

Expanding the Toolbox for Computational Analysis in
Rational Drug Discovery:
Using Biomolecular Solvation to Predict Thermodynamic,
Kinetic and Structural Properties of Protein-Ligand Complexes

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat)

dem
Fachbereich Pharmazie der
Philipps-Universität Marburg
vorgelegt von

M. Sc. Chem.
Tobias Hüfner
(geb. Wulsdorf)
aus Hünfeld

Marburg/Lahn 2019

Erstgutachter: Prof. Dr. Gerhard Klebe
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg

Zweitgutachter: Prof. Dr. Peter Kolb
Institut für Pharmazeutische Chemie
Philipps-Universität Marburg

Eingereicht am 22.08.2019

Tag der mündlichen Prüfung am 22.10.2019

Hochschulkennziffer: 1180

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. Gerhard Klebe am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von Juli 2015 bis Juli 2019 durchgeführt.

"The Computer is a bicycle for the mind." – Steve Jobs

Abbreviations

ACE	Acetyl (capping group)
ADME	Adsorption-Distribution-Metabolism-Excretion
ALR2	Aldose-Reductase type 2
b3lyp	Becke, 3-parameter, Lee-Yang-Parr
BB	Building Block
CPU	Central Processing Unit
CSD	Cambridge Structural Database
DAA	Diamanoid Amino Acid
DBI	Davies–Bouldin index
DFT	Density Functional Theory
DNA	Deoxyribonucleic acid
EP	Endothiapepsin
ESP	Electrostatic Potential
FBDD	Fragment-based Drug Discovery
FEP	Free Energy Perturbation
FES	Free Energy Surface
fs-IR	Femtosecond Infrared Spectroscopy
GAFF	Generalized Amber Force Field
GCMC	Grand-Canonical Monte Carlo
Gips	GIST-based Processing of Solvent Functionals
GIST	Grid Inhomogeneous Solvation Theory
GPCR	G-Protein Coupled Receptor
GPU	Graphics Processing Unit
HF	Hartree-Fock
HSA	Hydration Site Analysis
IMM	Impey-Madden-McDonald approach
IST	Inhomogeneous Solvation Theory
ITC	Isothermal Titration Calorimetry
JAWS	Just Add Water Molecules
LoCorA	Local Correlation Analysis
MC	Monte Carlo

MD	Molecular Dynamics
MMGBSA	Molecular Mechanics Generalized-Born Surface Area
MMPBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MOE	Molecular Operating Environment
MRBB	Minimal Representation Building Block
MRT	Mean Residence Time
MUE	Mean Unsigned Error
NICS	Nucleus Independent Chemical Shift
NME	N-Methyl (capping group)
NMR	Nuclear Magnetic Resonance Spectroscopy
NOE	Nuclear Overhauser Effect
NSGA2	Non Dominated Sorting Genetic Algorithm 2
PAINS	Pan-Assay Interference Compounds
PDB	Protein Data Bank
PKA	Protein Kinase A
PKI	Protein Kinase Inhibitor
PLS	Partial Least Squares
PMF	Potential Of Mean Force
psF	Pseudo F
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
RDF	Radial Distribution Function
RESP	Restricted Electrostatic Potential
RISM	Reference Interaction Site Model
RNA	Ribonucleic Acid
SAR	Structure Activity Relationship
s.d.	Standard Deviation
SBDD	Structure-based Drug Discovery
SLSQP	Sequential Least Squares Programming
SPR	Surface Plasmon Resonance
SSP	Stable State Picture
TCF	Time Correlation Function
TLN	Thermolysin

US	Umbrella Sampling
WBI	Wiberg Bond Index
WHAM	Weighted Histogramm Analysis Method

Zusammenfassung

Die meisten biomolekularen Interaktionen finden im wässrigen Medium statt. Daher ist es wichtig die Interaktionen zwischen Proteinen und Wassermolekülen in der Wirkstoff-Forschung zu berücksichtigen. Die Untersuchung dieser Interaktionen mittels experimenteller Methoden ist anspruchsvoll, daher werden häufig Computer-Simulationen verwendet um die molekularen Details von Protein-Wasser oder Ligand-Wasser-Interaktionen zu studieren.

Im zweiten Kapitel der vorliegenden Doktorarbeit wird die Entwicklung, Parametrisierung und Erprobung eines Ansatzes vorgestellt, der zur Berechnung der Solvatations-Beiträge in Protein-Ligand Bindungsreaktionen verwendet werden kann. Der Ansatz verwendet eine umfassende Menge an Trajektorien aus Moleküldynamik-Simulationen in Kombination mit GIST Berechnungen um Modelle zu erhalten, mit welchen die relativen Beiträge zur Protein-Ligand Solvatations-Thermodynamik vorhergesagt werden können. Um den Ansatz zu validieren wurde das Model System Thrombin mit einem Satz von 53 Liganden mit bekannter Kristallstruktur und ITC Profilen untersucht. Dabei wurde herausgefunden, dass die Bindungs-Thermodynamik von insgesamt 186 Paaren von Liganden genau vorhergesagt werden kann. Die relative Freie Energie der Bindung für diese 186 Paare kann dabei schon alleinig aus der Desolvatation des freien Liganden ermittelt werden. Im Weiteren werden vollständige thermodynamische Profile für Protein-Ligand Bindungsreaktionen korrekt vorhergesagt.

Im dritten Kapitel wird der zuvor vorgestellte Ansatz verwendet um eine Strategie zu entwickeln die es ermöglicht Wirkstoffe mit gewünschter Solvatations-Thermodynamik auszustatten. Für diesen Zweck werden die Thrombin-Liganden (gleiche Liganden Serie wie im vorangegangenen Kapitel 2) in kleinere molekulare Bausteine zerlegt. Im nächsten Schritt wird die Solvatations-Thermodynamik eines jeden Bausteins im Liganden ebenso wie für den isolierten Baustein in wässriger Lösung berechnet. Dabei wurden sehr diverse Eigenschaften für die verschiedenen Bausteine gefunden, was deren Potential zum Entwurf von Liganden mit einer großen Bandbreite von Solvatations-Charakteristika ermöglicht. Ebenso wurden Fernstrukturierungseffekte von Wassermolekülen entdeckt. Diese Effekte konnten nur durch die Zerlegung der Liganden und der korrespondierenden GIST-Integrale in einzelne Bausteine ermöglicht werden. Die Fernstrukturierungseffekte treten im ungebundenen Liganden auf und beschreiben die verstärkte Strukturierung von Solvens-molekülen auf einer Baueinheit bedingt durch das Vorhandensein einer anderen Baueinheit auf einer entfernten Seite des Liganden. Im Weiteren wurde gezeigt, dass die Fluorierung von Baueinheiten zu erhöhten unvorteilhaften

Desolvationseigenschaften führt. Die Fluorierung führt daher zu einer reduzierten Bindungsaffinität. Die Forschungsarbeiten aus Kapitel 2 und 3 wurden mit Hilfe des Computerprogramms *Gips* durchgeführt, welches im Zuge dieser Doktorarbeit entwickelt wurde. In Kapitel 4 wird der Mechanismus und die Zeitskala der Desolvation für eine Protein-Ligand Dissoziationsreaktion für die von Trypsin und Thrombin im Komplex mit Benzamidin und *N*-amidinopiperidin untersucht. Die Untersuchung wird durchgeführt mittels „Umbrella Sampling“ und *LoCorA* Rechnungen. *LoCorA* ist eine Methode zur Analyse von Besetzungszeiten von Wassermolekülen auf der Oberfläche von Aminosäuren. Damit wurde herausgefunden, dass Wassermoleküle ungefähr 1.3 ns in der *apo* Bindetasche von Thrombin verweilen, wohingegen sie in der *apo* Bindetasche von Trypsin um eine Größenordnung kürzer verweilen (0.3 ns). Dieser Unterschied wird mit Solvens-Kanälen im Falle von Thrombin, und mit einem Solvens-Reservoir im Falle von Trypsin erklärt. Die Solvens-Kanäle bedingen, dass Wassermoleküle die gleichen Besetzungszeiten für beide Komplexe zeigen im Falle von Thrombin. Durch das Fehlen dieser Kanäle in Trypsin gibt es hier jedoch unterschiedliche Besetzungszeiten für die beiden Komplexe. Der *LoCorA* Ansatz ist implementiert in das Computerprogramm *LoCorA* (gleicher Name wie der Ansatz selbst), welches im Zuge dieser Doktorarbeit entwickelt wurde.

Weitere Studien die im Zuge dieser Doktorarbeit durchgeführt und mit experimentellen Untersuchungen kombiniert wurden, sind in Kapitel 5 dieser Dissertation zu finden. Zu jeder dieser Studien ist eine separate Zusammenfassung und Erläuterung bezüglich der Eigenanteile vorangestellt zu finden.

Abstract

Most biomolecular interactions occur in aqueous environment. Therefore, one must consider the interactions between proteins and water molecules when developing a drug molecule against a target protein. The study of these interactions is challenging using experimental techniques alone, therefore computer simulations are commonly used to study the molecular details of protein-water or ligand-water interactions.

In the first study presented in this doctoral dissertation (Chapter 2), the development, parameterization and testing of an approach is presented that can be used to calculate the solvation contribution in protein-ligand binding thermodynamics. The approach uses an extensive amount of molecular dynamics trajectories in conjunction with GIST calculations in order to obtain models that can predict relative protein-ligand solvation thermodynamics. In order to validate the approach, the model system thrombin is investigated using a set of 53 ligands with experimentally characterized protein-ligand structures and ITC profiles. We found that the binding thermodynamics of 186 congeneric pairs of ligands can be accurately described using our solvation-based models. The relative free energy of binding for these 186 pairs can be calculated from the desolvation free energy of the ligand molecules alone. Furthermore, complete thermodynamic profiles for protein-ligand binding reactions (i.e. free energy, enthalpy and entropy of binding) are accurately predicted by incorporating GIST solvent data from the unbound ligand as well as the protein-ligand complex.

In Chapter 3, the aforementioned approach is applied to develop a strategy that enables to equip drug molecules with a desired set of solvation thermodynamics properties. For this purpose, the thrombin ligands (same ligand series as in previous Chapter 2) and the corresponding GIST integrals are decomposed into smaller building block molecules. In the next step, the solvation thermodynamics for the building blocks in the ligand molecule as well as the solvation thermodynamics for the isolated building block in aqueous solution are calculated. We found greatly varying solvation thermodynamics for the different building blocks, demonstrating their potential to design ligands with a wide range of solvation characteristics. Also, we found that the building block decomposition of ligand molecules and the corresponding GIST integrals can be readily used to understand remote solvent structuring effects. These effects occur in the unbound ligand molecule and describe the enhanced solvent structuring on a building block in the ligand molecule due to the presence of another building block at a distal site of the ligand. Furthermore, we demonstrated that the fluorination of building blocks leads to an increased

unfavorable desolvation free energy and thus disfavors binding for the presented dataset. The research presented in Chapter 2 and Chapter 3 was accomplished with the computer program *Gips* that was developed as part of this doctoral dissertation.

In the following Chapter 4, the mechanism and time scale of desolvation is being analyzed for the protein-ligand dissociation reaction of trypsin and thrombin in complex with benzamidine and *N*-amidinopiperidine. The analysis is carried out using umbrella sampling free energy calculations and *LoCorA* calculations. The *LoCorA* approach is a method for the analysis of residence times of water molecules on the surface of amino acids. It was found that water molecules reside approximately 1.3 ns in the binding pocket of thrombin, whereas in trypsin they are residing one order of magnitude shorter (0.3 ns). This difference is explained with special solvent channels that connect the interior of the binding pocket to bulk solvent environment. The solvent channels are present in thrombin but not in trypsin. Furthermore, the selectivity profiles of benzamidine and *N*-amidinopiperidine are related to a solvent-mediated free energy barrier that is present in thrombin but not trypsin. Also due to the presence of the solvent channels, the water molecules show similar residence time for both complexes in the case of thrombin but differing residence times in the case of the two trypsin complexes. The *LoCorA* approach is implemented in the computer program *LoCorA* (same name as the approach itself), which was developed as part of this doctoral dissertation.

In the course of this doctoral dissertation, further computational studies were carried out in combination with experimental ones. These can be found in chapter 5 of this dissertation. Each of these studies is preceded by a separate abstract and a statement concerning the author contribution.

Table of Contents

ABBREVIATIONS	V
ZUSAMMENFASSUNG	VIII
ABSTRACT	X
1 INTRODUCTION	3
1.1 DRUG DISCOVERY IS A MULTI-OBJECTIVE OPTIMIZATION PROBLEM	3
1.2 MOLECULAR RECOGNITION AS A RATIONALE TO DRIVE DRUG DISCOVERY	6
1.3 THE USE OF THERMODYNAMICS IN THE STUDY OF PROTEIN-LIGAND INTERACTIONS	7
1.4 BIOMOLECULAR SOLVATION: THE STRUCTURAL PERSPECTIVE	10
1.5 COMPUTERS AND MOLECULAR INTERACTIONS	12
2 PROTEIN-LIGAND COMPLEX SOLVATION THERMODYNAMICS: DEVELOPMENT, PARAMETERIZATION AND TESTING OF GIST-BASED SOLVENT FUNCTIONALS	19
2.1 ABSTRACT	19
2.2 INTRODUCTION	20
2.3 MATERIALS AND METHODS	22
2.4 THEORETICAL BACKGROUND	27
2.5 RESULTS AND DISCUSSION	41
2.6 COMPARATIVE ANALYSIS OF THE APPLIED FUNCTIONALS	60
2.7 CONCLUSION	62
2.8 SUPPORTING MATERIAL	64

3	MAPPING SOLVATION THERMODYNAMICS ON BUILDING BLOCKS: A STRATEGY TO DESIGN BETTER BINDERS	78
3.1	ABSTRACT	78
3.2	INTRODUCTION	79
3.3	RESULTS	82
3.4	DISCUSSION	99
3.5	CONCLUSION	101
3.6	METHODS	102
3.7	SUPPORTING MATERIAL	110
4	THE ROLE OF WATER MOLECULES IN PROTEIN-LIGAND DISSOCIATION: AN ANALYSIS OF THE MECHANISMS AND KINETICS OF BIOMOLECULAR SOLVATION USING MOLECULAR DYNAMICS	120
4.1	ABSTRACT	120
4.2	INTRODUCTION	121
4.3	THEORETICAL BACKGROUND	126
4.4	RESULTS	131
4.5	DISCUSSION	152
4.6	CONCLUSION	156
4.7	MATERIALS AND METHODS	157
4.8	SUPPORTING MATERIAL	161
5	ADDITIONAL STUDIES	168
6	REFERENCES	177
	ACKNOWLEDGEMENTS	190
	CURRICULUM VITAE	192
	ERKLÄRUNG	196

1 Introduction

1.1 Drug Discovery is a Multi-Objective Optimization Problem

Aspects of drug discovery in pre-clinical efforts consist mainly of research comprising the elucidation and identification of a single or multiple target proteins, screening of large compound libraries and the optimization of promising compounds. Also, new compounds are tested for potential to toxic side effects, which may prevent the initialization of subsequent clinical stages. At this stage of research and development, methods from multiple scientific disciplines (such as medicine, chemistry, physics and computer science) contribute to the collective research objective. This multitude of scientific disciplines is necessary due to the complexity of drug discovery itself, which must be treated as a multi-objective optimization problem. Often, vast amounts of data must be processed, filtered and interpreted in order to validate experimental findings or suggest new experiments that eventually lead to novel therapeutically active compounds.¹ The multi-objective character of pre-clinical drug discovery may be divided into three main aspects (this is by no means meant to be a comprehensive list):

- [A] **Identification** and **validation** of the target protein
- [B] Finding a drug molecule that binds **tightly** to the target protein
- [C] Finding a drug molecule that binds **selectively** to the target protein
- [D] Finding a drug molecule that meets **ADME-Tox** (Absorption-Distribution-Metabolism-Excretion-Toxicology) requirements

In the initial step, a drug target protein is identified and validated (aspect [A]). As this is the first step in a cascade of development steps, it is most crucial for the success of a drug discovery campaign. During this initial phase, *in vitro* experiments are used to select the potential drug target but also animal models such as the zebrafish are used.² In human cancer research, the vast knowledge of molecular mechanisms and pathophysiology is exploited for mechanism-based target identification strategies.³ As soon as a protein has been identified as a potential drug target, a bioassay is established that enables the assessment of its biological activity. This is an important step, as it is used in the following steps for the selection and optimization of lead compounds. Although target-based strategies are seemingly efficient they are often criticized as they are associated with a decline in the number of compounds that enter clinical phases.⁴

A good drug molecule must bind tightly to a target protein (aspect [B]), thus the molecule is optimized with respect to its specific set of molecular interactions to a target protein. Its interactions with the target protein result in an effect on the cellular level and thus may lead to a therapeutic effect. In the best possible case, information about molecular interactions are gained by studying the three-dimensional structure of the protein and the drug molecule using experimental techniques such as X-ray crystallography⁵⁻⁷ or NMR⁸⁻¹⁰ spectroscopy. However, in many real-world scenarios, experimentally valid information about the three-dimensional structure is not available. In these cases, researchers must use a homology model¹¹ of the target protein. A homology model is a computationally predicted structure of the target protein that is based on various data sources mostly extracted from previously characterized and structurally related proteins. These models can be obtained (almost entirely) based on the amino acid sequence. However, it must be noted that in some cases the identity of the target protein is not known at all. Nonetheless, it is still possible to design active molecules without precise knowledge of the target structure.¹²⁻¹⁵ In any case, i.e. whether structural data are available or not, it is important to have a design objective that is based on a rationally-driven hypothesis about the molecular interactions of the involved biomolecules (for instance proteins, DNA, RNA or tRNA) and a drug molecule. A rational design hypothesis is often driven by physics-based models of the drug molecule and the target protein. These models may represent molecules on various levels of detail, ranging from the electronic structure to the (coarse) semi-atomistic scale. Thus, it is quite common in contemporary drug discovery to use these models together with a massive integration of computational approaches and resources into routine research and development workflows.¹⁵⁻²¹ In cases where physics-based models cannot be derived straightforwardly, one usually tries to learn from well-studied model protein systems in order to extrapolate to the actual system under study. It must be noted that although the use of structural data is extremely convenient, also other approaches such as QSAR (quantitative structure activity relationship) or QSPR (quantitative structure property relationship) are successfully applied.^{20,22-25} These approaches do not necessarily require information about the structure of the target molecular system.

Another aspect of pre-clinical drug discovery is selectivity (aspect [C], see previous page).^{26,27} Selectivity can be defined as the property of a molecule to bind more preferentially to a single target protein than to another protein (or a group of other proteins). In an ideal scenario, a potential drug molecule must be able to discriminate its target protein and the corresponding binding site from other proteins and binding sites, at least in pre-clinical investigations to

validate a given target. Failure to do so may result in unwanted side effects in later clinical phases, which may cause the rejection of a candidate molecule from further assessments. Nevertheless, particularly in the field of GPCRs many cases are known where mixed action against a set of targets make the quality of the desired therapeutic action. One example of these so-called “dirty drugs” is the anti-psychotic drug Chlorpromazine.²⁸

In order to circumvent situations in which unwanted side effects occur, pre-clinical research and development efforts aim at designing clinical candidates with an optimal selectivity profile, also to elucidate their mode of action. Computational approaches can efficiently accelerate this part of the design process by incorporating models from different proteins into the optimization of a drug molecule.^{19,29} It is important to note that under high concentrations of a ligand molecule, binding to a non-preferred protein may occur to a therapeutically relevant amount. Thus, the concept of selectivity must not be treated as an absolute measure for the discrimination between proteins but as a relative one.

Lastly, ADME-Tox is a critical aspect in pre-clinical drug discovery (aspect [D]) that relates to other disciplines such as pharmacokinetics, pharmacology and toxicology. The acronym ADME-Tox stands for *absorption, distribution, metabolism, excretion* and *toxicology*. These properties are commonly linked to physical properties by the *Lipinski's rule of five*, which readily estimates a compound's drug-likeness based on its molecular weight, *logP* value and number of hydrogen bond donors/acceptors.^{30,31} The *absorption* of a drug molecule is described by the pathway that the drug undergoes while it enters the human body and different administration pathways can be selected. The pathway critically affects the bioavailability of the drug and thus is an important factor that must be taken into account early on in the drug development process. The bioavailability is often directly related to basic physical properties such as solubility, lipophilicity or pH stability.^{25,32} The aspect of *distribution* relates to the transport of the drug compound to its effector site. Usually the drug is first circulated through the body via the bloodstream and then gets distributed to the effector site(s). There are special cases where the distribution is hindered by barriers, such as the blood-brain barrier, which requires special strategies to be overcome effectively.³³ Once the drug has entered the body, it undergoes various paths of chemical decomposition, which are referred to as *metabolism*. Most of the known metabolic decomposition processes take place in the liver. In this organ, predominantly a special group of proteins, the cytochrome P450 enzymes, carry out the molecular modifications of drug molecules into smaller molecular species using a cascade of oxidation steps. These smaller molecular species are called metabolites and can be more active

than the parent drug or even toxic compounds can be generated. Thus, the metabolic paths as well as the identity of possible metabolites must be considered while developing a drug molecule. In some cases, the occurrence of metabolites is specifically desired as they have superior activity compared to the parent drug (e.g. pro-drug).

Excretion involves the various mechanisms by which a drug (also its metabolites) can exit the body. A major exit pathway runs via the kidneys, where drugs and metabolites are excreted in the form of urine. Other exit pathways involve the excretion via feces, lungs or the skin.

The final aspect of ADME-Tox is the toxicological behavior of the compound. A key parameter for the characterization of the toxicity of a drug compound is its lethal dose. In this context, various *in silico* approaches have emerged that attempt to predict the toxicity of a compound based on comprehensive data sets.^{34,35}

1.2 Molecular Recognition as a Rationale to Drive Drug Discovery

As already introduced in the previous subsection about the origin and need to design tight-binding drug molecules, molecular interactions are used as a fundamental concept to understand the behavior of a potential drug compound with respect to a target protein. When using the term “drug”, one usually refers to a functional representation of a molecule that is ultimately related to some sort of therapeutic use. However, in the context of molecular interactions, one must correctly refer to the term “ligand” (derived from the Latin word *ligandus*, which is the gerundive form of *ligo*, meaning “bind”), as one will only consider the fact that the molecule, i.e. the ligand, physically interacts (it “binds”) with the protein. The ligand and protein shape an assembly, termed protein-ligand complex (or for short “complex”), that is the basis for all thermodynamic and structural considerations.

The fact that ligands are able to bind to macromolecules with a specific set of interactions is often referred to as a molecular recognition process,³⁶⁻³⁸ which was also awarded with the 1987 Nobel Prize in Chemistry. The intuitively emerging picture in this context divides the reaction partners into a host (e.g. a protein) and a guest (e.g. a peptide substrate) molecule. The host and the guest molecule undergo molecular interactions based on their molecular complementarity. Based on this principle, very successful computational approaches, such as molecular docking, have emerged and are routinely applied in drug discovery pipelines in order to perform a so-called “virtual screening” of large compound libraries.^{18,19} The molecular interactions that effectively form any sort of molecular complementarity are electrostatic interactions, van der

Waals interactions, π - π interactions, halogen bonding or hydrogen bonding.^{39,40} It must be noted that these various types of interactions cannot always be strictly separated from each other as they are partly related to similar fundamental physical principles. The concept of molecular recognition is related to the simplified assumption of a lock-and-key-model as commonly employed to illustrate enzyme-substrate interactions. This implies that a ligand fits into a protein, as a key fits into a lock. This very static picture of protein-ligand interaction neglects the dynamic and highly coupled behavior of the large amount of molecular degrees of freedom that are present in macromolecular species (such as proteins) and the multiple solvent molecules. Moreover, it is known that some proteins are highly adaptive and can open additional (transient) subpockets upon binding of the ligand. Depending on whether the ligand induces the opening of the pocket or if the protein opens the pocket on its own, this process is called either induced-fit or conformatoinal selection. In any case, it is a superior model of protein-ligand complex formation as it directly relates to the various degrees of freedom given for a macromolecule such as a protein. Computational methods that explicitly consider the molecular degrees of freedom, such as molecular dynamics or Monte Carlo simulations, have emerged over the last years and are now an important part of drug discovery.⁴¹ These methods are suitable in cases where high-throughput processing is not desirable, as an enhanced level of molecular detail is necessary in order to understand the system under study. In the present doctoral dissertation, this concept was realized and will be further introduced in the section “Computers and Molecular Interactions”.

