin-3. *J. Am. Chem. Soc.* **2010**, *132*, 14577−14589.

(6) Tzeng, S. R.; Kalodimos, C. G. Protein Activity Regulation by Conformational Entropy. *Nature* **2012**, *488*, 236−240.

(7) Gill, M. L.; Byrd, R. A.; Palmer, A. G., III Dynamics of GCN4 Facilitate DNA Interaction: A Model-Free Analysis of an Intrinsically Disordered Region. *Phys. Chem. Chem. Phys.* **2016**, *18*, 5839−5849.

(8) MacRaild, C. A.; Daranas, A. H.; Bronowska, A.; Homans, S. W. Global Changes in Local Protein Dynamics Reduce the Entropic Cost of Carbohydrate Binding in the Arabinose-Binding Protein. *J. Mol. Biol.* **2007**, *368*, 822−832.

(9) Syme, N. R.; Dennis, C.; Bronowska, A.; Paesen, G. C.; Homans, S. W. Comparison of Entropic Contributions to Binding in a "Hydrophilic" versus "Hydrophobic" Ligand-Protein Interaction. *J. Am. Chem. Soc.* **2010**, *132*, 8682−8689.

(10) Haider, K.; Wickstrom, L.; Ramsey, S.; Gilson, M. K.; Kurtzman, T. Enthalpic Breakdown of Water Structure on Protein Active-Site Surfaces. *J. Phys. Chem. B* **2016**, *120*, 8743−8756.

(11) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. Entropy-Enthalpy Transduction Caused by Conformational Shifts Can Obscure the Forces Driving Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 20006−20011.

(12) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808−813.

(13) Wienen-Schmidt, B.; Jonker, H. R. A.; Wulsdorf, T.; Gerber, H. D.; Saxena, K.; Kudlinzki, D.; Sreeramulu, S.; Parigi, G.; Luchinat, C.; Heine, A.; Schwalbe, H.; Klebe, G. Paradoxically, Most Flexible Ligand Binds Most Entropy-Favored: Intriguing Impact of Ligand Flexibility and Solvation on Drug-Kinase Binding. *J. Med. Chem.* **2018**, *61*, 5922−5933.

(14) Saraboji, K.; HÅkansson, M.; Genheden, S.; Diehl, C.; Qvist, J.; Weininger, U.; Nilsson, U. J.; Leffler, H.; Ryde, U.; Akke, M.; Logan, D. T. The Carbohydrate-Binding Site in Galectin-3 Is Preorganized to Recognize a Sugar-like Framework of Oxygens: Ultra-High Resolution Structures and Water Dynamics. *Biochemistry* **2012**, *51*, 296−306.

(15) Manzoni, F.; Wallerstein, J.; Schrader, T. E.; Ostermann, A.; Coates, L.; Akke, M.; Blakeley, M. P.; Oksanen, E.; Logan, D. T. Elucidation of Hydrogen Bonding Patterns in Ligand-Free, Lactose- and Glycerol-Bound Galectin-3C by Neutron Crystallography to Guide Drug Design. *J. Med. Chem.* **2018**, *61*, 4412−4420.

(16) Johannes, L.; Jacob, R.; Leffler, H. Galectins at a glance. *J. Cell Sci.* **2018**, *131*, 1−9.

(17) Dumic, J.; Dabelic, S.; Flögel, M. Galectin-3: An Open-Ended Story. *Biochim. Biophys. Acta, Gen. Subj.* **2006**, *1760*, 616−635.

(18) Liu, F. T.; Rabinovich, G. A. Galectins: Regulators of Acute and Chronic Inflammation. *Ann. N. Y. Acad. Sci.* **2010**, *1183*, 158−182.

(19) Liu, F.-T.; Rabinovich, G. A. Galectins as Modulators of Tumour Progression. *Nat. Rev. Cancer* **2005**, *5*, 29−41.

(20) Di Lella, S.; Sundblad, V.; Cerliani, J. P.; Guardia, C. M.; Estrin, D. A.; Vasta, G. R.; Rabinovich, G. A. When Galectins Recognize Glycans: From Biochemistry to Physiology and Back Again. *Biochemistry* **2011**, *50*, 7842−7857.

(21) Blanchard, H.; Yu, X.; Collins, P. M.; Bum-Erdene, K. Galectin-3 Inhibitors: A Patent Review (2008-Present). *Expert Opin. Ther. Pat.* **2014**, *24*, 1053−1065.