1.3 The Use of Thermodynamics in the Study of Protein-Ligand Interactions

In the previous subsection, the concept of molecular interactions and its relationship to molecular recognition has been introduced. It was outlined, how this concept is critical in the development of drug molecules. However, so far it was not explained how exactly our considerations on the atomistic level relate to actual physical observables, such as equilibrium constants, turn-over rates or some read-out from a biophysical experiment.

The relation between atomistic considerations and actual experiments is established by using concepts from chemical thermodynamics. Thermodynamics as a branch of physics deals with measurable macroscopic physical quantities such as temperature, pressure, volume, heat or work. Relations between these quantities are established by an axiomatic set of laws (the four

laws of thermodynamics), which introduce important physical quantities such as internal energy or entropy. At this point, a fundamental equation for the calculation of protein-ligand thermodynamics is introduced:

$$\Delta G^0 = -RT \ln K_B \quad (1-1)$$

$$K_B = \prod_i^N a_i^{v_i} \quad (1-2)$$

In eq. (1-1), ΔG^0 is known as the standard Gibbs free energy and is a measure for the maximal amount of reversible work that can be performed by a system. It is calculated from the universal gas constant R ($8.3144 \text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$), the absolute temperature T and the equilibrium binding constant K_B . The equilibrium binding constant K_B is defined as the product of the activities a_i of all N species in the system with stoichiometric coefficients v_i (see eq. (1-2)) at standard conditions. For practical considerations, the activity of some species X_i can be well approximated by its equilibrium concentration $[X_i]$. Thus, for the case of protein-ligand interactions, the equilibrium binding constant for some binding reaction $P+L \rightarrow PL$ may be formulated as follows:

$$K_B = \frac{1}{K_D} = \frac{[PL]}{[P][L]} \quad (1-3)$$

In eq. (1-3), K_D is the dissociation constant, which is the inverse of the binding constant. The dissociation constant can be interpreted as the equilibrium concentration of ligand $[L]$, at which the equilibrium concentrations of the protein-ligand complex $[PL]$ and the free protein $[P]$ are equal. Thus, K_D is an intuitive measure for the ability of a ligand molecule to bind to a protein and can be readily obtained by measuring equilibrium concentrations.

As has been shown by eq. (1-1), there is a direct relationship between the binding constant and the standard Gibbs free energy. The standard Gibbs free energy can be decomposed into standard enthalpy, ΔH^0 , and standard entropy, ΔS^0 , contributions

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 \quad (1-4)$$

These contributions are especially insightful, as they are a means to the composition of the standard Gibbs free energy and consequently, also of the equilibrium constant (see eq. (1-1)).

The binding standard enthalpy, ΔH^0 , is generally assumed to be a measure of the change in internal energy upon binding due to molecular interactions (neglecting contributions from pressure-volume work). The difference between standard enthalpy and standard Gibbs free energy is the standard entropy multiplied by the absolute temperature of the system. If the temperature of the system is >0 K (which must be the case due to the third law of thermodynamics), then all atoms are under thermal motion. However, due to molecular interactions, the protein and ligand molecule are locked together and cannot move freely. Thus, the system is put under restraints which countervail the intrinsic random thermal motions of the system due to its temperature. These restraints lead to a decrease in entropy and consequently to an increase in standard Gibbs free energy (see eq. (1-4)). The standard enthalpy and entropy of binding can be readily obtained from experimental techniques, such as isothermal titration calorimetry. Thus, eq. (1-4) directly provides access to equilibrium binding properties on the molecular level.

From the argumentation above, it is evident that enthalpy and entropy are mutually coupled as both are dependent on the strength of molecular interactions. So far, one would intuitively assume that enthalpy and entropy must compensate each other, since an increase in molecular interaction energy would increase the restraints on the molecules, thus countervailing entropy. However, in many cases enthalpy and entropy do not compensate each other completely. Also, there are several cases where they actually reinforce each other. This has led to several controversies about our general understanding of entropy.⁴²⁻⁴⁴ As proteins are a system of many tightly coupled mechanical degrees of freedom, the binding of a ligand molecule can result in enhanced thermal motion in the actual binding site.⁴⁵

Another critical aspect of binding thermodynamics is arising from interactions with water molecules.⁴⁶⁻⁵⁰ Water is a ubiquitous substance and biomolecules are generally adapted to an aqueous environment.⁵¹ Consequently, also protein binding sites are to a certain amount filled with water molecules. The water molecules in protein binding sites generally have different properties than the unbound water molecules in bulk water phase. As soon as a ligand molecule (i.e. a drug molecule or substrate molecule) binds into the binding pocket, water molecules are released from the binding pocket into bulk water phase. Although not fully understood,⁵² the process is generally regarded as being entropically beneficial, as the water molecules experience fewer restraints in the bulk water phase as in the protein binding pocket. The entropic benefit is balanced with the gain or loss in interaction energy by the breaking or making of bonds in the binding pocket and bulk water phase (see also Figure 1-1). In this context,

hydrophobic surface patches in protein pockets are of special interest as water molecules become restrained as soon as they are bound, however at a minimal amount of solute-solvent interaction energy.^{46,53} Thus, the release of those water molecules into bulk water phase is expected to result in minimal energetic cost and maximal entropic gain. This concept is commonly referred to as *classical hydrophobic effect*. A shift in interactions of hydrophobically bound water molecules upon ligand binding can lead to a gain in binding affinity of several orders of magnitude.⁵³ Thus, an improved understanding of the thermodynamics and molecular interactions established by water molecules can greatly improve drug discovery efforts.

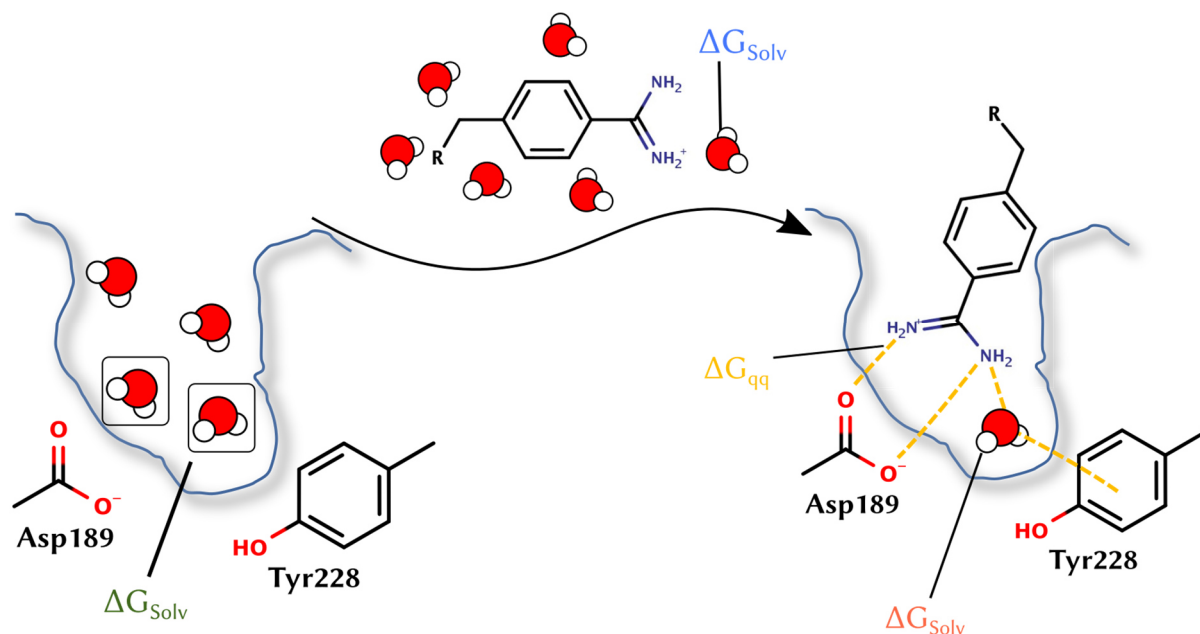


Figure 1-1: Schematic representation of a binding mechanism. The contributions to binding free energy due to various interactions are highlighted (ΔG_{Solv} : free energy from solvation contributions; ΔG_{qq} : free energy from charge-charge interaction contributions). This example depicts the S1 binding pocket of thrombin with a benzamidine head-group of a typical thrombin-inhibitor.

1.4 Biomolecular Solvation: The Structural Perspective

In the previous subsection, the relation between molecular interactions and thermodynamics was established. In the second part, the important contributions of water molecules to protein-ligand binding thermodynamics were explained. In the following subsection, several aspects relevant for the experimental elucidation of protein-water interactions will be introduced.

From the experimental structural perspective, water molecules are hard to capture. Even in modern high-resolution X-ray protein crystallography, hydrogen atoms are (usually) not resolved and consequently, the orientation of water molecules cannot be determined explicitly.

However, in many cases the environment of water molecules allows only for quite a limited number of possible orientations due to hydrogen bonding constraints. Protein crystallographic structure determination based on neutron scattering reveals the position of hydrogen atoms and thus also the orientation of water molecules.⁵⁴ However, these measurements usually take very long and are experimentally in many cases not feasible.

The completely different method NMR is also often used in drug development. It has many advantages compared to crystallography both with respect to sample preparation but also with respect to the fact that it captures the solution dynamics of the protein. However, it cannot be used in all cases and for meaningful evaluations even requires isotope labeling of the protein.⁸ Also, the resolution is often worse compared to structures determined by crystallographic experiments. As water molecules enter and leave the binding site of a protein at frequencies that are often faster than the timescale accessible by NMR, water-water or water-solute interactions cannot be resolved in most cases.

Experimental techniques have individual limitations that must be taken into account when used to rationalize the thermodynamics of protein-ligand binding reactions. In many cases, the interplay between biomolecular solvation and protein-ligand binding thermodynamics can be readily analyzed using crystallography. One popular example is the protein thrombin, an enzyme from the family of serine proteases, which is a well-studied model system. Furthermore, it is of therapeutic relevance due to its important role in the human blood coagulation cascade. A comparative example of two crystal structures of thrombin-inhibitor complexes can be seen in Figure 1-2. In this example, one can see that the polar *meta*-pyridyl moiety interacts with a water molecule in the S1 binding pocket (Figure 1-2A). This water molecule is able to further interact with two other water molecules and with Asp189. In the analogous derivative with a phenyl moiety (Figure 1-2B), only three water molecules are present due to missing polar interactions of water molecules with the aromatic ring. Due to the missing water molecule, also the other two water molecules have less interaction partners available in the S1 binding pocket. This lack of interactions causes an increase in the binding enthalpy value of $\Delta\Delta H = 4.1 \text{ kJ}\cdot\text{mol}^{-1}$ for the transition of the pyridyl moiety (Figure 1-2A) to the phenyl moiety (Figure 1-2B). At the same time, the value of the entropy contribution to the free energy of binding decreases by $-T\Delta\Delta S = -10.0 \text{ kJ}\cdot\text{mol}^{-1}$ due to missing restrictions imposed on the water molecules in the presence of the phenyl group. This illustrates the thermodynamic interpretation of molecular entropy as presented in the previous section.

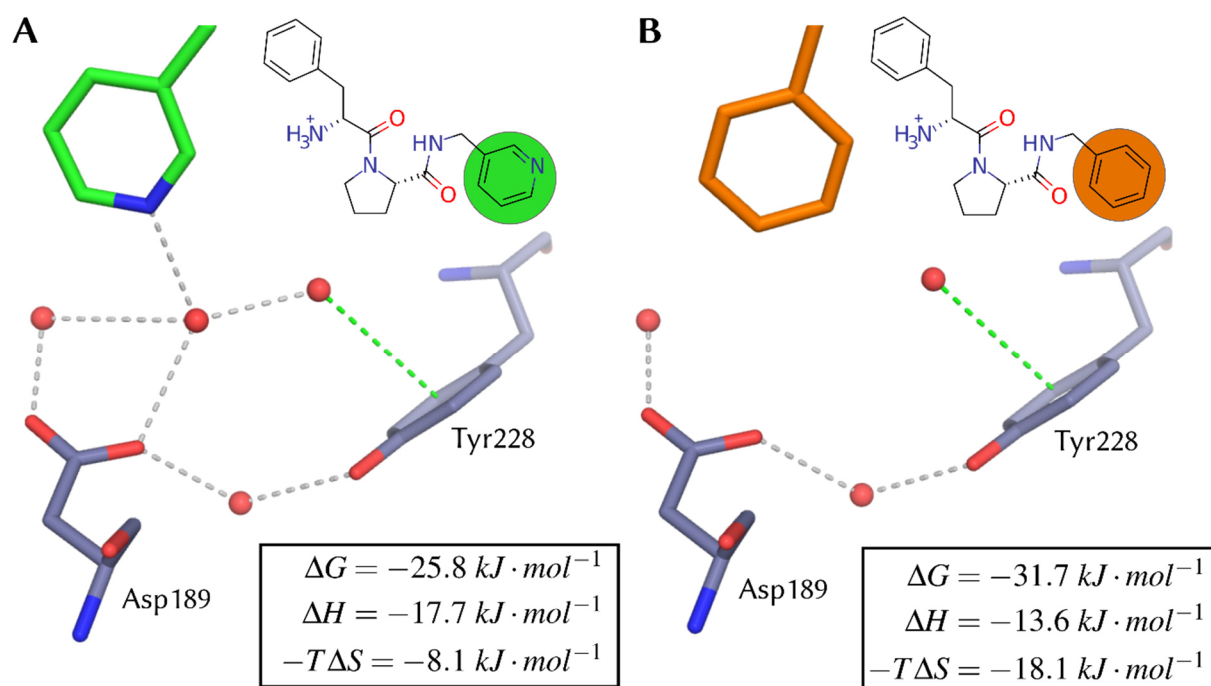


Figure 1-2: Example of the S1 binding pocket in crystal structures of thrombin-inhibitor complexes. The dashed grey lines indicate hydrogen bonding interactions, whereas the dashed green lines indicate interactions with the aromatic system. The part of the ligand that is shown in the crystal structure is also highlighted in the 2d depiction of the ligand. (A) Crystal structure 2ZFF; (B) Crystal structure 3P17.

1.5 Computers and Molecular Interactions

1.5.1 Approaches for the Treatment of Molecular Interactions in Computer Programs

In the previous section, the scientific field of drug discovery, specifically pre-clinical drug discovery, was introduced. The main focus of this section concentrated on the role of molecular interactions and biomolecular solvation in the context of molecular recognition. In order to gain insights into molecular interactions and biomolecular solvation, computers and computer simulations are routinely applied in drug discovery. In the following section, the main approaches for the treatment of molecular interactions in computational chemistry software packages are introduced.

In order to bridge the gap between experimentally determined structure and thermodynamic data, computational approaches are frequently applied. Depending on the features of the underlying system, different computational methods are used and, in many cases, a combination of multiple methods is applied. In the field of drug discovery, molecular dynamics simulations have emerged as a powerful computational technique, as they provide sufficient atomistic detail at reasonable computational costs. In molecular dynamics simulation, it is common to use a

molecular mechanics force field for the calculation of the interaction energy between atoms, but also for the internal mechanical degrees of freedom. These molecular mechanics force fields (or for short only “force fields”) are essentially an additive approach for the calculation of the system energy based on classical mechanics. Most force fields have the general functional form:

$$E_{MM}(\vec{x}) = E_{bonded}(\vec{x}) + E_{nonbonded}(\vec{x}) \quad (1-5)$$

$$E_{bonded}(\vec{x}) = E_{bond}(\vec{x}) + E_{angle}(\vec{x}) + E_{torsion}(\vec{x}) \quad (1-6)$$

$$E_{nonbonded}(\vec{x}) = E_{elec}(\vec{x}) + E_{vdw}(\vec{x}) \quad (1-7)$$

In eq. (1-5), the total system molecular mechanics energy, $E_{MM}(\vec{x})$, takes the configuration of the system, \vec{x} , as its argument (see eq. (1-5)). The total system energy is calculated from individual energetic contributions accounting for interactions between mutually bonded atoms (i.e. atoms that are one, two or three bonds apart), E_{bonded} , and interactions between nonbonded atoms (i.e. atoms that are more than two bonds apart or in different molecules), $E_{nonbonded}$. These two terms are further broken down into several individual contributions (see eqs. (1-6) and (1-7)), which have the following meaning:

- 1.) E_{bond} : The energy of bond stretching, e.g. a C-C bond in an alkyl chain. Typically calculated from Hooke’s law with the general functional form

$$E(d) = \frac{k}{2}(d - d_0)^2 \quad (1-8)$$

k : force constant

d_0 : equilibrium bond length

- 2.) E_{angle} : The energy of the angular stretching deformations of three consecutive atoms, e.g. the H-O-H angle in a water molecule. This energy functional is also approximated with Hooke’s law (see eq. (1-8)).

- 3.) $E_{torsion}$: The energy of a torsion potential based on four atoms, e.g. the O-C-N-H torsion in an amide group. The functional form for the calculation of this energy term slightly varies between different force fields. It is most common to use a series of cosine functions, which is expressed as

$$E(\tau) = \sum_{torsions} A(1 + \cos(n\tau - \phi)) \quad (1-9)$$

A : amplitude

n : periodicity

τ : torsion angle

ϕ : phase factor

- 4.) E_{elec} : The energy due to the pairwise interaction between the partial charges on two atoms, e.g. the electrostatic interaction between an oxygen atom of a water molecule and a hydrogen atom in an amide group. The electrostatic interaction energy between two atoms i and j , is modelled by a classical Coulomb potential of the general functional form

$$E(r_{ij}) = -\frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \quad (1-10)$$

q_i, q_j : atomic charge of i and j

ϵ_0 : vacuum electric permittivity

r_{ij} : separation between i and j

- 5.) E_{vdw} : The energy due to pairwise van-der-Waals interactions, e.g. between an sp² carbon atom in the tyrosine side chain and the oxygen atom in a water molecule. For a pair of atoms i and j , these interactions are typically modeled by a Lennard-Jones potential:

$$E(r_{ij}) = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \quad (1-11)$$

A_{ij}, B_{ij} : Lennard-Jones parameters atom pair ij

r_{ij} : separation between i and j

The various functional forms introduced above (eqs. (1-8)-(1-11)) require several parameters. A common approach for obtaining these parameters, is to fit the individual force field terms to high-level *ab initio* data.^{55,56} Other approaches use data from NMR experiments in order refine force field parameters,^{56,57} which is helpful for adjusting the stability of secondary structure elements in protein structures. Another commonly used approach is to fit the parameters to experimentally derived values of liquid state properties (such as density, heat of evaporation) for various compounds.^{58,59} In many cases, a mixture of all these types of parameterization have emerged and many different force field derivatives exist that are optimized for a specific set of physical conditions or class of molecules.

Most molecular mechanics force fields are not able to explicitly treat chemical reactions (although exceptions such as *ReaxFF* exist^{60,61}). However, quite often chemical reactions occur in addition to the non-covalent interactions and therefore must be taken into account. In these cases, molecular systems are treated based on quantum chemical calculations. Also, the mixed treatment of interactions using quantum chemical and molecular mechanics is quite popular,⁶² especially when investigating enzymatic reactions,⁶³ light-induced reactions⁶⁴ or protonation reactions.⁶⁵ However, quantum chemistry calculations are very time-consuming compared to force field type calculations. Therefore, one must decide whether it is worth using quantum chemistry calculations based on the expected insights gained by these calculations. As a popular alternative to high-level quantum chemistry calculations, semi-empirical quantum chemistry methods based on the AM1⁶⁶ or PM6⁶⁷ functionals have been developed and come at a reduced computational cost.

In a completely different approach, the molecular mechanics force field is entirely heuristic (or knowledge-based) and does not dictate an explicit functional form to the molecular potential.^{68,69} In this context, one usually uses the term “scoring function” instead of “force field”, as it is not based on the physical representation of forces. In heuristic scoring functions, large structural databases (such as the PDB or CSD) are scanned for the occurrence of specific interatomic separations of all sorts of atom types.^{68,69} From the distribution of these occurrences, one can calculate a score for all pairs of atom types based on their interatomic separation. The score for a pair of atoms essentially reports how “good” or how “bad” their current interatomic separation is with respect to the corresponding (knowledge-based)

distribution of interatomic separations. These scoring functions are applied for evaluating results from docking calculations or crystal structures. The main benefit of heuristic scoring functions is, compared to entirely physics-based force fields, their ability to judge a result based on actual experimental evidence from a manifold of experiments.⁷⁰ At the same time, this strength can also be seen as a caveat, as the accuracy and precision of any heuristic scoring function is entirely limited to and biased by the data it is derived from. Thus, care must be taken when using heuristic scoring functions outside the scope of their parameterization.

1.5.2 Computational Approaches for the Dissection of Molecular Solvation Thermodynamics

In the previous subsection, several general approaches for the calculation of molecular interactions using computer programs have been introduced. Special emphasis has been taken on force fields, which are mainly used in this doctoral dissertation. Other approaches, such as quantum chemistry and heuristic scoring functions were also explained briefly. In the following subsection, computational approaches that allow for the structural and thermodynamic characterization of water molecules are introduced.

Ever since researchers investigated protein-ligand interactions using X-ray crystallography, water molecules that mediate contacts between protein and ligand or solvate residues in an *apo* protein binding pocket have attracted computational chemists. In several successful attempts, the binding free energy contributions of these water molecules were estimated using alchemical methods.^{71,72} In the context of this class of methods, specialized approaches such as GCMC⁷³⁻⁷⁶ or JAWS^{74,77} have emerged. Despite their accuracy, these methods are usually quite time consuming and therefore do only allow for investigating few cases at a time. As an alternative to these computationally intensive methods, other approaches such as WaterMap,^{78,79} GIST,⁸⁰⁻⁸³ SZMAP^{84,85} or Grid Cell Theory⁸⁶⁻⁸⁸ have been developed. From these approaches, WaterMap and GIST have become quite popular in drug discovery. Both approaches are based on the theoretical framework of *inhomogeneous solvation theory*⁸⁹⁻⁹¹ (developed by Themis Lazaridis) and are used for post-processing of molecular ensembles generated from molecular dynamics or Monte Carlo simulations. In WaterMap and GIST, solvent energy and solvent entropy contributions are calculated relative to bulk solvent energy and bulk solvent entropy. Thus, one effectively calculates the energy and entropy calculations for transferring a water molecule from a specific position at the solute surface (e.g. a binding site) to pure bulk solvent. The fundamental difference between the two approaches is in their spatial representation of

solvation thermodynamics properties, i.e. density, enthalpy and entropy: In WaterMap, solvation properties are averaged over spherical regions (typically with radius 1 Å) called *hydration sites* (see Figure 1-3A). These hydration sites often over-simplify the non-spherical density distribution around a density maximum. Nonetheless, this approach has resulted in several successful studies, in which the insights on the solvation properties have greatly enhanced the process of drug optimization.^{50,92,93} In GIST (grid inhomogeneous solvation theory), the properties of the water molecules are spatially represented as a three-dimensional grid (see Figure 1-3B). This allows for approximating the non-spherical density distribution of water molecules by small grid cells, typically with dimensions of 0.5x0.5x0.5 Å. A caveat of the grid-based approach is the necessity of more sampling (i.e. longer timescales in the case of molecular dynamics simulations) than for hydration sites in order to achieve convergence. Typically, convergence of water properties estimated with hydration sites is achieved in less than 10 ns, whereas for a GIST grid up to 50-100 ns of simulation time are required.⁸² The use of GIST for the analysis of solvation properties using molecular dynamics simulations is becoming increasingly popular. In several investigations, GIST has been used successfully in order to improve virtual screening results^{81,94}. However, care must be taken since GIST results are usually limited to a single or only few conformations of a solute molecule. Thus, multiple GIST calculations with different solute conformations must be carried out in order to obtain a reasonable estimate for the different solvation properties.

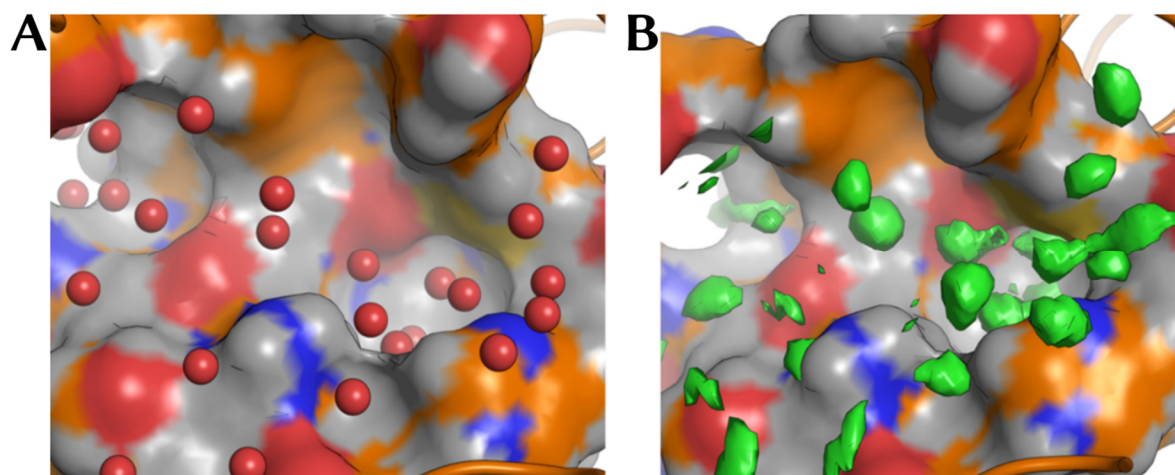


Figure 1-3: **A:** Hydration sites in the binding pocket of Caspase 3. **B:** Water occupancy map displayed at three times bulk solvent density as obtained from a GIST calculation. The figure is extracted from Haider et al.⁹⁵

2 Protein-Ligand Complex Solvation Thermodynamics: Development, Parameterization and Testing of GIST-based Solvent Functionals

2.1 Abstract

We present a set of solvent functional-based models which calculate the binding contributions resulting from solvation free energy, enthalpy and entropy for a set of 53 thrombin ligands. Our solvent functionals are based on molecular dynamics simulations in conjunction with GIST processing and are calibrated using accurate experimental data from ITC measurements. We found, that excellent agreement with experimentally derived enthalpy-entropy factorization can be achieved by considering both the solvation thermodynamics of the protein-ligand complex as well as the desolvation of the ligand molecule in solution. We demonstrate, that the desolvation free energy of the ligand drives the actual binding process, whereas contributions from the protein-ligand complex are necessary for the discrimination between individual ligands.

2.2 Introduction

The importance of molecular solvation and desolvation is undoubtedly highly appreciated in the field of drug design^{47,48,96}, but at the same time it has always been debated quite controversially^{97,98}. Over the last years, water molecules became recognized as an active contributor to protein-ligand binding and contributed to our understanding of molecular recognition⁵⁰, allosteric regulation⁹⁹ or preorganization phenomena.^{43,100} This gain in understanding is mainly due to important advancements in computational techniques such as WaterMap⁷⁹, SZMAP^{84,85}, GIST¹⁰¹, JAWS⁷⁷, GCMC⁷⁴ or SPAM¹⁰². Also, the impressive improvement of high-resolution crystallography in routine application of drug design projects using synchrotron radiation enhanced our current structure-based understanding of biomolecular solvation as the basis for binding thermodynamics. However, at the moment it is by far not straightforward to integrate solvation features into rationally derived Structure-Affinity-Relationships (SAR), although some early attempts with GIST already have been proven to be promising.⁹⁴ One of the main obstacles is the difficulty how to partition the overall binding free energy into contributions that solely come from interactions of solvent molecules, solute molecules or mixtures thereof. During a Structure-Based Drug Discovery (SBDD) campaign, it is crucial to know the precise location of water molecules in order to predict the next candidate ligand molecule with optimized binding properties. For that, it is necessary to characterize the water structure of all end-states during the ligand-binding reaction. However, the water-structure of the unbound ligand usually is not known *a priori*, which complicates SBDD in so far as it is not known if a potential water molecule in the protein-ligand bound structure is picked up during the binding process or if it is already bound to the ligand molecule in solution. In the first case, the water molecule might have a stronger impact on binding affinity than in the latter case. The pre-bound state of the ligand molecule in the bulk solvent phase is usually not investigated, although there are clear indications that this state can give the predominant contribution to the thermodynamic binding profiles.^{43,100} In the current work, we make use of the water structure from all end-states of the ligand binding reactions.

Usually, the configuration space of solvent molecules is strongly coupled to the configuration of the solute molecules, however this dependency decreases with increasing distance from the solute surface. With spatially resolved end-state approaches to solvation thermodynamics like Grid Inhomogeneous Solvation Theory (GIST), one can make use of this assumption by only considering the thermodynamics of the solvent molecules in proximity to the solvent surface. This simplifies the problem drastically. In this study, we will also make use of GIST and will

use this method for the construction of a rational SAR, based solely on solvent contributions. These solvent contributions are rationalized by building different physically motivated models which use data from molecular dynamics (MD) simulations and associated GIST calculations in order to predict solvation free energy and enthalpy (energy). The functional form of all these models was based on the previously described displaced-solvent functional.^{78,82} The term *solvent functional* simply refers to a mathematical formulation of a function that uses the three-dimensional distribution of solvent energy, entropy and density as independent variables. This functional employs different (initially unknown) parameters that are required to transform these distributions into scalar values for solvent free energy, energy or entropy. As an enhancement to the literature-described displaced-solvent functional based on GIST,⁸² we suggest novel displaced-solvent functionals that require fewer parameters than the original one while at the same time not compromising predictive power.

We selected a highly congeneric series of 53 ligands binding to thrombin for which high-resolution crystal structures are available and for which free energy, enthalpy and entropy data were determined by ITC and SPR (48 with ITC data and 5 with SPR data).^{49,103–111} This series of thrombin binders was sorted into matching pairs, such that the affinity difference between ligands within a given pair can predominantly be attributed to a difference in solvation. The resulting 186 pairs are used to further parameterization and testing of our solvent functionals. In the following, the term *GIST data* will be used in order to refer to the general combination of solvent energy, entropy and density distributions obtained from GIST calculations. Also, note that we will use the term “energy” when referring to computed energies and the term “enthalpy” when referring to experimentally determined enthalpies. Throughout this work, we will compare calculated energy values and measured enthalpy values with each other. However, it must be noted that they do not strictly correspond to the same physical quantity, since enthalpy includes a contribution from pressure-volume work. This term is usually negligible in condensed phase systems and therefore enthalpy can be well approximated by energy.

In the first part of this work, we will introduce the new solvent functionals and how they are applied to the different states (protein-ligand complex, ligand in aqueous solution) of the system. In the second part, we apply these solvent functionals in order to build models that (1) calculate solvation free energies based only on protein pocket desolvation (the already established displaced-solvent formalism), (2) calculate free energies based on both, the protein-ligand complex and the ligand molecule alone (full binding-displacement treatment), and (3) calculate free energies with optimized solvation enthalpies.

2.3 Materials and Methods

As a prerequisite of this study, we carried out MD simulations for the *apo* protein, each protein-ligand complex and each unbound ligand in solution. From these simulations, we calculated solvent energy, entropy and density using the GIST^{79,80,82} method and developed, parameterized and tested different solvent functionals. These solvent functionals calculate the solvation portion of the free energy, energy and entropy of protein-ligand association processes. The solute atoms in each MD simulation must be restrained to a reference structure. However, this positional fixation diminishes the influence of protein flexibility on the solvation thermodynamics. To cope with this, we assume that the effect of flexibility is most important for the *apo* protein, since in a protein-ligand complex, atoms become more firmly fixated due to interactions between the ligand and the protein. Therefore, we carried out unrestrained MD simulations of the *apo* protein and split the conformations observed along the trajectory into clusters. For the most representative structure from each cluster, MD simulations with positional restraints were carried out in triplicates and subsequently used as input for our GIST calculations. For the protein-ligand complexes as well as the unbound ligand molecules, only fully restrained MD simulations were carried out, keeping the complex spatially fixed to the conformation found in the crystal structure. They served directly as input for GIST. The complete workflow is outlined in Figure 2-1.

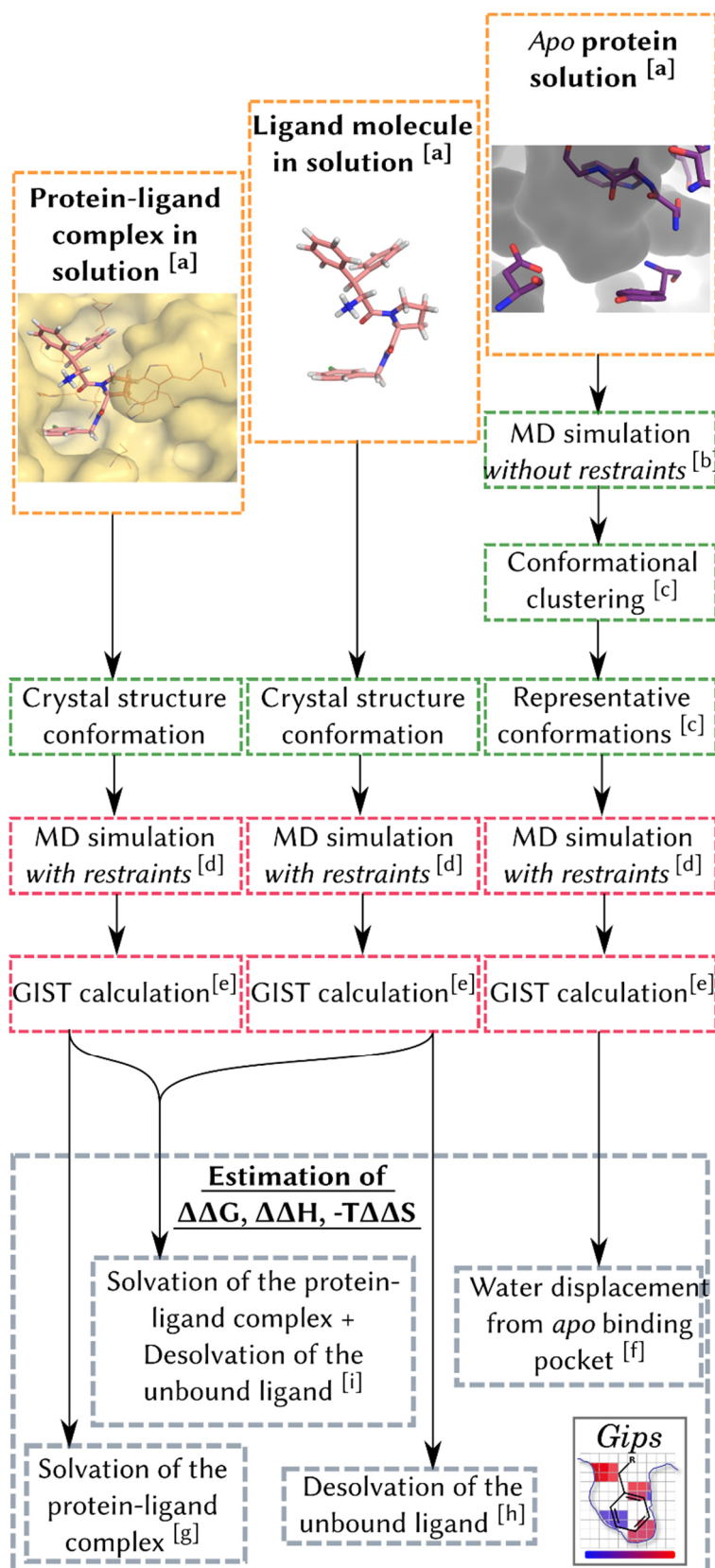


Figure 2-1: Overview of the workflow used in this study. The lower-case letters [a]-[h] refer to specific steps as referenced in the main text.

2.3.1 Structure Preparation

The pdb accession codes for the thrombin *apo* structure¹¹² as well as the ligand-bound thrombin structures^{49,103–111} are listed in the Supporting Information. All structures were prepared (curating for missing sidechains, assigning protonation states) using the structure preparation utility implemented in MOE¹¹³. For the ligand partial charge calculation, we initially decomposed the ligand molecules into amino-acid moieties with acetyl, N-methyl, N-dimethyl and methylsulfonate (-SO₂CH₃) capping groups (Figure 2-1, step [a]). This choice of capping groups is justified by the fact that all ligand molecules (see Figure 2-1, step [b] for a representative example or the Supporting Information for a complete list) contain amide and sulfonamide linker groups. Then, we performed a multimolecule and multiconformational *RESP* (restrained electrostatic potential) fitting based on these amino-acid moieties.^{55,114} The ESP (electrostatic potential) of these were obtained from the HF/6-31G* level of theory (b3lyp/6-31G* structure optimization) calculated using Gaussian09.¹¹⁵ Then, GAFF force-field parameters¹¹⁶ were assigned to the ligand molecules and missing force-field parameters were assigned using *antechamber* and *parmchk2* from the AmberTools17 package.¹¹⁷ The protein, ligand, structurally bound sodium ions as well as the water molecules from the crystal structure were combined and assigned force-field parameters using *tLEaP*. For the protein, we used the Amber FF14SB⁵⁷ force field together with the TIP4P-Ew water model.¹¹⁸ The system was embedded in a truncated octahedron simulation box filled with water molecules. The box was build such that the minimum distance between each solute and crystallographically determined water molecule and any box edge was at minimum 16 Å. The systems were neutralized by placing chloride counter-ions at random positions in the bulk water phase of the simulation boxes using the *addIonsRand* utility of *tLEaP*. After creating the simulation boxes and saving the parameter and starting structure files to disk, we randomly stripped off water molecules from the bulk phase (ca. 1% of all water molecules), such that all systems contained exactly the same total number of water molecules (13348). The complete building procedure was repeated for each of the three replicates per system (i.e. each system contained different positions of counter ions and initial water configurations).

The procedure was repeated analogously for the building of the simulation boxes of the *apo* structure with the same total number of water molecules as in the protein-ligand complexes. The simulation boxes of the ligand molecules were prepared analogously, however with a total number of 3500 water molecules for each ligand simulation box.

2.3.2 Molecular Dynamics Simulations of the *apo* Protein

For the *apo* protein, we initially performed MD simulations in order to obtain multiple representative conformations of the *apo* binding pocket. These conformations were used as starting structures for the latter GIST analysis (see [b] in Figure 2-1).

We performed an energy minimization (250 steps of steepest descent, 250 steps of conjugate-gradient optimization) of the system, keeping the solute atoms fixed at their crystallographic positions using a harmonic potential with a force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. After the first, a second minimization was carried out, using a force constant of only $5 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. In the next step, the system was kept harmonically restrained using a force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$ and heated to 300 K within 25 ps using an integration time-step of 1 fs. At this temperature, the system was equilibrated to a target pressure of 1 bar in an NPT ensemble within 100 ps using the Berendsen barostat¹¹⁹. During this NPT run, the positional restraints were removed gradually and the integration time step was switched to 2 fs. A final 1 ns equilibration run was carried out in the NVT ensemble. Triplicate production MD runs were carried out for 600 ns and coordinates were saved to disk every 10 ps.

During all runs, periodic boundary conditions were applied using the particle-mesh Ewald method with a real-space cutoff of 9 Å. We used the Langevin dynamics thermostat with a collision frequency $\gamma = 2 \text{ ps}^{-1}$ and different random seeds for each run. During all molecular dynamics runs, we applied the SHAKE¹¹⁹ algorithm to all bonds involving hydrogen atoms. We used *pmemd* and its GPU implementation *pmemd.cuda*¹²⁰⁻¹²² from the Amber16 package for energy minimization and molecular dynamics runs.¹¹⁷

2.3.3 Conformational Clustering along the Trajectory of the *apo* Protein

The conformations in the combined trajectories of the *apo* protein were clustered (see c) in Figure 2-1) based on RMSD using the average linkage clustering implementation of *cpptraj* (V17)¹²³. Only every 10th frame of each trajectory was included in the clustering using the *sievetoframe* utility from the clustering routine of *cpptraj*. Clustering was based on the non-hydrogen atoms of the following binding-site residues: D234, S235, V255, S256, W257, G258, E259, G260, C261, Y267, G268, F269, Y270, H73, Y77, W80, W122, E124, L125, L126, I209, D229, A230, C231, E232, G233. These protein binding site residues were selected, since they bear at least one atom within 4 Å of the ligand in the protein-ligand complex of PDB-code 3RML. Their choice of this ligand was arbitrary, but represents the binding pose of all ligands in the dataset reasonably well. The combined *apo* trajectories were found to be well described

by three clusters (see Supporting Information for plots showing Davies-Bouldin Index and pseudo F-statistic for different clustering solutions).

2.3.4 Molecular Dynamics Simulations for GIST Analysis

In this step, MD simulations were carried out (step [d] in Figure 2-1) in order to sample water configurations that later could be processed with the GIST approach.

Initially, the system energy was minimized using 2500 steps of steepest descent and 2500 steps of conjugate gradient minimization, while keeping the non-hydrogen atoms of the solute harmonically restrained to their starting positions with a force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. In a second minimization run, again 2500 steps of steepest descent and 2500 steps of conjugate gradient minimization were carried out with a weaker force constant of $2 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. Then, the system was heated to 300 K within 25 ps using an integration time-step of 1 fs and positional restraints with a force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. At a temperature of 300 K, the system was equilibrated to a target pressure of 1 bar using the Berendsen barostat¹¹⁹ within 5 ns. In a final equilibration run, the system was simulated for 5 ns in the NVT ensemble. Triplicate production MD runs were carried out for 50 ns each. The coordinates of the system were saved to disk every 2 ps.

For the simulations with positional restraints on the *apo* protein, only those non-hydrogen atoms that were considered in the clustering procedure (i.e. the binding site, see also step [c]), were also restrained during the complete minimization, equilibration and simulation procedure. All other non-hydrogen atoms were allowed to move freely. For the protein-ligand complexes and the separated ligand molecules in the water phase, all non-hydrogen atoms were considered for the restraining procedure.

For all energy minimization and MD runs, the same periodic boundary condition, thermostat and SHAKE settings were used as described for the unrestrained MD simulations of the *apo* protein.

2.3.5 Post-Processing of Trajectories and GIST Calculations

All molecular dynamics trajectories with positional restraints on the solute atoms were post-processed with the GIST^{79,101} routine (step [e] in Figure 2-1) as implemented in *cpptraj* (V17)¹²³. The GIST grids for each trajectory had 100x100x100 grid voxels with 0.5 x 0.5 x 0.5 Å side lengths per grid cell. The grid box was placed at the center of geometry of the ligand molecules in the case of protein-ligand complexes and ligand molecules in solution. For the *apo*

protein, the center of geometry defined by the amino acids D234, S235, V255, S256, W257, G258, E259, G260, C261, Y267, G268, F269, Y270, H73, Y77, W80, W122, E124, L125, L126, I209, D229, A230, C231, E232 and G233 was used as the center of the grid box. These residues were selected according to the same criteria as in step [c].

The 100x100x100 grid was evenly split into eight smaller grids of 52x52x52 (including one additional grid voxel in each dimension to account for missing entropy calculations in the outmost layer of grid voxels) with the *SplitVolume.py* python script.¹²⁴ This allowed us to effectively carry out the complete GIST calculation by means of eight significantly reduced “small” GIST calculations. The processed eight small GIST grids were finally combined back into the original GIST grid by using the data parsing routines of our *Gips* program. For visualization purposes, the GIST maps were processed with *gistpp*^{80,125} and load into *PyMOL*.^{126,127}

2.3.6 MM-GBSA and MM-3DRISM Calculations

The MM-GBSA¹²⁸ and MM-3DRISM¹²⁹ calculations were carried out using the *mmpbsa.py*¹²⁸ program from the AmberTools17 package. We processed frames extracted every 100ps from the trajectories that were generated for the processing with GIST. For the GBSA calculations, we used the *Onufriev, Bashford, Case* (OBC) variant with modified α, β and γ together with mbondi2 radii (igb = 5 option in sander).^{130,131} For the 3D-RISM calculations, we used the Gaussian Fluctuation^{132,133} approximation.

2.4 Theoretical Background

We will initially introduce three different so-called basic solvent functionals (termed F4, F5 and F6 according to the number of parameters), which bear resemblance to solvent functionals from the early work on WaterMap.⁷⁸ After that, we will explain how state-specific parameter settings are applied to the basic solvent functionals, in order to model solvent free energy and solvent energy at once. In the last part of this section, we will introduce the concept of global and state-specific parameter settings.

Generally, the solvent functionals use the raw solvent entropy, energy and density data from our GIST calculations of the *apo* protein (step [f] in Figure 2-1), the complex (step [g]), the ligand in solution (step [h]) or the combination of protein-ligand complex and unbound ligand (step [i]) as input data. From the input GIST data, the solvent functionals calculate solvent free

energy, entropy and energy for a solute species. We used different solvent functionals with varying combinations of global and state-specific parameter settings. A set of optimal parameters for the different solvent functionals was obtained by training with experimentally determined protein-ligand binding thermodynamics. In a first attempt, we trained our solvent functionals only with free energy data as it was commonly undertaken in similar approaches.^{78,82} However, within this approach the explicit entropy and energy terms in the solvent functionals can (freely) compensate/reinforce each other unless further boundary conditions are installed and consequently do not necessarily represent the actual enthalpy-entropy factorization, as obtained from experiment. Therefore, we also fitted the energy term in the solvent functionals to experimental enthalpy data, in addition to the fitting of free energy, in order to explicitly account for enthalpy-entropy factorization. Depending on the state of a protein-ligand binding reaction (i.e. the protein-ligand complex state or the unbound state of the ligand) that the GIST data are derived from, each state can have the same or specific set of parameters or individual parameters. When each state is described by the same set of parameters, the latter will be referred to as “global parameter setting”. When each state is described by individual parameters, then the latter will be referred to as “state-specific parameter setting”.

The solvent functionals for the calculation of solvent free energy, energy and entropy were constructed from the well-known Helmholtz free energy equation:

$$\Delta G_{Solv}^0 = \Delta H_{Solv}^0 - T\Delta S_{Solv}^0 \approx \Delta E_{Solv}^0 - T\Delta S_{Solv}^0 \quad (2-1)$$

Here, ΔG_{Solv}^0 , ΔH_{Solv}^0 and ΔS_{Solv}^0 are the standard solvation free energy, solvation enthalpy and solvation entropy of binding, respectively. The solvation enthalpy term is approximated by the solvation energy, as contributions from pressure-volume work are negligible. As GIST is a grid-based approach, each of the aforementioned quantities is calculated separately for each individual grid voxel k . Consequently, the solvation energy value of a single grid voxel k is calculated as

$$\Delta E_{Solv}(\vec{r}_k) = E_{SW}(\vec{r}_k) + 2\left(E_{WW}(\vec{r}_k) - E_{WW}^{(bulk)}\right) \quad (2-2)$$

In eq. (2-2), the solvation energy term, $\Delta E_{Solv}(\vec{r}_k)$, is the sum of the water-water interaction energy, $E_{WW}(\vec{r}_k)$, referenced to water-water interaction energy in bulk water, $E_{WW}^{(bulk)}$, and the solute-water interaction energy, $E_{SW}(\vec{r}_k)$. The factor of 2 accounts for the fact that by

convention GIST water-water energies are calculated as one-half the mean interaction energy of the water molecules in voxel k with all other water molecules in the system.⁸² The water-water interaction energy in bulk water is specific for the water model and calculated in separate MD simulations containing only water molecules. All energy terms are effectively calculated from the molecular mechanics force field of the MD simulation.