(22) Delaine, T.; Collins, P.; MacKinnon, A.; Sharma, G.; Stegmayr, J.; Rajput, V. K.; Mandal, S.; Cumpstey, I.; Larumbe, A.; Salameh, B. A.; Kahl-Knutsson, B.; van Hattum, H.; van Scherpenzeel, M.; Peiters, R. J.; Sethi, T.; Schambye, H.; Oredsson, S.; Leffler, H.; Blanchard, H.; Nilsson, U. J. Galectin-3-Binding Glycomimetics That Strongly Reduce Bleomycin-Induced Lung Fibrosis and Modulate Intracellular Glycan Recognition. *ChemBioChem* **2016**, *17*, 1759−1770.

(23) Mandal, S.; Nilsson, U. J. Tri-Isopropylsilyl Thioglycosides as Masked Glycosyl Thiol Nucleophiles for the Synthesis of S-Linked Glycosides and Glyco-Conjugates. *Org. Biomol. Chem.* **2014**, *12*, 4816−4819.

(24) Keller, S.; Vargas, C.; Zhao, H.; Piszczek, G.; Brautigam, C. A.; Schuck, P. High-Precision Isothermal Titration Calorimetry with Automated Peak-Shape Analysis. *Anal. Chem.* **2012**, *84*, 5066−5073.

(25) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes. The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1986.

(26) Freiburger, L.; Auclair, K.; Mittermaier, A. Global ITC Fitting Methods in Studies of Protein Allostery. *Methods* **2015**, *76*, 149−161.

(27) Salomonsson, E.; Larumbe, A.; Tejler, J.; Tullberg, E.; Rydberg, H.; Sundin, A.; Khabut, A.; Frejd, T.; Lobsanov, Y. D.; Rini, J. M.; Nilsson, U. J.; Leffler, H. Monovalent Interactions of Galectin-1. *Biochemistry* **2010**, *49*, 9518−9532.

(28) Ursby, T.; Unge, J.; Appio, R.; Logan, D. T.; Fredslund, F.; Svensson, C.; Larsson, K.; Labrador, A.; Thunnissen, M. M. The Macromolecular Crystallography Beamline I911−3 at the MAX IV Laboratory. *J. Synchrotron Radiat.* **2013**, *20*, 648−653.



DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

(29) Kabsch, W. XDS. In *International Tables for Crystallography, Vol. F: Crystallography of Biological Macromolecules*; Rossmann, M. G., Arnold, E., Eds.; Kluwer Academic Publishers: Dordrecht, 2010; Vol. *66*, pp 125−132.
(30) Evans, P. R.; Murshudov, G. N. How Good Are My Data and What Is the Resolution? *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *69*, 1204−1214.
(31) Murshudov, G. N.; Skubak, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A. REFMAC5 for the Refinement of Macromolecular Crystal Structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67*, 355−367.
(32) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(33) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. Electronic Ligand Builder and Optimization Workbench (ELBOW): A Tool for Ligand Coordinate and Restraint Generation. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 1074−1080.
(34) Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards Automated Crystallographic Structure Refinement with Phenix.Refine. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 352−367.
(35) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 486−501.
(36) Burnley, B. T.; Afonine, P. V.; Adams, P. D.; Gros, P. Modelling Dynamics in Protein Crystal Structures by Ensemble Refinement. *eLife* **2012**, *1*, No. e00311.
(37) Roe, D. R.; Cheatham, T. E., III PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Synamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.
(38) Wittekind, M.; Mueller, L. HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. *J. Magn. Reson., Ser. B* **1993**, *101*, 201−205.
(39) Bax, A.; Clore, G. M.; Driscoll, P. C.; Gronenborn, A. M.; Ikura, M.; Kay, L. E. Practical Aspects of Proton-Carbon-Carbon-Proton Three-Dimensional Correlation Spectroscopy of 13C-Labeled Proteins. *J. Magn. Reson.* **1990**, *87*, 620−627.
(40) Bax, A.; Clore, G. M.; Gronenborn, A. M. 1H-1H Correlation via Isotropic Mixing of 13C Magnetization, a New Three-Dimensional Approach for Assigning 1H and 13C Spectra of 13C-Enriched Proteins. *J. Magn. Reson.* **1990**, *88*, 425−431.
(41) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* **1995**, *6*, 277−293.
(42) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, P.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins: Struct., Funct., Genet.* **2005**, *59*, 687−696.
(43) Millet, O.; Muhandiram, D. R.; Skrynnikov, N. R.; Kay, L. E. Deuterium Spin Probes of Side-Chain Dynamics in Proteins. 1. Measurement of Five Relaxation Rates per Deuteron in 13C-Labeled and Fractionally 2H-Enriched Proteins in Solution. *J. Am. Chem. Soc.* **2002**, *124*, 6439−6448.
(44) Ahlner, A.; Carlsson, M.; Jonsson, B. H.; Lundström, P. PINT: A Software for Integration of Peak Volumes and Extraction of Relaxation Rates. *J. Biomol. NMR* **2013**, *56*, 191−202.
(45) Loria, J. P.; Rance, M.; Palmer, A. G. A Relaxation-Compensated Carr-Purcell-Meiboom-Gill Sequence for Characterizing Chemical Exchange by NMR Spectroscopy. *J. Am. Chem. Soc.* **1999**, *121*, 2331−2332.
(46) Mulder, F. A. A.; Skrynnikov, N. R.; Hon, B.; Dahlquist, F. W.; Kay, L. E. Measurement of Slow (Microseconds-Milliseconds) Time Scale Dynamics in Protein Side Chains by 15N Relaxation Dispersion NMR Spectroscopy: Application to Asn and Gln Residues in a Cavity Mutant of T4 Lysozyme. *J. Am. Chem. Soc.* **2001**, *123*, 967−975.
(47) Davis, D. G.; Perlman, M. E.; London, R. E. Direct Measurements of the Dissociation-Rate Constant for Inhibitor-Enzyme Complexes via the T1r and T2 (CPMG) Methods. *J. Magn. Reson., Ser. B* **1994**, *104*, 266−275.
(48) Carver, J. P.; Richards, R. E. A General Two-Site Solution for the Chemical Exchange Produced Dependence of T2 upon the Carr-Purcell Pulse Separation. *J. Magn. Reson.* **1972**, *6*, 89−105.
(49) Palmer, A. G.; Kroenke, C. D.; Loria, J. P. Nuclear Magnetic Resonance Methods for Quantifying Microsecond-to-Millisecond Motions in Biological Macromolecules. *Methods Enzymol.* **2001**, *339*, 204−238.
(50) d'Auvergne, E. J.; Gooley, P. R. The Use of Model Selection in the Model-Free Analysis of Protein Dynamics. *J. Biomol. NMR* **2003**, *25*, 25−39.
(51) d'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models II. A New Methodology for the Dual Optimisation of the Model-Free Parameters and the Brownian Rotational Diffusion Tensor. *J. Biomol. NMR* **2008**, *40*, 121−133.
(52) d'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models I. Minimisation Algorithms and Their Performance within the Model-Free and Brownian Rotational Diffusion Spaces. *J. Biomol. NMR* **2008**, *40*, 107−119.
(53) Mandel, A. M.; Akke, M.; Palmer, A. G. Backbone Dynamics of Escherichia Coli Ribonuclease HI: Correlations with Structure and Function in an Active Enzyme. *J. Mol. Biol.* **1995**, *246*, 144−163.
(54) Mittermaier, A.; Kay, L. E. Measurement of Methyl H-2 Quadrupolar Couplings in Oriented Proteins. How Uniform Is the Quadrupolar Coupling Constant? *J. Am. Chem. Soc.* **1999**, *121*, 10608−10613.
(55) Skrynnikov, N. R.; Millet, O.; Kay, L. E. Deuterium Spin Probes of Side-Chain Dynamics in Proteins. 2. Spectral Density Mapping and Identification of Nanosecond Time-Scale Side-Chain Motions. *J. Am. Chem. Soc.* **2002**, *124*, 6449−6460.
(56) Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, 5th ed.; Brooks/Cole Publishing Company: Monterey, 1999.
(57) Li, D. W.; Brüschweiler, R. A Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131*, 7226−7227.
(58) Jarymowycz, V. A.; Stone, M. J. Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences. *Chem. Rev.* **2006**, *106*, 1624−1671.
(59) Akke, M. Conformational Dynamics and Thermodynamics of Protein-Ligand Binding Studied by NMR Relaxation. *Biochem. Soc. Trans.* **2012**, *40*, 419−423.
(60) Caro, J. A.; Harpole, K. W.; Kasinath, V.; Lim, J.; Granja, J.; Valentine, K. G.; Sharp, K. A.; Wand, A. J. Entropy in Molecular Recognition by Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 6563−6568.
(61) Case, D. A.; Berryman, J. T.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T E.; LeGrand, S.; Li, P.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Monard, G.; Needham, P.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Salomon-Ferrer, R.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; York, D. M.; Kollman, P. A. *AMBER 2015*; University of California: San Francisco, 2015.
(62) Genheden, S.; Diehl, C.; Akke, M.; Ryde, U. Starting-Condition Dependence of Order Parameters Derived from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2010**, *6*, 2176−2190.
(63) Genheden, S.; Akke, M.; Ryde, U. Conformational Entropies and Order Parameters: Convergence, Reproducibility, and Transferability. *J. Chem. Theory Comput.* **2014**, *10*, 432−438.
(64) Uranga, J.; Mikulskis, P.; Genheden, S.; Ryde, U. Can the Protonation State of Histidine Residues Be Determined from Molecular Dynamics Simulations? *Comput. Theor. Chem.* **2012**, *1000*, 75−84.
(65) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of