The solvation entropy term is approximated by the one-body translational and orientational entropy contributions of voxel k , calculated as

$$-T\Delta S_{Solv}(\vec{r}_k) = -T\Delta S_{trans}(\vec{r}_k) - T\Delta S_{orient}(\vec{r}_k) \quad (2-3)$$

The derivation of those contributions will not be repeated here, since it has already been comprehensively introduced in the work by *Lazaridis et al.*^{89,91} and *Kurtzman et al.*⁷⁹. Note, that we will use the term entropy, ΔS_{Solv} , in order to refer to the entropic contribution to free energy, $T\Delta S_{Solv}$, which is effectively the entropy scaled by the thermodynamic temperature of the system. Throughout this work, all systems are treated at 300 K and the entropic contributions to free energy are all calculated at this temperature.

In GIST, the thermodynamic quantities are essentially obtained by spatially integrating over probability, entropy and energy distributions. However, the direct integration of these distributions is problematic, since some regions do not contribute to binding and others suffer from high level of noise in energy and entropy (particular regions with a low occupancy of water molecules). Therefore, the individual grids are not integrated directly, but coupled to a filtering mechanism, which uses a simple step-function formalism as will be explained in the following.

2.4.1 The F4 Solvent Functional

The basic solvent functional employed a set of four parameters for the calculation of solvation energy and entropy contributions:

$$\Delta E_{solv}^{(F4)} = \rho^0 \sum_k^G V_k g(\vec{r}_k) \Delta E_{solv}(\vec{r}_k) e_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-4)$$

$$-T\Delta S_{solv}^{(F4)} = -\rho^0 \sum_k^G V_k g(\vec{r}_k) T\Delta S_{solv}(\vec{r}_k) s_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-5)$$

$$e_s(\vec{r}_k) = \begin{cases} 1, & \text{if } \Delta E_{solv}(\vec{r}_k) > e_{co} \\ 0, & \text{otherwise} \end{cases} \quad (2-6)$$

$$g_s(\vec{r}_k) = \begin{cases} 1, & \text{if } g(\vec{r}_k) > g_{co} \\ 0, & \text{otherwise} \end{cases} \quad (2-7)$$

$$s_s(\vec{r}_k) = \begin{cases} 1, & \text{if } -T\Delta S_{solv}(\vec{r}_k) > s_{co} \\ 0, & \text{otherwise} \end{cases} \quad (2-8)$$

Here, ρ^0 is the one-body solvent density of pure water at 25°C and 1 bar for the water model in use (in this work, TIP4P-Ew¹¹⁸) in units of \AA^{-3} and V_k is the volume of grid voxel k . The quantities $\Delta E_{solv}(\vec{r}_k)$, $\Delta S_{solv}(\vec{r}_k)$ and $g(\vec{r}_k)$ represent the solvent energy, entropy and density values at grid point k located at \vec{r}_k on the grid G . The solvent density, $g(\vec{r}_k)$, is calculated as the average number of water molecules in grid voxel k , normalized to the average number of water molecules in the same volume in pure bulk water. Consequently, the solvent density at k is given in multiples of bulk water density, ρ^0 (0.0332 \AA^{-3} for the TIP4P-Ew water model used in this study). The thermodynamic quantities solvent energy and entropy referenced herein, are all normalized to the average number of water molecules in the respective grid voxel (for a discussion about normalized energies and entropies, please refer to reference ⁷⁹). In eq (2-4), $e_s(\vec{r}_k)$ and $g_s(\vec{r}_k)$ are the solvent energy and density step functions, which evaluate to 1 if the solvent energy or density at \vec{r}_k exceed the cutoff value e_{co} or g_{co} , respectively (see eqs. (2-6) and (2-7), respectively). The entropy term, ΔS_{solv} , uses the entropy step function $s_s(\vec{r}_k)$, which evaluates to 1 if the entropy at \vec{r}_k exceeds the cutoff value s_{co} and evaluates to 0 otherwise. The

volume step function, $v_s(\vec{r}_k)$, evaluates to 1 if grid point \vec{r}_k is inside the molecular volume of the ligand molecule and goes to 0 if grid point k is outside the molecular volume (these step functions were already explained elsewhere⁸²). In other words, it explicitly defines the boundaries for the volume integration of the GIST data. In this work, we could not use the molecular volume of the ligand molecules directly, since for ligand-bound structures or the free ligand in solution, the molecular volume of the ligand would be part of the solvent-excluded volume and therefore not account for GIST-based solvation contributions. For this reason, we used a volume that included the molecular volume of the ligand molecule plus its first solvation shell, which includes all water molecules within a distance of 3\AA from the molecular surface of the ligand molecule (see also below, *Soft Solvent Surfaces*). This distance was used for the calculation of the primary solvation layer volume for all ligands. Although this is not strictly accurate, since the thickness of a layer of water molecules varies depending on the roughness of the surface and the conformation of the solute molecule.

2.4.2 The F6 Solvent Functional

Instead of directly using the value of the energy, $\Delta E_{solv}(\vec{r}_k)$, or entropy, $\Delta S_{solv}(\vec{r}_k)$, quantities at grid point k , weighting parameters for energy, E_{aff} , and entropy, S_{aff} , are introduced. With these weighting parameters, an effective energy and entropy contribution to binding affinity is assigned to each grid value that exceeds the cutoff criteria introduced above. This approach was already introduced in prior work⁸² and is formulated as follows:

$$\Delta E_{solv}^{(F6)} = E_{aff} \sum_k^G e_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-9)$$

$$-T\Delta S_{solv}^{(F6)} = -S_{aff} \sum_k^G s_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-10)$$

Equations (2-9) and (2-10) are analogously formulated to eqs. (2-4) and (2-5), but contain the scalar weighting parameters E_{aff} and S_{aff} instead of the actual grid quantities $\Delta E_{solv}(\vec{r}_k)$ and $\Delta S_{solv}(\vec{r}_k)$. Since the weighting parameters are not known *a priori*, they must be obtained during a parameter optimization process.

2.4.3 The F5 Solvent Functional

In another attempt, we simplified eqs. (2-9) and (2-10), by using only a single general weighting parameter K_{aff} , for both energy and entropy:

$$\Delta E_{solv}^{(F5)} = K_{aff} \sum_k^G e_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-11)$$

$$-T\Delta S_{solv}^{(F5)} = -K_{aff} \sum_k^G s_s(\vec{r}_k) g_s(\vec{r}_k) v_s(\vec{r}_k) \quad (2-12)$$

2.4.4 Soft Solvent Surfaces

If the spatial integration volume of the ligand molecule cuts through a region highly occupied by water molecules, which also contributes a notable fraction to the total solvation thermodynamics, the final value of $\Delta E_{solv}(\vec{r}_k)$ or $\Delta S_{solv}(\vec{r}_k)$ depends on only few grid points. Depending on whether these grid points are included in the evaluated region, i.e. whether v_s evaluates to 1 or not, they might have a large impact on the final result. In such cases, the value of v_s would depend on the relative position and orientation of the grid used for evaluation with respect to the ligand. Of course, one could counteract this by using increasingly smaller grid voxels, but that would have a negative impact on the convergence of the GIST quantities. Therefore, we used *soft surfaces*, which replace the strict classification of “inside” or “outside” a given surface volume with a fuzzy classification which allows for an attenuation of v_s in the vicinity of the surface. For that, we scaled v_s with a distance-dependent exponential function:

$$v_s(\vec{r}_k) = \begin{cases} 1 & \text{if } d(\vec{r}_k) < 0 \\ \exp\left(-\frac{d(\vec{r}_k)}{s}\right) & \text{if } 0 < d(\vec{r}_k) < c \\ 0 & \text{if } d(\vec{r}_k) > c \end{cases} \quad (2-13)$$

$$d(\vec{r}_k) = \min_{a \in M} (|\vec{r}_k - \vec{r}_a| - R_a) \quad (2-14)$$

In eq. (2-13), $d(\vec{r}_k)$ is the distance between grid point k and the surface of the molecule and s is a softness parameter that controls how strong $v_s(\vec{r}_k)$ decays in the vicinity of the surface (see Figure 2-2 for a graphical representation). The softness cutoff parameter c determines how far the softness of the surface reaches into the remaining part of the grid. We found empirically that values of $s=1$ and $c=2$ give reasonable results. The distance $d(\vec{r}_k)$ is calculated using eq. (2-14)

which gives the distance of grid point k and its closest atom a with radius R_a in molecule M . Negative values of $d(\vec{r}_k)$ indicate that the grid point is inside the occupied volume of the molecule, whereas positive values indicate that the grid point is outside the volume.

Alternatively, one could also use Gaussian distributions centered on each grid voxel, as commonly used in molecular interaction field analysis.⁷⁰ Another common strategy to prevent overly dominant contributions from only few grid points is to smooth the three-dimensional distribution by assigning the average over all neighboring grid values to each grid point.

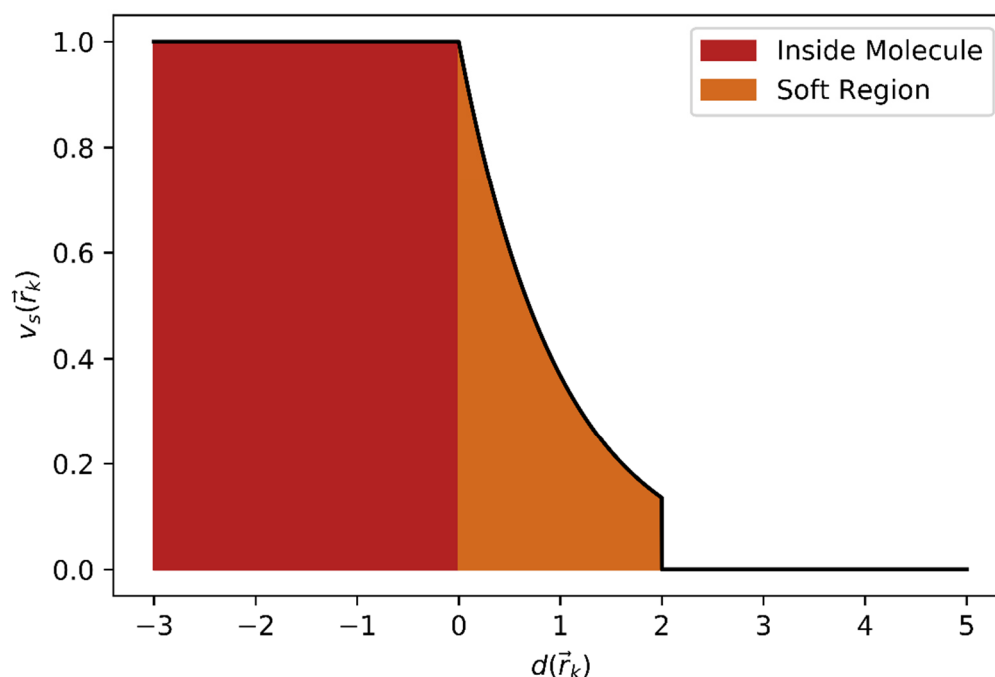


Figure 2-2: Illustrating the soft surface approach. The red region is inside the molecule and the volume indicator function evaluates to one. The soft region starts at the surface of the molecule (where $d(\vec{r}_k)$ becomes zero) and exponentially decreases up to the cutoff c . Beyond the cutoff, the volume indicator function is set to zero.

2.4.5 Displaced-Solvent Functionals

The use of displaced-solvent functionals was pioneered in the work of Abel et al.⁷⁸, using hydration sites, and later by Kurtzman et al.⁸², using GIST, in order to calculate the desolvation free energy contribution of the protein binding site. In our work, we also used this rational approach in order to correlate the desolvation of the protein binding site with relative free energies of matched ligand pairs. In addition, we also separately tested how well binding affinity can be correlated with ligand desolvation or protein-ligand complex solvation contributions. We will generally refer to a solvent functional with the nomenclature S/F, where S indicates the state (PL for protein-ligand complex, L for ligand in aqueous solution, P for apo

protein) of the system that was used for generating the GIST data. F indicates the basic solvent functional (F4 to F6) that was applied to the GIST data. For instance, the solvent functional P/F6 is the solvent functional that was used in prior work of Kurtzman et al.⁸² If GIST data from two states, (bound and unbound) were employed in the GIST functional, we will refer to it as S1-S2/F. Here, S1 and S2 refer to the initial and final state, respectively (e.g. S1: PL and S2: L) and the thermodynamic quantities are calculated from the differences between those states accordingly, e.g. $\Delta G(S1 - S2) = \Delta G(S1) - \Delta G(S2)$.

It is worth noting that the raw GIST data actually reports quantities that describe the process of solvation in the direction of actual solvation (i.e. solute-solvent association) and not desolvation (i.e. solute-solvent dissociation). Although the direction of the process is allowed to vary freely for the P/F5, P/F6, L/F5 and L/F6 functionals by not explicitly imposing a sign on the weighting parameters ($E_{\text{aff}}/S_{\text{aff}}$, and K_{aff}), a constant sign must be set for the solvent functionals involving P/F4 and L/F4. As can be seen from eqs. (2-4)-(2-5), basic functional F4 is calculated from the actual grid values, therefore these values must reflect the direction of the process (solvation vs. desolvation) in order to give a physically correct representation. For this reason, the signs of $\Delta E_{\text{Solv}}^{(F4)}$ and $-T\Delta S_{\text{Solv}}^{(F4)}$ (see eqs. (2-4), (2-5), respectively) were inverted for solvent functionals P/F4 and L/F4. This is necessary since all solvent functionals of type S/F, with S=P, L are based on protein-pocket desolvation or ligand desolvation, respectively. Also, it must be noted that the signs of $\Delta E_{\text{Solv}}(\vec{r}_k)$ and $-T\Delta S_{\text{Solv}}(\vec{r}_k)$ in eqs. (2-4),(2-5) were not inverted at all.

In the case of P/F4, P/F5 and P/F6 GIST data from multiple conformations of the protein were considered (as outlined in the MD protocol above). Therefore, the GIST data from these individual conformations were weighted according to their populations (calculated from the clustering).

2.4.6 Enhancing the Solvent Functionals by Employing State-Specific Parameter Settings

In order to let the solvent functional capture more of the different solvent distributions in the binding pocket and in the unbound state, we developed different schemes in which the protein-ligand complex state was assigned to different cutoff parameters in comparison to the unbound state of the ligand. In these so-called state-specific parameter settings, the parameters still have the same physical meaning in all states, which allows for their comparison across the different states. The general form of solvent functionals that employ either state-specific or global parameter settings is S1-S2/F/R, where S1 and S2 are different states, F is the basic functional

and R is the parameter setting. For instance, $R = \{g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(g)}, s_{CO}^{(g)}\}$ defines a set of parameters with individual solvent density cutoff parameters, $g_{CO}^{(PL)}, g_{CO}^{(L)}$, for the protein-ligand complex and ligand in aqueous solution, but uses global (i.e. the same in all states) energy and entropy cutoff parameters, $e_{CO}^{(g)}, s_{CO}^{(g)}$. The rationale behind this idea is that a high solvent density cutoff for the protein-ligand complex is necessary in order to identify the highly populated solvent regions in the binding pocket that actually contribute to solvation free energy and those that do not contribute. In the unbound state, water molecules are bound less tightly to the ligand's surface than in the binding pocket and a much lower solvent density cutoff must be applied. Therefore, a solvent density cutoff that is empirically adjusted to both, the protein-ligand complex and the unbound state, will not adequately represent either of them.

Furthermore, we tested a solvent functional which used individual parameters for each of the three cut-off values. In that case, the parameter settings are defined as $R = \{g_{CO}^{(PL)}, e_{CO}^{(PL)}, s_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(L)}, s_{CO}^{(L)}\}$. Note that the weighting parameters, K_{aff} (basic solvent functional F5) and E_{aff}, S_{aff} (basic solvent functional F6) are always global parameters and therefore identical for all states.

Table 2-1: List of all solvent functionals that were tested and parameterized in this work.

Solvent Functional	F ^{a)}	S1 ^{b)}	S2 ^{b)}	R Global parameters ^{c)}	R State- specific S1 ^{d)}	R State- specific S2 ^{d)}
P/F4	F4	P	—	n.a.	n.a.	n.a.
P/F5	F5	P	—	n.a.	n.a.	n.a.
P/F6	F6	P	—	n.a.	n.a.	n.a.
PL/F4	F4	PL	—	n.a.	n.a.	n.a.
PL/F5	F5	PL	—	n.a.	n.a.	n.a.
PL/F6	F6	PL	—	n.a.	n.a.	n.a.
L/F4	F4	L	—	n.a.	n.a.	n.a.
L/F5	F5	L	—	n.a.	n.a.	n.a.
L/F6	F6	L	—	n.a.	n.a.	n.a.
PL-L/F4/ $\{g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}\}$	F4	PL	L	$g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}$	n.a.	n.a.
PL-L/F5/ $\{g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}\}$	F5	PL	L	$g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}$	n.a.	n.a.
PL-L/F6/ $\{g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}\}$	F6	PL	L	$g_{co}^{(g)}, e_{co}^{(g)}, s_{co}^{(g)}$	n.a.	n.a.
PL-L/F4/ $\left\{ \begin{matrix} g_{co}^{(PL)}, g_{co}^{(L)} \\ e_{co}^{(g)}, s_{co}^{(g)} \end{matrix} \right\}$	F4	PL	L	$\{e_{co}^{(g)}, s_{co}^{(g)}\}$	$\{g_{co}^{(PL)}, g_{co}^{(L)}\}$	n.a.
PL-L/F5/ $\left\{ \begin{matrix} g_{co}^{(PL)}, g_{co}^{(L)} \\ e_{co}^{(g)}, s_{co}^{(g)} \end{matrix} \right\}$	F5	PL	L	$\{e_{co}^{(g)}, s_{co}^{(g)}\}$	$\{g_{co}^{(PL)}, g_{co}^{(L)}\}$	n.a.
PL-L/F6/ $\left\{ \begin{matrix} g_{co}^{(PL)}, g_{co}^{(L)} \\ e_{co}^{(g)}, s_{co}^{(g)} \end{matrix} \right\}$	F6	PL	L	$\{e_{co}^{(g)}, s_{co}^{(g)}\}$	$\{g_{co}^{(PL)}, g_{co}^{(L)}\}$	n.a.
PL-L/F4/ $\left\{ \begin{matrix} g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)} \\ g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)} \end{matrix} \right\}$	F4	PL	L	n.a.	$\{g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)}\}$	$\{g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)}\}$
PL-L/F5/ $\left\{ \begin{matrix} g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)} \\ g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)} \end{matrix} \right\}$	F5	PL	L	n.a.	$\{g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)}\}$	$\{g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)}\}$
PL-L/F6/ $\left\{ \begin{matrix} g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)} \\ g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)} \end{matrix} \right\}$	F6	PL	L	n.a.	$\{g_{co}^{(PL)}, e_{co}^{(PL)}, s_{co}^{(PL)}\}$	$\{g_{co}^{(L)}, e_{co}^{(L)}, s_{co}^{(L)}\}$

- a) The basic solvent functionals used for this solvent functional. For the basic solvent functionals we applied eqs. (2-4),(2-5) for F4, eqs. (2-11),(2-12) for F5 and eqs. (2-9),(2-10) for F6.
- b) The states S1 and S2 can be either P (*apo* protein binding pocket), PL (protein-ligand complex) or L (ligand molecule in aqueous solution). For some functionals, S2 cannot be assigned due to the nature of the functional. In these cases, S2 is specified as “—”.
- c) Parameter settings for global parameters
- d) Parameter settings for the state- specific parameter settings of states S1 and S2.
- e) Parameter settings are designated as “not available” (n.a.), if the functional does not allow for a parameter setting of this type (global, S1 state-specific or S2 state-specific).

2.4.7 Objectives and Parameter Optimization

In order to adjust optimal sets of parameters for the different basic solvent functionals (represented by eqs. (2-4)-(2-5), (2-9)-(2-10) and (2-11)-(2-12)) that are employed in the solvent functionals, we constructed appropriate training and test datasets (see next section for details on the datasets). The parameters were obtained by fitting the solvent functionals to relative experimental free energies or to relative experimental free energies and enthalpies simultaneously. In the first case, the optimization problem is singular and the underlying objective is the squared sum of residuals of the free energy differences for all N pairs in the dataset:

$$\operatorname{argmin} \left(\sum_i^N \left(\Delta\Delta G_{\text{solv}}^{(X)}(AB^{(i)}) - \Delta\Delta G_{\text{solv}}^{(\text{exp})}(AB^{(i)}) + C_G \right)^2 \right) \quad (2-15)$$

In eq. (2-15), we included a constant C_G , which accounts for systematic deviations between calculated and experimental data. In order to solve the optimization problem depicted by eq. (2-15), we used the basin-hopping optimization strategy¹³⁴ as implemented in *PyGMO*¹³⁵. Briefly, basin-hopping is an optimization algorithm, which uses a Metropolis Monte Carlo search in parameter space in combination with a local minimization of the objective function, eq. (2-15). We used 2500 Monte Carlo steps in combination with SLSQP local minimization¹³⁶. The basin-hopping optimizer was allowed to stop earlier if no improvement of the objective function was found after 100 iterations. The local minimizer stopped when a minimization step changed no parameter by more than the 10^{-8} th of the respective parameter value at that step. A partial brute-force search, as carried out in a previous study⁸², was not applicable in our work, since each individual protein-ligand complex as well as each individual ligand in solution was associated with its own GIST dataset. The multitude of GIST data that is processed therefore leads to a drastic increase in computing time for a single evaluation of the solvent functional. The simultaneous optimization of energy and entropy is a multi-objective optimization problem and therefore requires a different optimization treatment. For this, we used *NSGA2*,¹³⁷ a genetic optimization algorithm, in order to simultaneously optimize free energy and enthalpy (energy). The two objectives consist of the minimization of the squared sum of residuals of free energy (eq. (2-15)) and energy (eq. (2-16)) for all N pairs:

$$\operatorname{argmin} \left(\sum_i^N \left(\Delta\Delta E_{\text{solv}}^{(X)}(AB^{(i)}) - \Delta\Delta E_{\text{solv}}^{(\text{exp})}(AB^{(i)}) + C_E \right)^2 \right) \quad (2-16)$$

Both eqs. (2-15) and (2-16) are coupled to each other, as they both contain the solvent energy. Both free energy and energy can have individual systematic deviations between calculated and experimental data, individual additive constants C_E and C_G were applied for the free energy and energy objective, respectively.

For the genetic optimization algorithm, we used 200 populations with a distribution index for crossover, η_c , and mutation, η_m , of 10. The populations were evolved over 2500 generations. After the optimization was complete, we saved the final parameters from each population and divided them into non-dominated fronts using the *fast_non_dominated_sorting* routine from the *PyGMO* library. The front with the lowest domination level was sorted using the *sort_population_mo* routine from the *PyGMO* library. From this front, we kept only the first solution per optimization attempt (note that a single best solution cannot be obtained in multiobjective optimization).

The final parameters for both single and double objective optimization were trained and tested from 10 random sets of 5-fold cross validation. We applied the same set of random splits in the optimization of every solvent functional. In order to further evaluate the model performance, we also trained the solvent functionals based on 10 random sets obtained by shuffling the dependent data (i.e. free energy and enthalpy (energy)).

During the optimization calculations, the parameters of the solvent functionals were allowed to vary freely within predefined boundaries. For the weighting parameters E_{aff} , S_{aff} (F6) and K_{aff} (F5) the allowed parameter range was $[-3,+3]$ kcal·mol⁻¹ and for the energy and entropy cutoff parameters it was $[-10,+10]$ kcal·mol⁻¹. Note, that as the entropy contribution already contains the thermodynamic temperature of the system (300 K in all cases, see also the section Theoretical Background), the unit of the entropy-dependent parameters is given as kcal·mol⁻¹. For the solvent density cutoff parameter we applied $[+1,+8]$ ρ^0 (with ρ^0 being the bulk solvent density, as introduced above) and for the constants C_G and C_E $[-3,+3]$ kcal·mol⁻¹. With these boundaries, we cover the density distribution in the binding site. The energy distribution is covered in the range of approximately ± 5 standard deviation units around the mean value (-0.04 kcal·mol⁻¹). In order to allow for similar coverage of the energy and entropy cutoff parameter space, we allowed both values to vary in the same boundaries. The boundaries are wide enough, such that the optimized parameters do not include much bias towards a specific parameter range.

2.4.8 The Thrombin Dataset

In this study, we investigate the contribution of solvation to binding affinity for a set of 53 ligands binding to the serine-protease thrombin for which crystal structures and experimental thermodynamic profiles (see Figure 2-3 for an overview) were obtained from different sources^{49,103-111} (see Supporting Information for a detailed list).

Since it is very difficult to distinguish between those contributions to binding affinity which come from solvent molecules and those which do not, it is most convenient to work with pairs of molecules. The pairs under investigation must be combined such that the overwhelming part of the difference in binding affinity between the paired molecules originates from small changes likely linked to a difference in the solvation pattern. In order to ascertain that the pairs are aligned with respect to this strategy, we applied several filtering criteria in order to find appropriate pairs among all possible pairs of the dataset. Specifically, the following filtering criteria were applied:

Charged Head Group: Many thrombin binders contain a positively charged P1 head group (since thrombin is a serine-protease, we apply the nomenclature of *Schechter and Berger*¹³⁸ in order to refer to the different portions of the ligand), which interacts with the charged D189 deeply buried in the S1 binding pocket.⁴⁹ This charge-charge interaction imposes an important feature to the binding properties of these compounds, which cannot be found similarly in ligands that do not bear this charged P1 head group. This could be problematic with respect to our solvent functionals for two reasons: 1) The underlying physical principle of the solvent functionals employed in this work does not explicitly account for charge-charge interactions of the solute. 2) The parameters of the solvent functional could be falsely trained considering particularly this charge-charge interaction. Therefore, only pairs in which both ligands bear a positively charged head group that interacts with D189 are regarded as a valid pair.

N-Terminus at the distal P3/4 site: Another common feature of the compounds in the dataset is the distal ligand portion, which binds to the S3 and S4 pockets of thrombin. Only in cases where the ligands contained a charged ammonium group (formally the N-terminus of the peptide-like ligand scaffold) at both molecules, they were handled as a valid pair.

Sulfonamide Group at the distal P3/4 site: The introduction of a sulfonamide linker between a glycine and terminal benzyl group gives rise to considerable preorganization of the ligands (formation of a β -turn-type conformation) in solution prior to binding (unpublished results). Since the imposed effect of preorganization is also not covered by the underlying physical

model applied here, we only considered pairs where both molecules had a sulfonamide moiety at this position as in that way preorganization effects will cancel out in the pairwise comparison. Total Charge: The total charge of the compounds dictates the total number of counter ions necessary to provide charge-neutral simulation boxes. Since desolvation of the counter ions is not considered in the solvent functional, we allowed only such pairs where both ligands had the same total charge.

This resulted in a total number of 253 pairs. We further reduced this dataset by eliminating all pairs that had a *Tanimoto* fingerprint similarity (based on the Daylight-like fingerprints as descriptors, calculated using *RDKit*¹³⁹) of less than 0.7. This last step effectively filtered out all pairs that differed in size and had chemically very dissimilar P1 and P3 portions. The dataset resulting from the last filtering step consisted of 186 pairs (see Supporting Information). This final set showed a broad distribution of free energies and enthalpies as outlined in Figure 2-3).

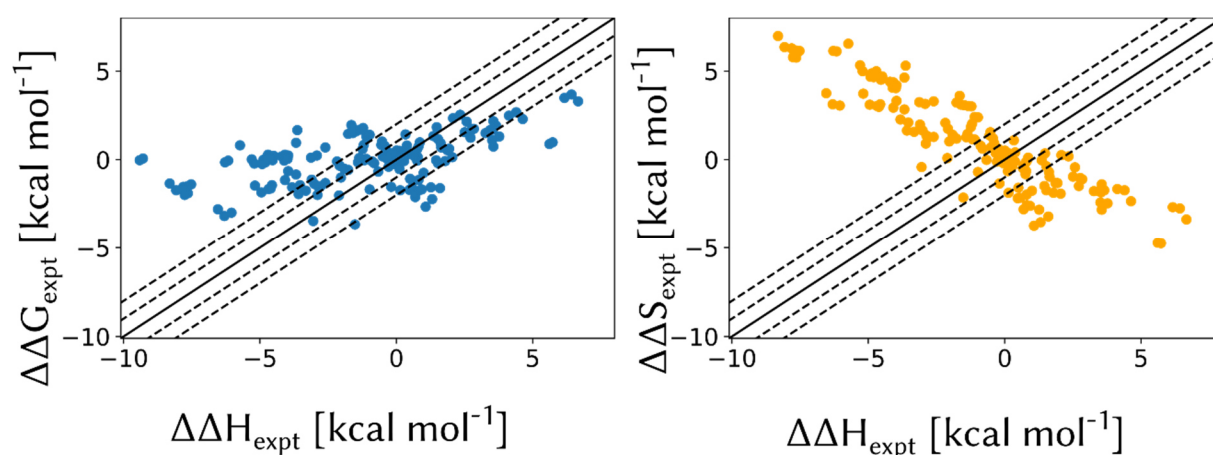


Figure 2-3. Plots showing the free energy versus energy (left) and entropy versus energy (right) distributions for all 186 matching ligand pairs. The solid black lines indicate zero difference in entropy (left) and zero difference in free energy (right). The dashed lines indicate ± 1 kcal·mol⁻¹ and ± 2 kcal·mol⁻¹ difference in entropy (left) and free energy (right).

2.5 Results and Discussion

In this section, solvation free energy calculations are presented based on the solvent functionals introduced above (see also Table 2-1). In the first part, we will present solvent functionals that calculate solvent free energies based on solvent molecules from a single state, S/F, this includes the so-called displaced-solvent approach. In order to assess the accuracy and predictivity of these S/F type solvent functionals, they will be trained and tested with experimentally determined binding free energy data. In the second part, we will use contributions from two different states, S1-S2/F/R, for the calculation of solvent free energies and energies. For the S1-S2/F/R type solvent functionals, we will use experimentally determined binding free energy in conjunction with binding enthalpy data in order to train and test the parameters in our solvent functionals. This approach will be called the ‘full binding-displacement approach’ and includes GIST data from the protein-ligand complex and the ligand. Hence, it effectively captures the displacement of the water molecules from the unbound state of the ligand and the solvent molecules picked up during binding. This approach will be necessary for accounting explicitly for experimentally observed solvent energy contributions alongside the experimentally observed free energies.

In this section we will justify the performance of the individual solvent functionals based on the correlation of the test data as obtained from five-fold cross validation. The corresponding performance of the solvent functionals with training data can be found in the Supporting Information.

2.5.1 Solvent Free Energy Calculated from a Single State Approach (S/F type)

Assessing Model Performance. Initially, we investigated the GIST solvation free energy obtained from either the uncomplexed protein binding pocket (*apo* form of the protein), the protein-ligand complex or the ligand alone in the bulk phase. In all cases, we used the same set of (randomly chosen) splits that divided the dataset into training and test data. The performance of the solvent functionals was evaluated using five-fold cross validation. We found that the ligand alone in aqueous solution gives GIST data that are best suited to establish a solvent functional that accurately predicts the binding free energy (see Figure 2-4A for an overview of the test set performance). The solvent functionals L/F4, L/F5 and L/F6 (Figure 2-4A, red) give the highest correlation and clearly perform better than similar functionals that were trained using shuffled data (see Figure 2-4C). For all solvent functionals, the mean unsigned error (MUE) is considerably low ($<0.5 \text{ kcal}\cdot\text{mol}^{-1}$, see Figure 2-4B). This demonstrates that our

solvent functionals exhibit excellent accuracy. However, care must be taken since the functionals, especially the ones obtained from the ligand molecules (L/F4 to L/F6), reveal low MUE for the shuffled datasets, as well (see Figure 2-4D). This phenomenon can possibly be explained by the spread of data of the considered dataset and an enhanced accumulation of data points in the $[-1,+1]$ kcal·mol⁻¹ region. For comparison, if one would use the mean of the experimental values from the training data sets in order to predict the test data sets, one would still have an MUE of 1.1 kcal·mol⁻¹. The solvent functionals based on GIST data from the protein binding pocket, P/F4 to P/F6 (Figure 2-4, blue), performed well but worse than the ones based on the ligands alone. From the ones based on the protein binding pocket, the F6 basic solvent functional performs best. This is the functional, which was also employed in previous work.⁸² Interestingly, all the solvent functionals derived with GIST data from the protein-ligand complex, PL/F4, PL/F5 and PL/F6 (Figure 2-4, green), performed worst amongst all the single-state displaced-solvent approaches. Unfortunately, they do not perform significantly worse using shuffled data. This result is somewhat surprising, since the considered high-resolution crystallographic data on protein-ligand complexes reflect the best available experimentally validated structural information about the water molecules. At the same time, it is also the most complex system, which probably suffers most from the physically artificial positional restraints used throughout the MD simulation applied to generate the GIST data. In the specific case of thrombin, the side chain of Glu192, located on top of the S1 binding pocket, demonstrates pronounced flexibility in most of the crystal structures, but as the simultaneously recorded ITC data show, the orientation of Glu192 is crucial for the affinity of the formed complex.¹⁰⁹ By restraining the protein to its crystallographic coordinates, the conformational flexibility of this amino acid is most likely not captured adequately. Consequently, the solvent thermodynamics may be biased towards a single conformation, which is not adequate for some protein-ligand complexes.

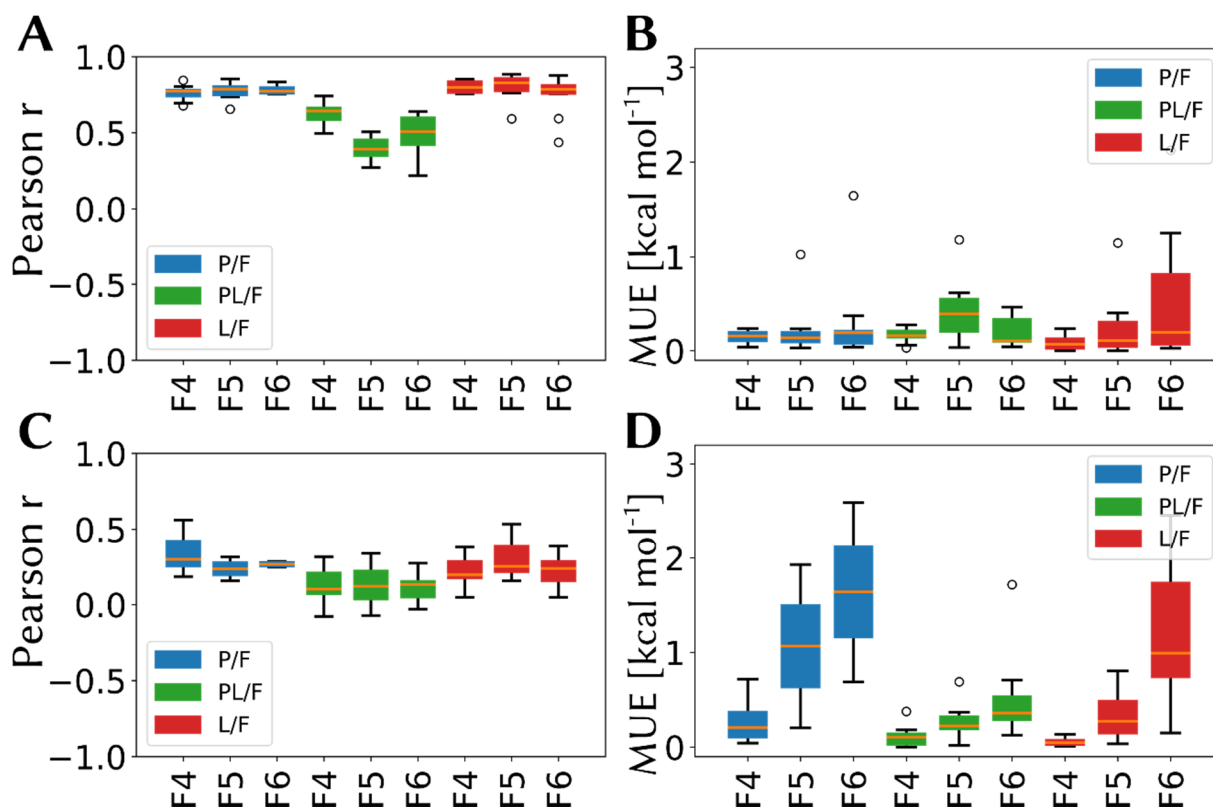


Figure 2-4. Boxplots showing correlation based on the test data from five-fold cross validation and MUE for displaced-solvent functionals ($X=\{F4,F5,F6\}$) P/FX (blue), PL/FX (green) and L/FX (red). The correlations and MUE are based on 10 random replicates of five-fold cross validation. **A:** Pearson r for the actual dataset ($p<0.05$ for all correlation coefficients); **B:** MUE for the actual dataset; **C:** Pearson r for shuffled data; **D:** MUE for the shuffled dataset.

Evaluating the Model Parameters. Since the investigated solvent functionals have physicochemical motivation, their parameters can be interpreted in a way to gain insights into the physical processes that they try to capture by the applied solvent functional. As can be seen in Figure 2-5, the fluctuations of some parameters are quite large. This is in part due to the rather extended parameter range, which allowed the energy and entropy cutoff parameters to vary between -10 to $+10$ kcal·mol⁻¹ and the solvent density cutoff from 1 to 8 ρ^0 . This is a fundamental difference to prior work,⁸² which allowed only for positive energy and entropy cutoff values, e_{CO} and s_{CO} , respectively. However, we think that the inclusion of negative values is justified, due to the fact that we investigate relative differences between ligands. In this situation, the exact nature of the reference state (bulk water phase) becomes obsolete and consequently the exact zero-point of ΔE_{solv} and $T\Delta S_{solv}$ is not relevant. But nonetheless, the sign of the cutoff values is important for evaluating the quality of the water molecules at a specific location in the binding site. We interpret the median of the parameter values, instead of their mean value, since several parameter distributions have a long tail and therefore it is not

meaningful to report the mean value. A graphical overview of the median and quartile values can be found in Figure 2-5. The numerical values of the individual median, 1st and 2nd quartile values for all functionals can be found in the Supporting Information.

For the solvent functionals based on the *apo* protein, the median energy cutoff parameter, e_{CO} , is negative for P/F5 and P/F6 (-6.56 and -4.35 kcal·mol⁻¹, respectively) and positive for P/F4 (0.62 kcal·mol⁻¹). Thus, solvent functionals P/F5 and P/F6 capture water molecules with an energy lower than the mean in the bulk solvent, whereas solvent functional P/F4 captures water molecules with an energy slightly above bulk solvent energy. Also, for P/F5 and P/F6, e_{CO} is found to have a tail into the positive regime, indicated by the 75 percentile values of 0.88 kcal·mol⁻¹ and 2.97 kcal·mol⁻¹, respectively. The energy weighting parameter, E_{aff} as well as the universal weighting parameter K_{aff} , are found at negative median values for P/F5 and P/F6, respectively. This seems to be counterintuitive, since the energy cutoff parameter, e_{CO} , allows only for water molecules that are by -6.56 and -4.35 kcal·mol⁻¹ more stable than in bulk water phase. Consequently, the displacement of energetically stable water molecules should not be favorable in terms of free energy. However, ligands which are able to displace water molecules that are energetically more stable in the binding pocket than in bulk phase, replace a protein-water interaction by a stable protein-ligand interaction. Most probably, this is the reason for the good performance of this approach. The negative sign in the energy cutoff together with the negative weighting parameter indicates that these water molecules can be used to probe for stable protein-ligand interactions. A result that would not be possible by only considering contributions from energetically *unfavorable* (i.e. $e_{CO} > 0$ kcal·mol⁻¹) water molecules. In contrast to P/F5 and P/F6, the much simpler P/F4 functional with its positive energy cutoff parameter, correctly identifies only water molecules which are energetically less stable than in bulk water phase.

A similar behavior, although with different sign, is found for the entropy cutoff parameter, s_{CO} . For this parameter, median values of 5.57, 4.41 and 0.08 kcal·mol⁻¹ were found for functionals P/F4, P/F5 and P/F6, respectively. Although this parameter does not fluctuate much for P/F4, we found more pronounced fluctuations for P/F5 and P/F6. The negative entropy weighting parameter, S_{aff} , is expected to a certain degree as it indicates that the displacement of entropically unfavorable water molecules leads to a gain in binding free energy. The solvent density cutoff, g_{CO} , was found to have only small fluctuations and median values of 2.07, 4.46 and 5.41 ρ^0 for the three functionals. This already indicates that solvent density values in the range of 4.5 to 5.5 ρ^0 are appropriate for evaluating the contributions of water molecules in the

apo pocket using the P/F5 and P/F6 functionals. For the P/F4 functional, a lower solvent density value of about $2 \rho^0$ is appropriate. Likely, the low solvent density value reflects the high energy cutoff value found for this functional.

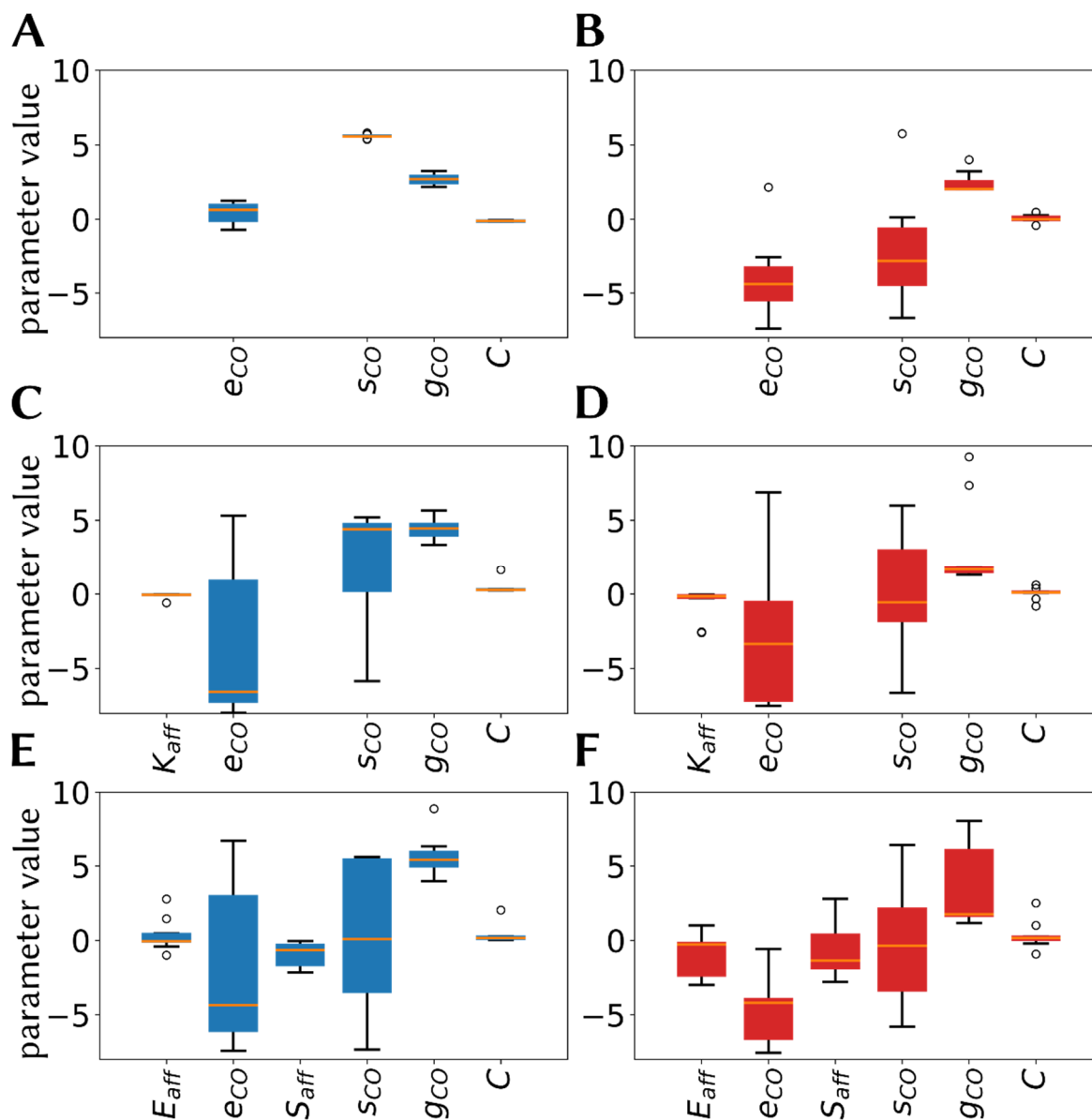


Figure 2-5. Boxplots showing distribution of the parameters for different functionals and datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. All parameters are in units $\text{kcal}\cdot\text{mol}^{-1}$, except g_{CO} , which is given in multiples of bulk density ρ^0 . **A, C, E:** P /F4, P /F5, P/F6 (blue); **B, D, F:** L/F4, L/F5, L/F6 (red).

We refrain from interpreting the parameters of the solvent functionals derived from the protein-ligand complexes (PL/F4, PL/F5 and PL/F6), since these did not result in adequate correlations with the experimental data (see Figure 2-4). Additionally, they did not significantly outperform

the solvent functionals trained with shuffled data, which speaks against the predictive power of these functionals.

For the L/F4, L/F5 and L/F6 functionals, only GIST data from MD simulations of the ligand molecules in aqueous solution were used. The energy as well as the entropy cutoff for L/F4 are low (-4.35 and -2.79 kcal·mol⁻¹, respectively). With an energy and entropy cutoff being that low, all major solvation sites on the ligand's surface contribute to binding affinity. Consequently, this solvent functional suggests that binding affinity is significantly determined by contributions from solvation energy and entropy of the ligand alone. For the other solvent functionals, L/F5 and L/F6, the energy cutoff parameters are observed at negative median values (-3.35 and -4.20 kcal·mol⁻¹, respectively), whereas the entropy cutoff parameters are close to zero for both functionals (-0.57 and -0.36 kcal·mol⁻¹, respectively). Contrary to solvent functional L/F4, rather large fluctuations are observed. The overall negative sign for the energy cutoff parameter in solvent functionals L/F4, L/F5 and L/F6 indicates that water molecules with an energy lower than in bulk solvent are effectively considered in the free energy score. As the entropy cutoff parameters of L/F5 and L/F6 are close to zero, water molecules with mostly entropically beneficial displacement upon binding are considered. Furthermore, the solvent density cutoff was found to be close to $2 \rho^0$ for all three solvent functionals. The fluctuations for this parameter were low for L/F4 and L/F5, but can increase up to $8 \rho^0$ in the case of L/F6. The negative sign for the median values for the weighting parameter, K_{aff} and E_{aff} (-0.16 and -0.27 kcal·mol⁻¹), indicates, as in the case of P/F5 and P/F6, a favorable free energy contribution from the displacement of water molecules that are bound energetically favorable compared to bulk water phase. The large and negative entropy weighting parameter, S_{aff} (-1.36 kcal·mol⁻¹), indicates that the displacement of entropically unfavorable water molecules leads to a large gain in free energy. Consequently, according to the solvent functional, entropy and energy seem to reinforce each other in the context of ligand desolvation.

It should be noted that the binding affinity has only negligible correlation of $r = 0.15$ with the logP value (calculated with the *Crippen*¹⁴⁰ logP implementation in *RDKit*¹³⁹) of the molecules studied in this work. Consequently, we assume that the solvent functionals do not capture the solubility or hydrophobicity of the ligand molecules, but the actual solvation thermodynamics within the binding process.

Distribution of the Water Molecules in the S1 Subpocket. The S1 subpocket plays a key role in the substrate recognition process of thrombin. It contains D189 deeply buried at the bottom

of this pocket and this amino acid interacts with the positively charged amino acid sidechains of the bound substrate. Another amino acid, Y228, next to D189, does not interact directly with the substrate, but poses an apolar counterpart to the charged D189 residue. As such, it accommodates weakly bound water molecules on top of the aromatic side chain, which can lead to a boost in binding affinity upon displacement.^{54,104} This residue also plays a crucial role for the development of inhibitors as it can be involved in favorable chlorine- π -interactions.