DOI: 10.1021/jacs.8b11099
J. Am. Chem. Soc. 2019, 141, 2012−2026

Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(66) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665−9678.
(67) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(68) Bayly, C. C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(69) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.
(70) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2016.
(71) Seminario, J. M. Calculation of Intramolecular Force Fields from Second-Derivative Tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271−1277.
(72) *Turbomole* V7.01; Turbomole GmbH, 2007; http://www.turbomole.com.
(73) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. An Automatic Method to Generate Force-Field Parameters for Hetero-Compounds. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 274−289.
(74) Genheden, S.; Ryde, U. A Comparison of Different Initialization Protocols to Obtain Statistically Independent Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 187−195.
(75) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(76) Wu, X.; Brooks, B. R. Self-Guided Langevin Dynamics Simulation Method. *Chem. Phys. Lett.* **2003**, *381* (3−4), 512−518.
(77) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(78) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089.
(79) Prompers, J. J.; Brüschweiler, R. General Framework for Studying the Dynamics of Folded and Nonfolded Proteins by NMR Relaxation Spectroscopy and MD Simulation. *J. Am. Chem. Soc.* **2002**, *124*, 4522−4534.
(80) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* **1984**, *51*, 1011−1028.
(81) Trbovic, N.; Cho, J.-H.; Abel, R.; Friesner, R. A.; Rance, M.; Palmer, A. G. Protein Side-Chain Dynamics and Residual Conformational Entropy. *J. Am. Chem. Soc.* **2009**, *131*, 615−622.
(82) King, B. M.; Tidor, B. MIST: Maximum Information Spanning Trees for Dimension Reduction of Biological Data Sets. *Bioinformatics* **2009**, *25*, 1165−1172.
(83) King, B. M.; Silver, N. W.; Tidor, B. Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. *J. Phys. Chem. B* **2012**, *116*, 2891−2904.
(84) Fogolari, F.; Maloku, O.; Dongmo Foumthuim, C. J.; Corazza, A.; Esposito, G. PDB2ENTROPY and PDB2TRENT: Conformational and Translational-Rotational Entropy from Molecular Ensembles. *J. Chem. Inf. Model.* **2018**, *58*, 1319−1324.
(85) Nguyen, C. N.; Kurtzman Young, T.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril. *J. Chem. Phys.* **2012**, *137*, 973−980.
(86) Peterson, K.; Kumar, R.; Stenström, O.; Verma, P.; Verma, P. R.; Håkansson, M.; Kahl-Knutsson, B.; Zetterberg, F.; Leffler, H.; Akke, M.; Logan, D. T.; Nilsson, U. J. Systematic Tuning of Fluoro-Galectin-3 Interactions Provides Thiodigalactoside Derivatives with Single-Digit nM Affinity and High Selectivity. *J. Med. Chem.* **2018**, *61*, 1164−1175.
(87) Halle, B.; Wennerström, H. Interpretation of Magnetic Resonance Data from Water Nuclei in Heterogeneous Systems. *J. Chem. Phys.* **1981**, *75*, 1928−1943.
(88) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. I. Theory and Range of Validity. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.
(89) Schichman, S. A.; Amey, R. L. Viscosity and Local Liquid Structure in Dimethyl Sulfoxide-Water Mixtures. *J. Phys. Chem.* **1971**, *75*, 98−102.
(90) Catalan, J.; Diaz, C.; Garcia-Blanco, F. Characterization of Binary Solvent Mixtures of DMSO with Water and Other Cosolvents. *J. Org. Chem.* **2001**, *66*, 5846−5852.
(91) Showalter, S. A.; Bruschweiler, R. Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field. *J. Chem. Theory Comput.* **2007**, *3*, 961−975.
(92) Li, D. W.; Showalter, S. A.; Brüschweiler, R. Entropy Localization in Proteins. *J. Phys. Chem. B* **2010**, *114*, 16036−16044.
(93) Haider, K.; Cruz, A.; Ramsey, S.; Gilson, M. K.; Kurtzman, T. Solvation Structure and Thermodynamic Mapping (SSTMap): An Open-Source, Flexible Package for the Analysis of Water in Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2018**, *14*, 418−425.
(94) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817−2831.


DOI: 10.1021/jacs.8b11099
*J. Am. Chem. Soc.* 2019, 141, 2012−2026

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

ISBN: 978-91-7422-662-1

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University