As can be seen from Figure 2-6A and Figure 2-6B, the regions which contribute with favorable entropy (red) to the free energy of binding are located mainly on top of the carboxylate group of D189 for the P/F4 and P/F5 functionals, respectively. Since the entropy cutoff value for P/F4 is $s_{CO}=5.6 \text{ kcal}\cdot\text{mol}^{-1}$, only entropically unfavorable water molecules contribute to free energy. Interestingly, these water molecules do not contribute to solvation energy, as suggested by the missing scoring regions in the energy map (blue, Figure 2-6A right) on top of D189. The energy cutoff value is at $e_{CO} = 0.6 \text{ kcal}\cdot\text{mol}^{-1}$ for P/F4, consequently it includes energy contributions, $\Delta E_{solv}(\vec{r}_k)$, from solvent molecules in regions that can only decrease the solvation energy (consider the negative sign for $\Delta E_{solv}^{(F4)}$ in P/F4, as explained in the Methods section). These regions are located on top of the sidechain of Y228 and reflect the fact that the solvation of this sidechain is accompanied by weak solute-water interactions as well as an unfavorable arrangement of water molecules in the binding pocket.

Solvent functional P/F5 does have the lowest energy cutoff parameter value ($e_{CO} = -6.6 \text{ kcal}\cdot\text{mol}^{-1}$), of all the protein pocket desolvation solvent functionals. Consequently, it includes the energetically favorable water molecules (region with quite low ΔE_{solv}) on top of D189 (Figure 2-6B, right). As already mentioned before, this likely reflects the fact that water molecules on top of the charged side chain serve as a probe for energetically favorable interactions between protein and ligand at this site.

According to P/F6, the water molecules that are favorable to displace with respect to entropy, are also in proximity to the carboxylate group of D189. However, these entropy scoring regions also distribute across the binding pocket (see Figure 2-6C, left) which is due to the lower entropy cutoff for this functional as compared to P/F4 and P/F5. Regarding solvent energy, the P/F6 functional favors water molecules bearing an energy lower than in bulk solvent ($e_{CO} = -4.4 \text{ kcal}\cdot\text{mol}^{-1}$). However, the cutoff value for this solvent functional is just high enough, such that water molecules on top of D189 are not considered for scoring in the energy term (see Figure 2-6C, right). However, the water molecules on top of Y228 are energetically unfavorable with respect to bulk water phase, which was also observed for solvent functional P/F4.

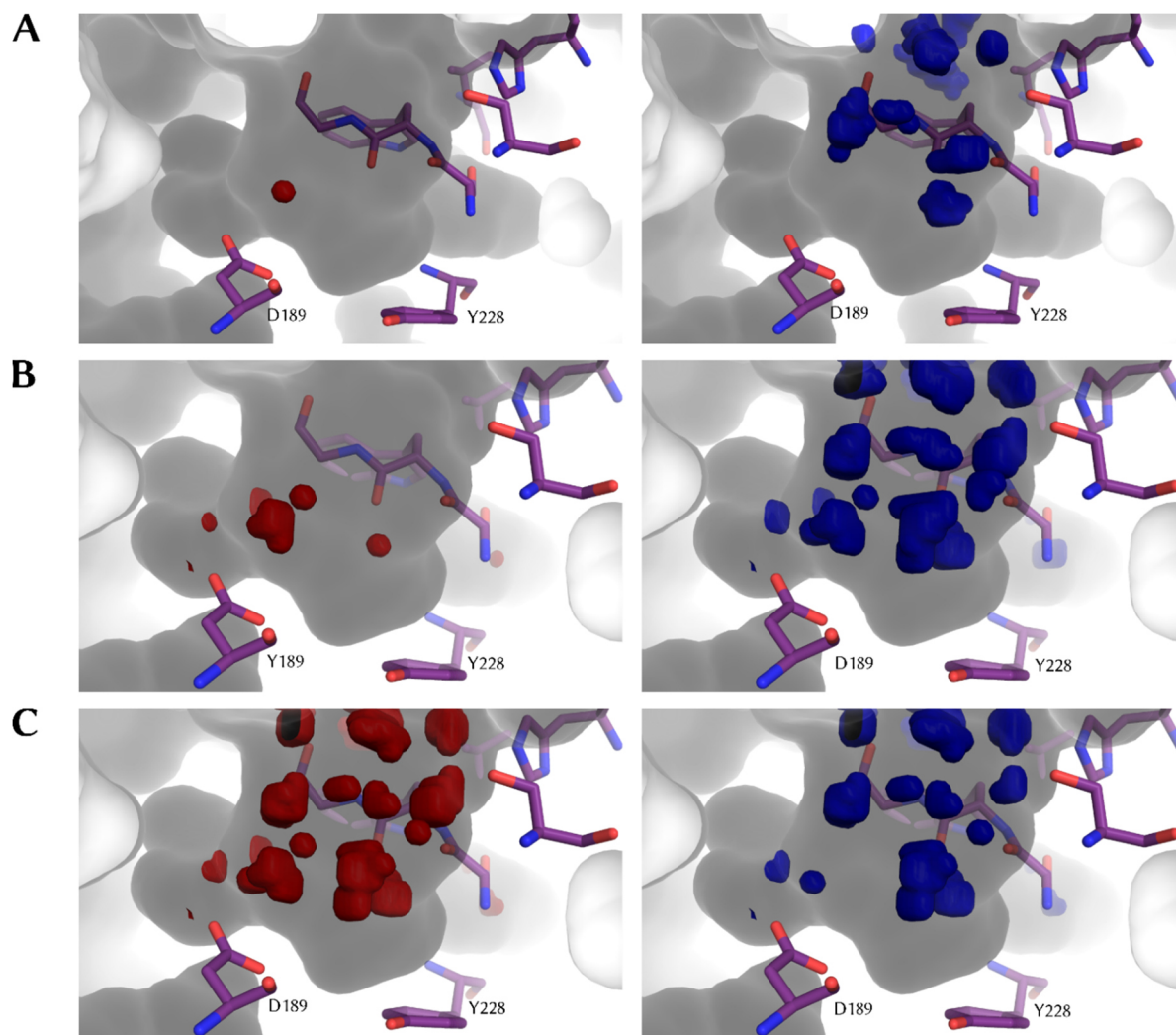


Figure 2-6. Entropy (red) and energy (blue) maps in the S1 subpocket of thrombin contoured at the median cutoff parameter values for the different solvent functionals. **A:** Representation of the distribution of the sum in eqs. (2-4) and (2-5) for the P/F4 functional contoured at $e_{CO} = 0.6 \text{ kcal}\cdot\text{mol}^{-1}$, $s_{CO} = 5.6 \text{ kcal}\cdot\text{mol}^{-1}$, $g_{CO} = 2.7 \rho^0$. **B:** Representation of the distribution of the sum in eqs. (2-11) and (2-12) for the P/F5 functional contoured at $e_{CO} = -6.6 \text{ kcal}\cdot\text{mol}^{-1}$, $s_{CO} = 4.4 \text{ kcal}\cdot\text{mol}^{-1}$, $g_{CO} = 4.5 \rho^0$. **C:** Representation of the distribution of the sum in eqs. (2-9) and (2-10) for the P/F6 functional contoured at $e_{CO} = -4.4 \text{ kcal}\cdot\text{mol}^{-1}$, $s_{CO} = 0.1 \text{ kcal}\cdot\text{mol}^{-1}$, $g_{CO} = 5.4 \rho^0$.

2.5.2 Solvent Free Energy from Two-State Full Binding-Displacement Treatment (S1-S2/F type)

Performance Considerations. In the second part of this work, we considered contributions from both the protein-ligand complex and the ligand molecule in the same calculation. The performance of functionals PL-L/F4, PL-L/F5 and PL-L/F6 show a considerably worse performance than the corresponding ones based on the individual displacement treatments. The functional PL-L/F4 shows a median correlation coefficient of only 0.4, which corresponds to just the same performance that was observed for optimization using shuffled data for this functional (see Figure 2-7A and C). This speaks against any predictive power of this functional. However, the other functionals PL-L/F5 and PL-L/F6 did perform better and slightly outperformed those trained on shuffled data. In line with this, the solvent functionals obtained with PL-L/F4 could also not achieve better accuracy than the ones trained on shuffled data, indicated by the MUE (see Figure 2-7B and D). For F5 and F6 however, improved accuracy over shuffled data was observed, indicated by the low MUE values. Due to the not quite satisfying performance of these functionals, we refrain from interpreting the parameters further with respect to potential implications for the water molecules in the system.

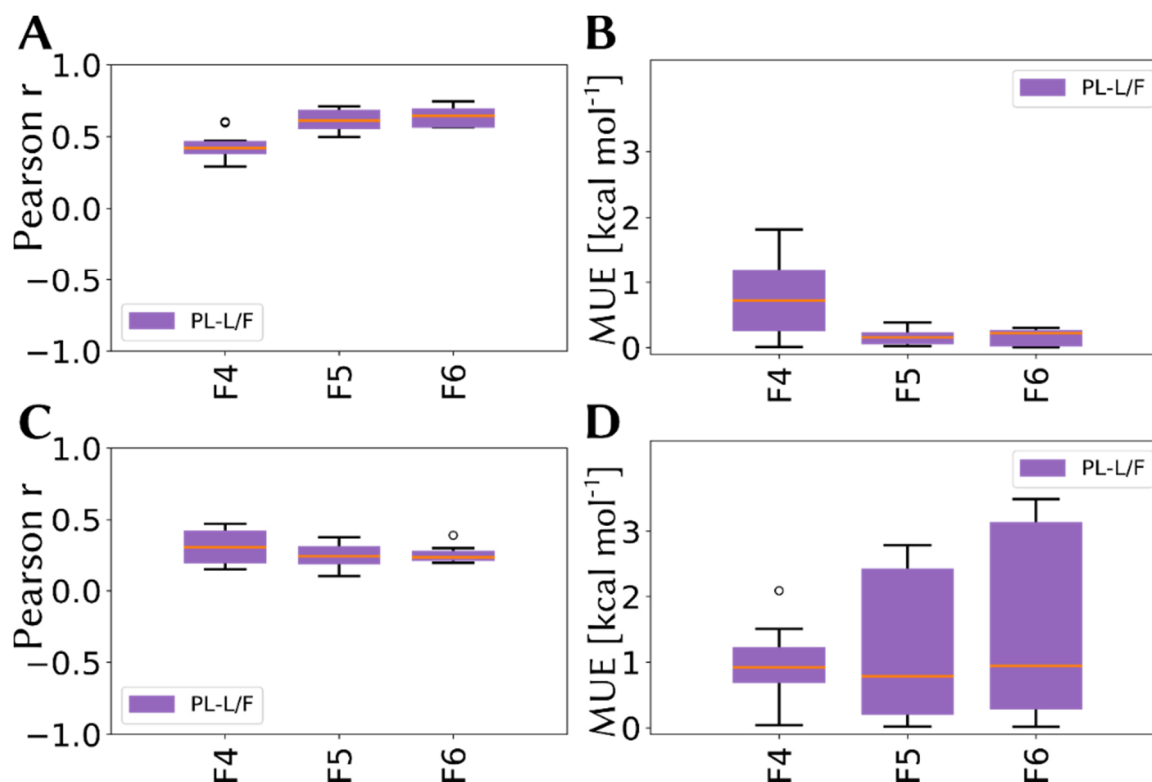


Figure 2-7: Boxplots showing Pearson correlation coefficient and MUE for solvent functionals PL-L/F4 to PL-L/F6 using GIST from both the protein-ligand complex and the ligand molecule in solution. A: Pearson r for the actual dataset ($p < 0.05$ for all correlation coefficients); B: MUE for the actual dataset; C: Pearson r for shuffled data; D: MUE for the shuffled dataset.

2.5.3 Solvation Free Energy including Explicit Optimization of Solvation Energy

The solvent free energy methods discussed so far are not parameterized using any explicit consideration of experimentally derived enthalpy or entropy contributions. Thus, no correlation between experimental and calculated enthalpy-entropy factorization is observed for the simple solvent functionals such as P/F6 (see Figure 2-8). In order to achieve correct enthalpy-entropy factorization, we performed optimization of the GIST functionals for the free energy data and the energy simultaneously using multiobjective optimization (see Methods section). In this approach, the free energy is calculated in the same way as it is in the single-objective approach, except that the parameters that affect the solvation energy (i.e. eco , gco , E_{aff} and K_{aff}) are optimized for the experimental enthalpy differences as well. By this, the solvent functional uses the entropy of solvation as a compensation for the difference between free energy and energy.

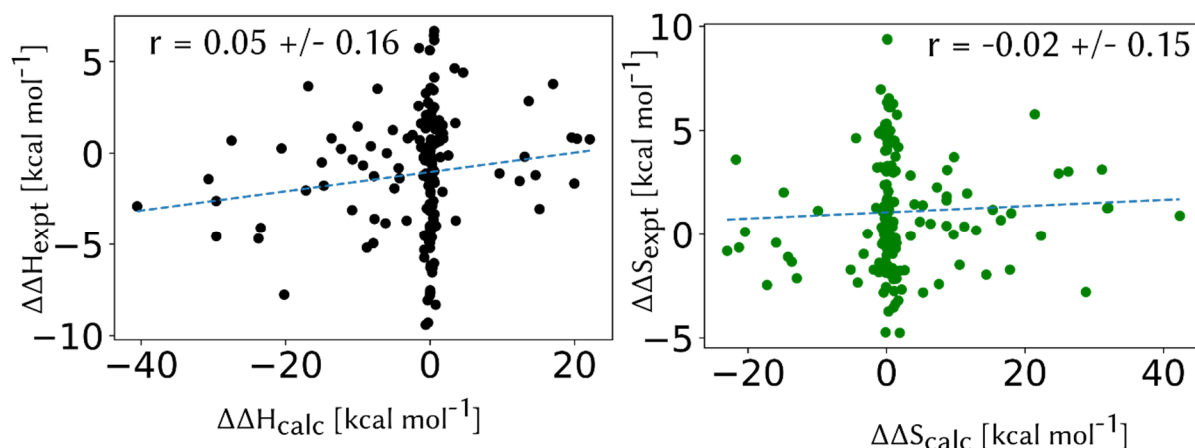


Figure 2-8: Comparison of calculated and experimental relative enthalpy values (left) and relative entropy values (right) calculated with the P/F6 solvent functional.

Performance of the Solvent Functionals. The solvent functionals P/F4, P/F5, P/F6, PL/F4, PL/F5, PL/F6, L/F4, L/F5 and L/F6 generally do not perform in a satisfactory way with respect to reproducing values of the experimental free energies (see Figure 2-9A) as obtained from multiobjective optimization. Although some agreement between experiment and calculation was found for the energy of solvation from P/F6 and L/F6, no clear correlation was found for PL/F4 to PL/F6 (see Figure 2-9B). We argue that the unsatisfactory agreement between calculated and experimental values is due to differently dominating contributions to solvation energy and entropy resulting from the unbound state or the bound state of the ligand. Specifically, we assume that either the ligand desolvation or the solvation of the protein-ligand complex contribute the lion's share to specific quantities. For this assumption, the two states must be handled differently, using individual cutoff values for the solvent density, entropy and energy, as introduced in the Theoretical Background section above. We tested two different functionals, one that employs different cutoff values for the solvent density in each state, but global (i.e. similar) entropy and energy cutoff values for each state ($R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(g)}, s_{CO}^{(g)}$) for each solvent functional PL-L/F4/R, PL-L/F5/R and PL-L/F6/R). The other functionals use different solvent density, energy and entropy cutoff values in each state ($R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$) for each solvent functional PL-L/F4/R, PL-L/F5/R and PL-L/F6/R). As a reference, we also analyzed the global (i.e. they are the same in each state) cutoff parameter setting for each state ($R = g_{CO}^{(g)}, e_{CO}^{(g)}, s_{CO}^{(g)}$) for each solvent PL-L/F4/R, PL-L/F5/R and PL-L/F6/R). These solvent functionals are similar to PL-L/F4, PL-L/F5 and PL-L/F6, which were already investigated in the first part of the Results section (see Figure 2-7 for an overview of their performance based on training/testing with free energy data).

For the solvent functional that employs individual solvent density cutoff values for each state ($R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(g)}, s_{CO}^{(g)}$), we observed little agreement of free energy and energy with experimental values for the F4 basic functional (see Figure 2-9 C,D magenta). For basic functionals F5 and F6, equal median performance was observed for the free energy, although F6 shows somewhat higher fluctuations than its F5 counterpart. In contrast to the solvent free energy, the solvent energy was observed to be in better agreement with F5. This is a bit puzzling, since the performance of F6 should not be worse than the performance of F5, as F5 can be treated as a subtype of F6, where $E_{aff} = S_{aff}$.

In the case of the solvent functional that uses individual solvent density, energy and entropy parameter settings for each state ($R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$), satisfactory performances were found for both solvent free energy as well as solvent energy (see Figure 2-9 C,D grey) for basic functionals F5 and F6. It is worth noting that no significant correlation could be determined for these functionals with shuffled data (see Supporting Information). Lastly, it must be emphasized that the functional with all-global parameter settings ($R = g_{CO}^{(g)}, e_{CO}^{(g)}, s_{CO}^{(g)}$), did not result in anything that could reliably reproduce the experimental free energy or enthalpy (see Figure 2-9 C,D brown).

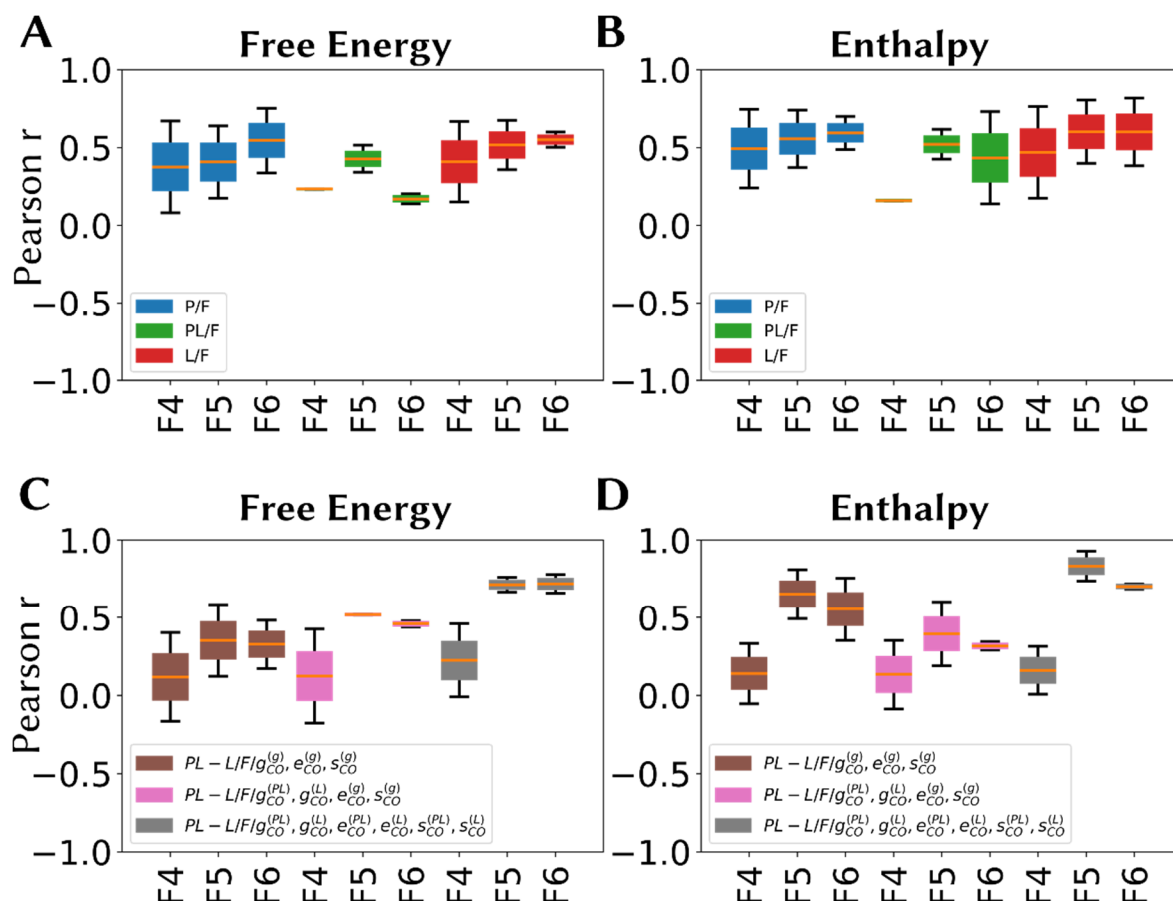


Figure 2-9: Boxplots showing the Pearson correlation coefficient for free energies and enthalpies calculated using different solvent functionals and multiobjective optimization. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Solvent free energy and energy calculated with displaced-solvent functionals based on (with $F=\{F4,F5,F6\}$) P/F (blue), PL/F (green) and L/F (red); **C,D:** Solvent free energy and energy calculated with full binding-displacement treatment.

The Parameters of the Functionals. In the following, the functionals with explicit multiobjective training of solvent free energy and energy are discussed. We only discuss the ones which actually were able to reproduce the experimentally observed enthalpy-entropy factorization. Therefore, only the parameter settings $R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$ with basic functionals F5 and F6 will be discussed.

As can be seen from Figure 2-10B, basic functional F6 results in a broad scatter of the ligand entropy cutoff parameter, $s_{CO}^{(L)}$, whereas the other parameters do not fluctuate strongly. For basic functional F5, overall less fluctuations than for basic functional F6 were observed (see Figure 2-10A). The median value of K_{aff} in F5 ($0.22 \text{ kcal}\cdot\text{mol}^{-1}$) as well as both E_{aff} and S_{aff} of F6 (0.17 and $0.01 \text{ kcal}\cdot\text{mol}^{-1}$) are positive. Thus, the solvent functionals score the solvation contributions from the protein-ligand complex to oppose binding, whereas the desolvation contributions from the ligand molecule boost binding (i.e. they lower the free energy). Furthermore, the value of

the entropy weighting parameter, S_{aff} , is almost 20 times lower than the energy weighting parameter, E_{aff} , with basic solvent functional F6. This can be due to the fact that during multiobjective optimization, entropy was only considered implicitly as the compensating difference between free energy and enthalpy. But most likely, this reflects the fact that the spread of experimental entropy is by far less than the corresponding spread in enthalpy (cf. Figure 2-3) in our dataset.

For both basic functionals F5 and F6, the median values for solvent energy, entropy and density cutoff parameters are found to be quite similar (see also the Supporting Information for a complete list of the numerical values). The median value of the energy cutoff parameter for the ligand molecule, $e_{CO}^{(L)}$, is slightly negative ($-0.95 \text{ kcal}\cdot\text{mol}^{-1}$), whereas the value for the protein-ligand complex is positive and very high in value ($8.03 \text{ kcal}\cdot\text{mol}^{-1}$). The low energy cutoff for the ligand together with the high solvent density cutoff value ($6.93 \rho^0$) for the ligand effectively allows only highly occupied regions around the ligand that are energetically only slightly stabilized. In line with the high density cutoff parameter values, also high values for the entropy cutoff parameters for the protein-ligand complex, $s_{CO}^{(PL)}$, as well as the ligand, $s_{CO}^{(L)}$, are observed (7.83 and $3.95 \text{ kcal}\cdot\text{mol}^{-1}$, respectively).

Most of the regions with high solvent density and energy (e.g. next to apolar patches) are recognized to fix water molecules that can be easily removed upon a favorable gain in solvation energy and free energy of binding. These regions can be large in size, thus indicating that the desolvation of the ligand molecule has a large contribution to the (negative) total free energy. These contributions can actually be overwhelming and thus overcompensate other contributions from the protein-ligand complex (see also Figure 2-11 for an overview of the ligand free energy factorization). As already mentioned, the protein-ligand complex contributions seem to oppose binding due to the positive sign of K_{aff} and E_{aff} for basic functionals F5 and F6. This is further substantiated by the high positive energy cutoff value for the protein-ligand complex as well as the high solvent density cutoff value ($9.97 \rho^0$). Thus, these regions contain (partly) entrapped water molecules that are unfavorable in energy with respect to bulk solvent. These water molecules should be rather replaced in the protein-ligand complex in order to gain free energy. Since the solvent energy contributions from the protein-ligand complex do oppose binding it is suggested that they rather discriminate between the individual ligand molecules and by that, contribute to the selectivity of the individual ligands with respect to the binding to the target protein.

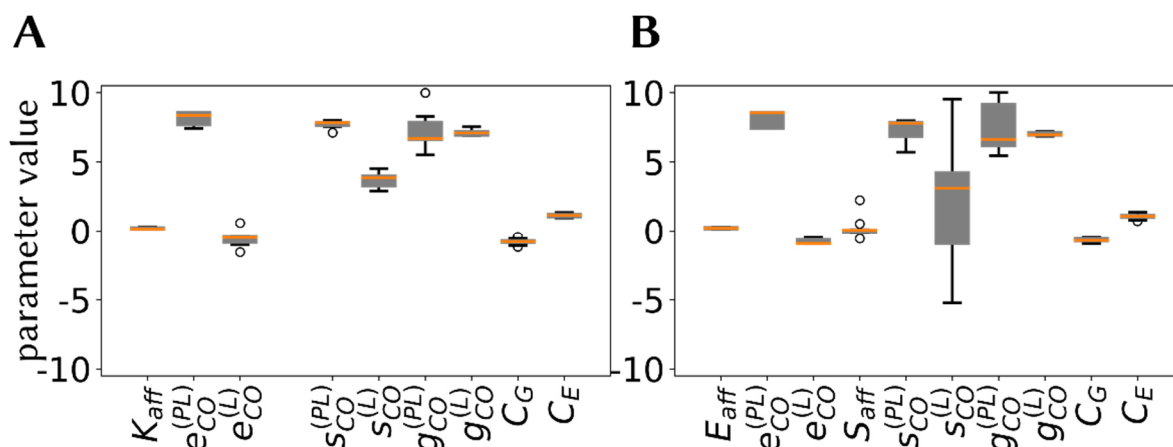


Figure 2-10: Boxplots showing the parameters for different solvent functionals obtained from multiobjective optimization with parameter settings $R = g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. All parameters are in units $\text{kcal}\cdot\text{mol}^{-1}$, except $g_{CO}^{(L)}/g_{CO}^{(PL)}$, which are given in multiples of bulk density ρ^l . **A:** Parameters for the F4 basic functional; **B:** Parameters for the F5 basic functional; **C:** Parameters for the F6 basic functional.

It must be emphasized that the weighting parameters E_{aff} , S_{aff} or K_{aff} were allowed to vary freely in the interval $[-3;+3]$ $\text{kcal}\cdot\text{mol}^{-1}$ during parameter optimization. Thus, the fact that the solvation of the protein-ligand complex opposes binding and the desolvation of the ligand favors binding was not enforced at any point during parameter optimization. Also, it must be noted that the solvent free energy values of the protein-ligand complex alone do not correlate strongly with experiment for the solvent functionals discussed in this section ($r = 0.30$). Whereas for the ligand alone reasonable correlation with experimental free energy was found ($r = 0.75$). These correlations are quite similar to the observed performance of solvent functionals PL/F6 and L/F6. However, the individual energies of the protein-ligand complex and the ligand do not appear to correlate with experimental enthalpy (both $r = 0.40$). This indicates that the free energy of solvation for the binding reaction can readily be calculated from the ligand in solution alone but rather not from the protein-ligand complex. However, for the calculation of solvation energy, the contribution of both states, the protein-ligand complex and the ligand, are necessary in order to obtain reasonable correlation with experiment.

As already noted, the desolvation of the pre-bound state of the ligand contributes the lion's share to the solvation free energy. This is also illustrated in Figure 2-11, which displays an overview of the solvent free energies and energies calculated with PL-L/F6/ $g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$. From this overview it is apparent that the desolvation energy of the ligand molecule has the largest impact on the total solvation free energy. The contributions of the protein-ligand complex are smaller (approximately one third the amount of

the contributions of the ligands), but they scatter more pronouncedly and therefore contribute significantly to the discrimination of ligand molecules that have similar desolvation behavior.

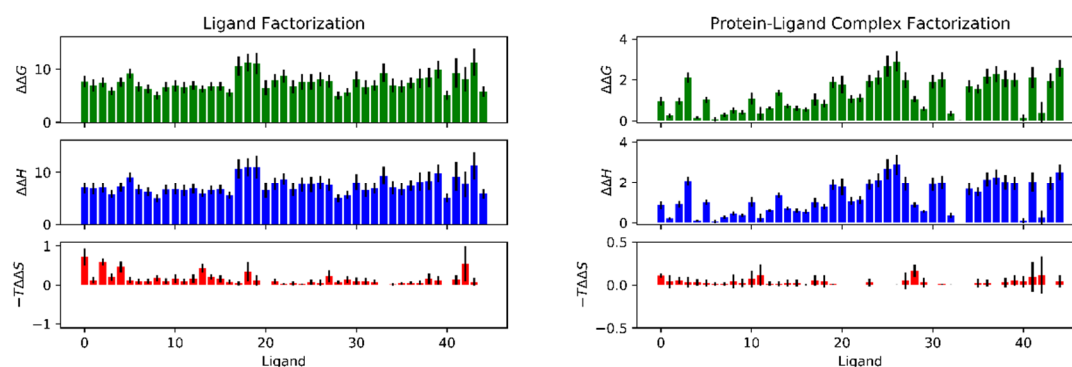


Figure 2-11: Overview of the energy-entropy factorization as obtained from the PL-L/F6/ $g_{CO}^{(PL)}$, $g_{CO}^{(L)}$, $e_{CO}^{(PL)}$, $s_{CO}^{(L)}$, $e_{CO}^{(PL)}$, $s_{CO}^{(L)}$ solvent functional. All quantities are units kcal·mol⁻¹. The error bars indicate the confidence interval at the 95% level.

2.5.4 Spatial Distribution of the Solvent Molecules in the Unbound and Bound State

As can be seen from Figure 2-12B and D, solvent molecules scatter around the positively charged terminal amino group exposed to the surface of the ligands **1** and **2** in their unbound state. This amino group is present in many of the thrombin ligands with a D-Phe-Pro scaffold considered in the evaluated data set. In the case of **2**, considerably more water molecules than for **1** are found in proximity to this amino group in the unbound state of the ligand. Furthermore, solvent molecules matching the energy cutoff are found on top of the aromatic portion in the unbound state of **2**, whereas they are clearly missing in the unbound state of **1**. In the bound state of both ligands (see Figure 2-12A and C), water molecules occupy a hydrophobic subpocket in the vicinity of W60, below the so-called 60s loop. These water molecules are also found in the *apo* crystal structure of the protein, however at rather shifted positions in the protein-ligand crystal structures. The occupation of this region with water molecules opposes binding and compensates the overall beneficial desolvation of both ligands in an unfavorable way. Ligand **1** also entraps a water molecule between its pyridine group and Y228, which is missing for ligand **2**. The energetic contribution of this entrapped water molecule is highly unfavorable with respect to the bulk water phase. Thus, **2** binds tighter to the protein than **1**, as shown by the calculated free energy difference of $\Delta\Delta G^{(\text{calc})}(\mathbf{1}\rightarrow\mathbf{2}) = -2.4\pm 0.8$ kcal·mol⁻¹ accompanied by a change in solvation energy of $\Delta\Delta H^{(\text{calc})}(\mathbf{1}\rightarrow\mathbf{2}) = -2.3\pm 0.6$ kcal·mol⁻¹. Overall, the process is driven by solvent energy, which is in agreement with the experimental free energy

difference of $\Delta\Delta G^{(\text{exp})}(\mathbf{1}\rightarrow\mathbf{2}) = -2.5\pm 0.2 \text{ kcal}\cdot\text{mol}^{-1}$ and the dominating experimental enthalpy difference of $\Delta\Delta H^{(\text{exp})}(\mathbf{1}\rightarrow\mathbf{2}) = -2.3\pm 0.2 \text{ kcal}\cdot\text{mol}^{-1}$.

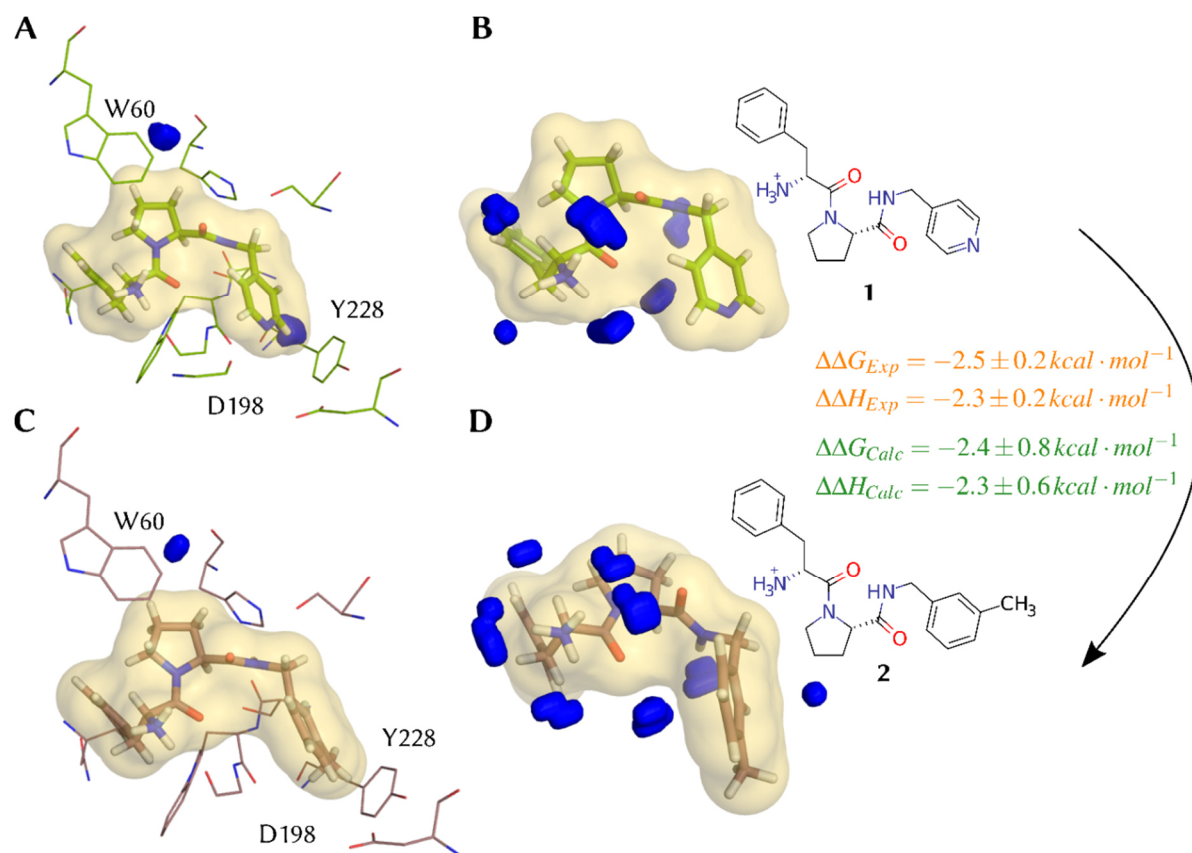


Figure 2-12: Example of unfavorable solvent energy regions calculated for ligands **1** and **2** in their unbound and bound states (both in the crystallographically observed binding pose). The maps were generated with the PL-L/F6/ $g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$ functional. The parameter values are the median values found for this particular functional. **A, B:** Solvent energy map for the bound and unbound state of **1** (PDB 3SV2)¹⁰³; **C, D:** Solvent energy for the bound and unbound state of **2** (PDB 2ZF0)¹⁰⁴. The bound states are countered at $e_{CO}^{(PL)} = 8.03 \text{ kcal}\cdot\text{mol}^{-1}$ and $g_{CO}^{(PL)} = 9.97 \rho^0$. The unbound states are contoured at $e_{CO}^{(L)} = -0.95 \text{ kcal}\cdot\text{mol}^{-1}$ and $g_{CO}^{(L)} = 3.95 \rho^0$. The errors for free energy and enthalpy (energy) display 1 stand. dev. (both for the experimental and calculated values).

2.5.5 Comparison with Other Methods

In order to see if our approach is in principle comparable to other methods, we applied the generalized Born surface area implicit solvation method (GBSA) and the 3D reference interaction site model (3D-RISM) to our systems. For both methods, we calculated their agreement with relative free energy differences from the experiment. Also, we benchmarked their performance with the addition of the internal molecular mechanics energy of the solute molecules (the MM-GBSA and MM-3DRISM approach). We used these methods for

comparison, since the required computing time for them lies between GIST and advanced free energy methods such as the Free Energy Perturbation (FEP) technique.

Once the MD simulations and GIST calculations are carried out, the computing time for our approaches depends heavily on the applied solvent functional and the parameter range that is allowed during the optimization of the parameters. For a typical solvent functional like P/F4, our program *Gips* obtains a set of converged (for convergence and termination criteria, see Methods section) parameter values after 1 hr of computing time using 20 cores on an Intel Xeon Skylake Gold 6148 processor at the Goethe-HLR compute cluster located at Goethe University Frankfurt. A multiobjective optimization, as with the $PL-L/F6/g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$ functional, usually takes about 10 hrs with the same processor type and number of cores. Thus, our approach is quite compatible to MMGBSA or MM-3DRISM approaches in terms of computing time. The comparison with advanced free energy methods based on alchemical transformation would not be feasible in the context of this study, particularly considering the required computational efforts.

The different methods and their correlation with experimental data are presented in Table 2-2 together with a comparison with our solvent functionals. The implicit solvation methods, both with and without the addition of the solute energy, were not able to reproduce the trend in binding free energy. In detail, the widely-used MM-GBSA approach (entry MM-GBSA / PL-L) based on the contributions of the protein-ligand complex and the free ligand molecule in solution did not capture the trend in binding free energy correctly. Only the consideration of the ligand molecule itself, both with and without consideration of the solute energy, were able to achieve moderate correlations of 0.58 and 0.59, respectively. With the 3D-RISM approach, no correlation was found, with and without the solute energy, if the protein-ligand complex was considered in the calculations. However, with the ligand molecule alone, we found considerable correlation of 0.75 with the 3D-RISM approach.

Table 2-2: Overview of GBSA and 3D-RISM performance compared with our solvent functionals.

Method ^{a)}	$\Delta\Delta G$ Pearson r ^{c)}	$\Delta\Delta H$ Pearson r ^{c)}
<i>P/F4</i> ^{b)}	0.76	-0.15
<i>P/F5</i> ^{b)}	0.78	0.00
<i>P/F6</i> ^{b)}	0.78	0.05
<i>L/F4</i> ^{b)}	0.81	-0.07
<i>L/F5</i> ^{b)}	0.86	-0.06
<i>L/F6</i> ^{b)}	0.88	-0.06
<i>PL-L/F5</i> / $g_{co}^{(PL)}, g_{co}^{(L)}, e_{co}^{(PL)}, s_{co}^{(L)}, e_{co}^{(PL)}, s_{co}^{(L)}$ ^{b)}	0.71	0.72
<i>PL-L/F6</i> / $g_{co}^{(PL)}, g_{co}^{(L)}, e_{co}^{(PL)}, s_{co}^{(L)}, e_{co}^{(PL)}, s_{co}^{(L)}$ ^{b)}	0.72	0.70
GBSA / PL	-0.27	0.16
GBSA / L	0.59	0.18
GBSA / PL-L	-0.26	0.17
MM-GBSA / PL	-0.40	0.06
MM-GBSA / L	0.58	0.25
MM-GBSA / PL-L	-0.35	-0.35
3D-RISM / PL	0.12	0.00
3D-RISM / L	0.75	0.26
3D-RISM / PL-L	0.13	0.01
MM-3D-RISM / PL	0.07	-0.02
MM-3D-RISM / L	0.69	0.29
MM-3D-RISM / PL-L	0.09	0.01

a) PL: Based only on the protein-ligand complex; L: Based only the ligand molecule; PL-L: Based on both the protein-ligand complex and the ligand. The PL-L approach corresponds to the standard 2-trajectory strategy in MMPBSA-type end-state analysis.

b) This work.

c) Pearson correlation between calculated and experimental free energy and enthalpy. The correlation is based on the pairwise relative differences as used throughout this work.

2.6 Comparative Analysis of the Applied Functionals

We demonstrated that displaced-solvent functionals like P/F6 can be used to calculate the solvation contribution of the free energy of binding based on the displacement of solvent molecules from the protein binding pocket. However, much simpler functionals like P/F4, which only require four parameters, can be applied to calculate the same quantity and achieve similarly satisfactory correlation with experimental data. Apart from the advantage that a functional with fewer parameters potentially requires less fine-tuning in the individual application, the simpler functional showed less fluctuations. Overall, this indicates that such a functional is less dependent on the quality and distribution of the training data. The fact that solvent displacement from the protein-binding pocket is such a good predictor for binding free energy suggests that displaceable solvent molecules are likely found at positions in the protein binding pocket that can be substituted by favorable interactions to a bound ligand molecule.

Surprisingly, an even increased performance, as compared to the solvent displacement based on the protein pocket desolvation, can be achieved by considering the contributions of the ligand molecules alone. This conclusion is particularly remarkable, as it suggests that the binding free energy differences across the series of considered thrombin ligands can be described entirely by the desolvation of the ligand molecules alone. However, this would suggest that these solvent functionals (L/F4, L/F5 and L/F6) assign the same binding free energy towards any arbitrary protein. On first sight, this suggestion cannot be correct, since it is well known that many thrombin ligands have strongly deviating binding properties already towards other related serine-proteases such as trypsin or factor Xa.¹⁴¹⁻¹⁴³ It is much more likely that the solvent functional parameters are trained on regions across the surface of the unbound ligand molecule which do contribute to solvation free energy such that they effectively correlate with binding free energy. Of course, would our ligands be trained with binding thermodynamic data towards a different protein (e.g. trypsin or factor Xa), then different parameters would be found and consequently different regions on the surface of the unbound ligand molecule would be affected, as already shown with 3D-QSAR models.^{24,143} Our findings imply that the water molecules and their thermodynamic properties across the surface of the unbound ligand molecule already constitute a blueprint of the binding free energy potentially gained by the interactions with the protein binding pocket. With other words, ligands capable to shed off tightly bound water molecules upon binding, must replace the lost water -ligand interactions by stable protein-ligand interaction. Most probably, this is the reason for the good performance of the L/F4, L/F5 and L/F6 functionals.

Taken together, the desolvation of water molecules from the binding pocket as well as the desolvation of the ligand molecule from the bulk water phase correlate well with the relative differences in binding free energy obtained from experiment, since they both serve sufficiently well as structural and thermodynamic representation of the interactions gained upon binding. In light of these considerations, it is understandable that the analysis of only the protein-ligand complex will not be able to capture the overall free energy of binding as it does not include any interactions of the ligand in the bulk phase prior to the formation of the protein-ligand complex. This is further underlined by the lack of a clear correlation between the binding free energy and the solvation free energy calculated using PL/F4, PL/F5 and PL/F6.

However, the formed protein-ligand complex is important for the consideration of the energetic contributions to the binding process. With the solvent functional that was trained with GIST data from both the protein-ligand complex and the ligand molecule, we were able to describe the free energy of binding as well as the enthalpy (energy) of binding. The free energy of binding seems to be mainly driven by the energetically dominating ligand desolvation contributions, however, the actual discrimination between ligands results from the enthalpic inventory of both, the complex and the ligand. The functionals suggest that ligand desolvation is essentially driven by the shedding of tightly bound water molecules with unfavorable energy compared to the bulk phase. In the protein-ligand complex, bound water molecules only contribute to binding if they are very unfavorable and consequently high in energy with respect to bulk water.

The striking performance of the solvent functionals based on GIST ligand data in our approach to predict differences in free energy of binding across a series of molecules is reminiscent of a very popular method developed about 30 years ago, the so-called 3D-QSAR method (e.g. CoMFA¹⁴⁴ and CoMSIA¹⁴⁵). In this approach a set of ligands has to be mutually aligned with conformations assumed to resemble the bound ligand geometries at the binding site of a target protein. With increasing availability of crystal structures of protein-ligand complexes, the mutual alignment has been assisted more and more by modeling the ligands into the binding site of the crystallographically characterized proteins. For data evaluation, the thus aligned ligands are embedded into an equally-spaced grid and by use of a molecular probe along with a distance dependent functional form, interaction potential values were assigned to the intersections of the surrounding grid. The correlation of the binding affinity with trends in the data assigned to the various grid points is achieved by PLS analysis. Besides the correlation of binding affinity with

the aligned molecules, the coefficients obtained at the different grid intersections allowed to spatially detect areas that vary and thus explain trends in the affinity data across the data set.

As functional form to map the binding properties of the ligands, various functionals have been applied, among them potentials taken from force fields (Coulomb, Lennard-Jones, potentials from Goodford's GRID¹⁴⁶ or the HINT¹⁴⁷ program). In our current work, we map the Amber force-field by an MD simulation and the obtained maps provide the input for our GIST analysis. As an advantage, a water probe detects donor and acceptor properties and via the local populations produced insights into the entropic aspect are made available.

The 3D-QSAR approaches achieve impressive predictive power even though their conceptual limitations with respect to features of the surrounding protein binding site, solvation properties or entropic considerations are evident, leaving the persisting question why 3D QSAR performs so well. Possibly our current work provides some answers to this nasty topic. As pointed out, our GIST analysis using only the ligand data allows screening the solvation thermodynamic properties across the ligand surfaces and thus constitutes a kind of blueprint of the binding free energy potentially gained by the interactions formed by the ligands in the protein binding pocket. Via the indication of displaceable solvent molecules, likely positions are found that can be substituted by favorable interactions once a ligand is bound in the protein binding pocket. Clearly, 3D QSAR does not capture features involving differences in the solvation patterns of the protein-ligand complexes. However, our GIST analysis of the protein-ligand complexes shows that bound water molecules only contribute to binding if they are very unfavorable and consequently high in energy with respect to bulk water phase. Likely, these situations are less frequent across congeneric series of ligands, but definitely, when present, they will contribute to false correlations in the 3D QSAR evaluations.

Admittedly, our study is based on one comprehensive data set and further evaluations of other data sets have to show the general validity of our considerations. Nonetheless, our approach has proven to show excellent agreement with a broad range of experimental thermodynamic data covered by our dataset. Moreover, our approach is comparable in computing time with other well-studied free energy methods, such as MM-GBSA and MM-3D-RISM, but considerably outperforms those methods in terms of accuracy and predictive power.

2.7 Conclusion

In this work, we presented a significant advance in considering solvation phenomena in drug discovery. Our work is based on GIST calculations and demonstrates how this method can be

used to develop models that are able to explain experimental enthalpy-entropy factorization. Furthermore, we demonstrated how our approach is used to partition solvent thermodynamics into individual contributions from the protein-ligand complex and the ligand in the bulk phase prior to complex formation. Furthermore, we introduced a much simpler form of the already widely used displaced-solvent functionals. This new form uses fewer parameters and is demonstrated to be less sensitive towards the composition of the training set. Admittedly, the method has been developed and assessed only on one comprehensive data set of thrombin ligands. We believe the approach has potential for general applicability, since we used a comprehensive dataset of ligand molecules covering a variety of chemical features. However, it is necessary to carry out further investigations in order to elucidate the general applicability of the approach. Nevertheless, the study allows some insights why methods such as 3D-QSAR analysis provide results with high predictive power.

We hope that our work will further stimulate the consideration and subsequently the implementation of solvation-based design strategies in the arsenal of tools for the design of novel drug molecules. Explicit solvation models are an underestimated and often poorly understood aspect in our current design strategies of late-stage drug discovery. We seek to make such methods more transparent and hopefully enhance their use in the future with the present study.

The methods developed in this work are available within the *Gips* (GIST-based processing of solvent functionals) software project. It is available free of charge to the scientific community from the GitHub page of the lead author (<https://github.com/wutobias>).

2.8 Supporting Material

2.8.1 PDB Accession Codes

Ligand bound structures:

Reference [¹⁰⁴] 2ZC9, 2ZDA, 2ZDV, 2ZF0, 2ZFF.

Reference [¹⁰⁵] 2ZFP, 3DHK, 2ZGX, 2ZO3, 3DUX.

Reference [¹¹⁰] 3BIU, 3BIV.

Reference [¹⁰³] 3P17, 3QTO, 3SI3, 3SI4, 3SV2, 3QTV, 3SHC, 3QWC, 3QX5.

Reference [⁴⁹] 3RLW, 3RLY, 3RM0, 3RM2, 3RML, 3RMM, 3RMN, 3RMO, 3T5F, 3UWJ.

Reference [¹¹¹] 3UTU.

Reference [¹⁰⁸] 4BAK, 4BAM, 4BAN, 4BAO, 4BAQ.

Reference [¹⁰⁹] 4UD9, 4UDW, 4UE7, 5AF9, 5AFZ.

Reference [^{107,148}] 6GBW, 5JFD, 5LCE, 5JZY, 5LPD

Reference [¹⁰⁶] CC01, CC04, CC05, CC08, CC10, CC11.

Reference [¹¹²] *Apo* structure: 2UUF.

2.8.2 Ligand Smiles Codes

Table S2-3: Smiles codes for all ligand molecules in this study

PDB	Smiles Code
2ZDV	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cccc(F)c1</chem>
2ZC9	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cccc(Cl)c1</chem>
2ZF0	<chem>Cc1cccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2cccc2)c1</chem>
2ZFF	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cccc1</chem>
2ZFP	<chem>CC[C@@H]([NH3+])C(=O)N1CCC[C@H]1C(=O)NCc1cccc(Cl)c1</chem>
2ZGX	<chem>CC[C@@H]([NH3+])C(=O)N1CCC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1</chem>
2ZDA	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2cccc2)cc1</chem>
2ZO3	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])C(c2cccc2)c2cccc2)cc1</chem>
3BIV	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)C[NH2+][C2CCCC2])cc1</chem>
3BIU	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)C[NH2+][C2CCCC2])cc1</chem>
3DHK	<chem>[NH3+][C@@H](C(=O)N1CCC[C@H]1C(=O)NCc1cccc(Cl)c1)C(c1cccc1)c1cccc1</chem>
3DUX	<chem>[NH3+][C@H](CC1CCCC1)C(=O)N1CCC[C@H]1C(=O)NCc1cccc(Cl)c1</chem>
3P17	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccnc1</chem>
3QTV	<chem>C[n+]1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2cccc2)cc1</chem>
3QTO	<chem>C[n+]1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2cccc2)c1</chem>
3QWC	<chem>C[n+]1ccc(Cl)c(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2cccc2)c1</chem>
3QX5	<chem>C[n+]1ccc(Cl)cc1CNC(=O)[C@@H]1CCCN1C(=O)[C@H]([NH3+])Cc1cccc1</chem>
3RLY	<chem>C[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1</chem>
3RLW	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)CNS(=O)(=O)Cc2cccc2)cc1</chem>
3RM0	<chem>CC(C)[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1</chem>
3RM2	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@@H](CC2CCCC2)NS(=O)(=O)Cc2cccc2)cc1</chem>
3RMM	<chem>C[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1C[NH3+]</chem>
3RML	<chem>[NH3+][C@H]1ccc(Cl)cc1CNC(=O)[C@@H]1CCCN1C(=O)CNS(=O)(=O)Cc1cccc1</chem>
3RMN	<chem>CC(C)[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1C[NH3+]</chem>
3RMO	<chem>[NH3+][C@H]1ccc(Cl)cc1CNC(=O)[C@@H]1CCCN1C(=O)[C@@H](CC1CCCC1)NS(=O)(=O)Cc1cccc1</chem>
3SHC	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccn1</chem>
3SI3	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cccc1</chem>
3SI4	<chem>C[n+]1cccc1CNC(=O)[C@@H]1CCCN1C(=O)[C@H]([NH3+])Cc1cccc1</chem>
3SV2	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccnc1</chem>
3T5F	<chem>CC(C)C[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1C[NH3+]</chem>
3UTU	<chem>COc1ccc(S(=O)(=O)N[C@@H](CC(=O)NCc2ccc(C#N)cc2)C(=O)N2CCC[C@H]2C(=O)NCc2ccc(C(N)=[NH2+])cc2)cc1Cl</chem>
3UWJ	<chem>CC(C)C[C@@H](NS(=O)(=O)Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1</chem>
4BAK	<chem>CCNC(=O)C[NH2+][C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1)C1CCCC1</chem>
4BAN	<chem>CNC(=O)C[NH2+][C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1)C1CCCC1</chem>
4BAM	<chem>CN(C)C(=O)C[NH2+][C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1)C1CCCC1</chem>
4BAO	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCN2C(=O)[C@H]([NH2+])CC(N)=O)C2CCCC2)c1</chem>
4BAQ	<chem>CCNC(=O)C[NH2+][C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(C(N)=[NH2+])cc1)C1CCCC1</chem>
4UDW	<chem>[NH3+][C@H](Cc1cccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1Cl</chem>
4UE7	<chem>NC(=[NH2+])N1CCCC1</chem>
5AF9	<chem>COc1ccc(C(=O)Nc2cccn2)cc1</chem>
4UD9	<chem>NC(=O)c1ccc(Cl)s1</chem>

5AFZ	<chem>NC(=[NH2+])c1ccc(CNC(=O)CNC(=O)[C@@H](Cc2ccccc2)NS(=O)(=O)Cc2ccccc2)cc1</chem>
5LCE	<chem>[NH3+][C@H](CC1CCCCC1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1CO</chem>
5JZY	<chem>NC(=[NH2+])c1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])CC2CCCCC2)cc1</chem>
5LPD	<chem>[NH3+][C@H](Cc1ccc(Cl)cc1)CNC(=O)[C@@H]1CCCN1C(=O)[C@H]([NH3+])CC1CCCCC1</chem>
CC01	<chem>NC1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2ccccc2)ccn1</chem>
CC04	<chem>NC1ccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2ccccc2)ccn1</chem>
CC05	<chem>COc1cccc(CNC(=O)[C@@H]2CCCN2C(=O)[C@H]([NH3+])Cc2ccccc2)c1</chem>
CC08	<chem>[NH3+][C@H](Cc1ccccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(O)cc1</chem>
CC10	<chem>[NH3+][C@H](Cc1ccccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(Cl)s1</chem>
CC11	<chem>[NH3+][C@H](Cc1ccccc1)C(=O)N1CCC[C@H]1C(=O)NCc1ccc(O)c1</chem>
6GBW	<chem>NC(=[NH2+])NCCC[C@H](NS(=O)(=O)Cc1ccccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1C[NH3+]</chem>
5JFD	<chem>NC(=[NH2+])NCCC[C@@H](NS(=O)(=O)Cc1ccccc1)C(=O)N1CCC[C@H]1C(=O)NCc1cc(Cl)ccc1C[NH3+]</chem>

2.8.3 Experimental Thermodynamic Data

Table S2-4: Experimental Thermodynamic profile for the thrombin binders in this work.

PDB	ΔG [kcal·mol ⁻¹]	ΔH [kcal·mol ⁻¹]	ΔTS [kcal·mol ⁻¹]
2ZC9	-8.46	-8.86	-0.41
2ZDA	-11.01	-9.58	1.46
2ZDV	-7.48	-3.13	4.35
2ZF0	-8.31	-6.81	1.50
2ZFF	-7.57	-3.25	4.32
2ZFP	-7.48	-8.00	-0.53
2ZGX	-9.58	-9.24	0.33
2ZO3	-11.58	-11.35	0.21
3BIU	-8.46	-4.04	4.42
3BIV	-8.65	-2.51	6.14
3DHK	-9.46	-10.89	-1.43
3DUX	-9.20	-7.86	1.34
3P17	-6.16	-4.23	1.93
3QTO	-5.71	-4.47	1.24
3QTV	-5.73	-5.64	0.10
3QWC	-5.80	-5.35	0.45
3QX5	-5.66	-5.57	0.10
3RLW	-10.72	-3.42	7.31
3RLY	-10.17	-3.80	6.38
3RM0	-11.25	-3.30	7.95
3RM2	-12.83	-2.72	10.10
3RML	-11.42	-8.46	2.96
3RMM	-11.27	-7.67	3.61
3RMN	-12.95	-8.65	4.30
3RMO	-13.02	-6.85	6.16
3SHC	-7.52	-7.24	0.29
3SI3	-5.97	-4.73	1.24
3SI4	-5.11	-3.70	1.41
3SV2	-5.78	-4.47	1.31
3T5F	-12.97	-7.19	10.08
3UTU	-14.09	-9.60	4.49
3UWJ	-12.40	-2.48	9.91
4BAK	-12.10	-5.55	6.55
4BAM	-12.14	-5.49	6.64
4BAN	-11.61	-5.48	6.13
4BAO	-11.33	-5.10	6.23
4BAQ	-12.01	-5.77	6.24
4UD9	-4.54	-7.09	-2.56

4UDW	-8.96	-10.77	-1.82
4UE7	-5.45	-3.68	1.77
5AF9	-4.11	-3.75	0.36
5AFZ	-9.27	-4.94	4.32
5JFD	-10.39	-12.44	-2.05
5LCE	-8.91	-12.56	-3.65
5JZY	-9.34	-13.21	-3.87
5LPD	-10.89	-13.04	-2.01
CC01	-8.38	-19.35	-10.82
CC04	-8.77	-19.35	-10.58
CC05	-7.71	-8.05	-0.32
ZC08	-7.28	-7.33	-0.05
CC10	-7.62	-9.41	-1.79
CC11	-7.50	-12.54	-5.02
6GBW	-9.29	-8.93	0.33

2.8.4 Clustering Statistics

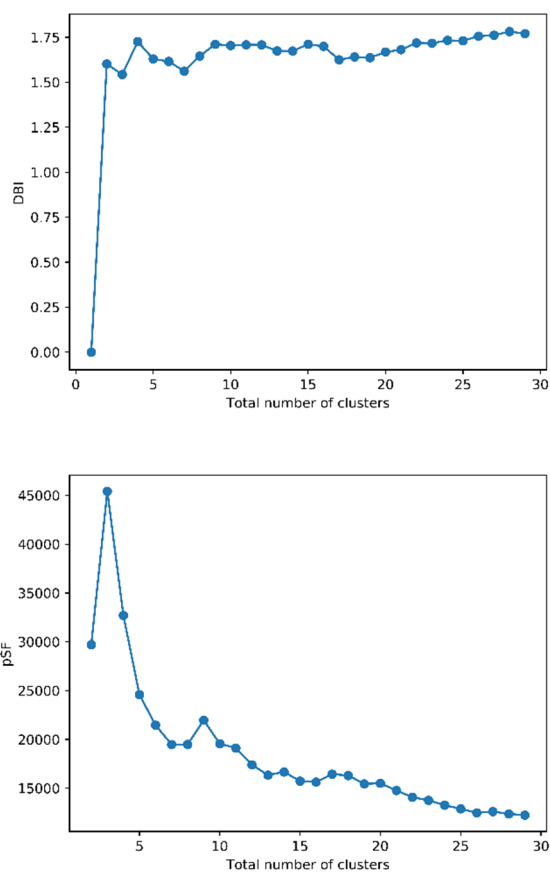


Figure S2-13: Davies-Bouldin index (left) and pseudo F-statistics (right) for clustering solutions with $N=1, \dots, 30$ using the average linkage algorithm as outlined in the main text.

2.8.5 Parameter Statistics for Displaced-Solvent Calculations

Table S2-5. Parameters for single-state solvent functional trained with free energy.

Solvent Functional ^{a)}		E_{aff} ^{b)}	e_{co}	S_{aff}	s_{co}	g_{co}	C
P	F4	N.A.	0.62/ -0.09/ 0.96	N.A.	5.57/ 5.56/ 5.61	2.70/ 2.43/ 2.91	-0.11/ -0.14/ -0.08
	F5	-0.05/ -0.06/ -0.04	-6.56/ -7.22/ 0.88	N.A.	4.41/ 0.21/ 4.75	4.46/ 4.00/ 4.76	0.28/ 0.26/ 0.30
	F6	-0.05/ -0.07/ 0.43	-4.35/ -6.06/ 2.97	-0.65/ -1.65/ -0.31	0.08/ -3.45/ 5.42	5.41/ 4.98/ 5.92	0.16/ 0.12/ 0.24
PL	F4	N.A.	-3.53/ -4.04/ -1.93	N.A.	-4.59/ -7.08/ -0.93	1.36/ 1.15/ 1.76	-0.01/ -0.02/ 0.05
	F5	-0.61/ -0.85/ -0.37	-3.67/ -5.36/ -0.80	N.A.	-0.60/ -5.39/ 4.23	6.52/ 2.77/ 8.94	0.20/ -0.27/ 0.63
	F6	2.22/ 1.88/ 2.67	3.06/ 2.93/ 3.34	-0.76/ -0.87/ -0.37	-3.49/ -4.99/ -0.65	3.11/ 2.19/ 3.84	0.19/ 0.14/ 0.21
L	F4	N.A.	-4.35/ -5.47/ -3.25	N.A.	-2.79/ -4.38/ -0.61	2.04/ 2.02/ 2.53	-0.00/ -0.03/ 0.16
	F5	-0.16/ -0.26/ -0.10	-3.35/ -7.15/ -0.55	N.A.	-0.57/ -1.80/ 2.96	1.68/ 1.47/ 1.80	0.10/ 0.07/ 0.13
	F6	-0.27/ -2.36/ -0.18	-4.20/ -6.60/ -3.96	-1.36/ -1.86/ 0.37	-0.36/ -3.35/ 2.11	1.75/ 1.65/ 6.05	0.15/ 0.05/ 0.24

The units of all parameters are expressed in units of kcal·mol⁻¹, except g_{CO} , which is given in multiples of ρ^0 . The reported values are the value for median, the 1st quartile and the 2nd quartile.

- a) Functional F4 to F6 are as described in the main text used for fitting procedure.
b) For functional F5, this is the generic affinity parameter K_{aff} .

Table S2-6. Parameters for two-state solvent functionals trained with free energy.

Solvent Functionals ^{a)}		E_{aff} ^{b)}	e_{co}	S_{aff}	s_{co}	g_{co}	C
PL/L	F4	N.A.	-1.01/ -3.09/ 4.51	N.A.	7.23/ -1.29/ 7.40	7.26/ 6.54/ 8.50	-0.86/ -1.83/ 0.52
	F5	0.09/ 0.06/ 0.10	3.08/ 2.71/ 3.26	N.A.	5.84/ 5.60/ 6.12	9.36/ 8.53/ 9.63	-0.07/ 0.14/ -0.03
	F6	0.08/ 0.05/ 0.10	3.03/ 2.79/ 3.12	0.08/ 0.04/ 0.10	5.91/ 5.22/ 6.05	8.72/ 6.25/ 9.59	-0.05/ -0.07/ -0.02

The units of all parameters are expressed in units of kcal·mol⁻¹, except g_{co} , which is given in multiples of ρ^0 . The reported values are the value for median, the 1st quartile and the 2nd quartile.

a) Functional F4 to F6 are as described in the main text used for fitting procedure.

b) For functional F5, this is the generic affinity parameter K_{aff} .

Table S2-7. Parameters for single-state solvent functionals trained with free energy and enthalpy.

Solvent Functionals ^{a)}		E_{aff} ^{b)}	e_{co}	S_{aff}	s_{co}	g_{co}	C_G	C_E
P	F4	N.A.	-5.88/ -7.47/ -5.85	N.A.	5.94/ 5.86/ 7.09	5.12/ 4.30/ 6.13	-0.86/ -0.93/ -0.80	1.12/ 1.06/ 1.18
	F5	-0.12/ -0.14/ -0.11	-7.26/ -8.38/ -6.83	N.A.	5.15/ 5.05/ 5.19	7.39/ 6.29/ 7.54	-0.93/ -0.99/ -0.85	1.11/ 1.05/ 1.14
	F6	-0.09/ -0.30/ -0.03	-4.70/ -7.31/ 0.78	-8.29/ -9.73/ -5.61	6.01/ 5.96/ 6.04	4.38/ 2.87/ 8.39	-0.94/ -1.05/ -0.83	1.05/ 1.03/ 1.11
PL	F4	N.A.	-4.04/ -4.04/ 5.89	N.A.	0.03/ 0.00/ 2.61	1.14/ 1.14/ 5.44	-1.11/ -1.19/ -1.05	1.14/ 1.11/ 1.22
	F5	6.81/ -2.92/ 7.99	2.80/ 0.22/ 2.84	N.A.	5.97/ 4.28/ 6.44	3.42/ 3.40/ 9.61	-1.16/ -1.23/ -1.10	1.24/ 1.17/ 1.29
	F6	2.12/ -2.33/ 6.66	1.53/ 0.22/ 2.84	3.98/ -1.57/ 10.00	4.90/ 3.18/ 7.36	3.53/ 3.10/ 8.06	-1.16/ -1.23/ -1.08	1.21/ 1.06/ 1.28
L	F4	N.A.	5.25/ 4.90/ 5.29	N.A.	0.00/ -0.00/ -0.00	1.00/ 1.00/ 1.03	-1.15/ -1.18/ -1.05	1.14/ 1.08/ 1.22
	F5	-9.95/ -10.00/ -9.64	-0.56/ -0.56/ -0.52	N.A.	4.00/ 4.00/ 4.00	7.59/ 7.17/ 7.59	-0.91/ -0.94/ -0.85	1.13/ 1.09/ 1.19
	F6	-10.00/ -10.00/ -8.91	-0.58/ -0.94/ -0.56	2.37/ -4.91/ 3.28	2.65/ 2.34/ 3.27	6.99/ 6.94/ 7.45	-0.92/ -1.01/ -0.82	1.12/ 1.07/ 1.13

The units of all parameters are expressed in units of kcal·mol⁻¹, except g_{CO} , which is given in multiples of ρ^0 . The reported values are the value for median, the 1st quartile and the 2nd quartile.

- a) Functional F4 to F6 are as described in the main text used for fitting procedure.
b) For functional F5, this is the generic affinity parameter K_{aff} .

Table S2-8. Parameters for two-state solvent functionals trained with free energy and enthalpy.

Solvent Functional ^{a)}		$E_{aff}^{b)}$	$e_{CO}^{(g)}$ $e_{CO}^{(PL)}$	$e_{CO}^{(L)}$	S_{aff}	$s_{CO}^{(g)}$ $s_{CO}^{(PL)}$	$s_{CO}^{(L)}$	$g_{CO}^{(g)}$ $g_{CO}^{(PL)}$	$g_{CO}^{(L)}$	C_G	C_E
PL-L/ $e_{CO}^{(g)}, s_{CO}^{(g)}$ $g_{CO}^{(g)}$	F4	N.A.	10.00/ 10.00/ 10.00	N.A.	N.A.	9.46/ 8.81/ 9.68	N.A.	9.99/ 9.91/ 9.99	N.A.	0.00/ 0.00/ 0.00	0.00/ 0.00/ 0.04
	F5	0.41/ 0.19/ 0.45	8.55/ 7.71/ 8.62	N.A.	N.A.	7.03/ 6.42/ 7.42	N.A.	5.96/ 5.94/ 8.71	N.A.	0.00/ 0.00/ 0.00	0.11/ 0.01/ 0.19
	F6	0.10/ 0.09/ 0.11	8.55/ 2.85/ 8.56	N.A.	0.72/ 0.53/ 0.78	7.54/ 7.54/ 7.54	N.A.	6.35/ 5.96/ 9.90	N.A.	0.00/ 0.00/ 0.00	0.00/ 0.00/ 0.01
PL-L/ $e_{CO}^{(g)}, s_{CO}^{(g)}$ $g_{CO}^{(PL)}, g_{CO}^{(L)}$	F4	N.A.	9.99/ 9.99/ 10.00	N.A.	N.A.	9.14/ 8.76/ 9.67	N.A.	9.61/ 9.38/ 9.98	2.76 / 2.54 / 2.80	-1.16/ -1.39/ -1.08	1.04/ 1.00/ 1.17
	F5	0.12/ 0.10/ 0.18	0.79/ -5.51/ 7.66	N.A.	N.A.	6.20/ 6.20/ 7.07	N.A.	9.48/ 9.48/ 9.79	6.54 / 2.75 / 6.95	-1.06/ -1.15/ -0.88	1.17/ 1.04/ 1.23
	F6	0.08/ 0.07/ 0.12	-0.56/ -4.26/ 1.50	N.A.	0.27/ -5.44/ 0.46	7.55/ 7.49/ 7.97	N.A.	9.73/ 9.49/ 9.94	4.96 / 2.52 / 6.49	-1.01/ -1.14/ -0.85	1.13/ 1.07/ 1.17
PL-L/ $e_{CO}^{(PL)}, s_{CO}^{(L)}$ $s_{CO}^{(PL)}, s_{CO}^{(L)}$ $g_{CO}^{(PL)}, g_{CO}^{(L)}$	F4	N.A.	10.00/ 10.00/ 10.00	- 1.17/ - 1.17/ -1.17	N.A.	9.39/ 8.71/ 9.76	4.54/ 4.53/ 4.58	9.37/ 9.37/ 9.38	3.25 / 3.25 / 3.26	-1.31/ -1.41/ -1.14	1.08/ 1.04/ 1.21
	F5	0.22/ 0.15/ 0.29	7.80/ 7.43/ 8.25	- 0.75/ - 0.95/ -0.47	N.A.	7.88/ 7.54/ 7.96	4.08/ 3.99/ 4.52	6.37/ 5.47/ 9.15	6.94 / 6.92 / 7.03	-0.70/ -0.78/ -0.57	1.05/ 1.02/ 1.11
	F6	0.17/ 0.16/ 0.29	8.03/ 7.43/ 8.56	- 0.95/ - 0.96/ -0.65	0.01/ -0.14/ 0.61	7.83/ 6.83/ 7.89	3.95/ - 0.96/ 5.15	9.97/ 5.63/ 9.99	6.93 / 6.88 / 6.95	-0.65/ -0.79/ -0.57	1.03/ 0.96/ 1.14

The units of all parameters are expressed in units of $\text{kcal} \cdot \text{mol}^{-1}$, except g_{CO} , which is given in multiples of ρ^0 . The reported values are the value for median, the 1st quartile and the 2nd quartile.

a) Functional F4 to F6 are as described in the main text used for fitting procedure.

b) For functional F5, this is the generic affinity parameter K_{aff} .

2.8.6 Correlation Statistics for Training Data Based on Actual Datasets

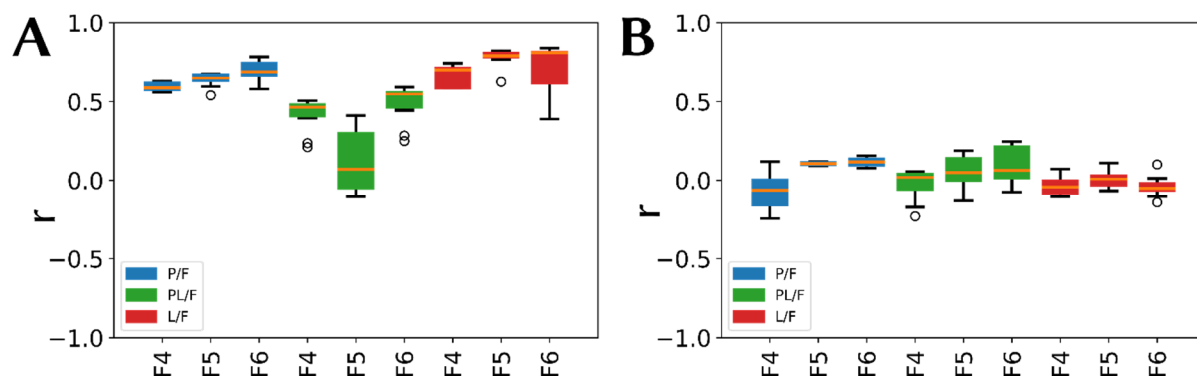


Figure S2-14: Boxplots showing correlation based on the training data from five-fold cross validation for solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$) using the actual datasets. **A:** Pearson r for the actual dataset; **B** Pearson r for the shuffled dataset.

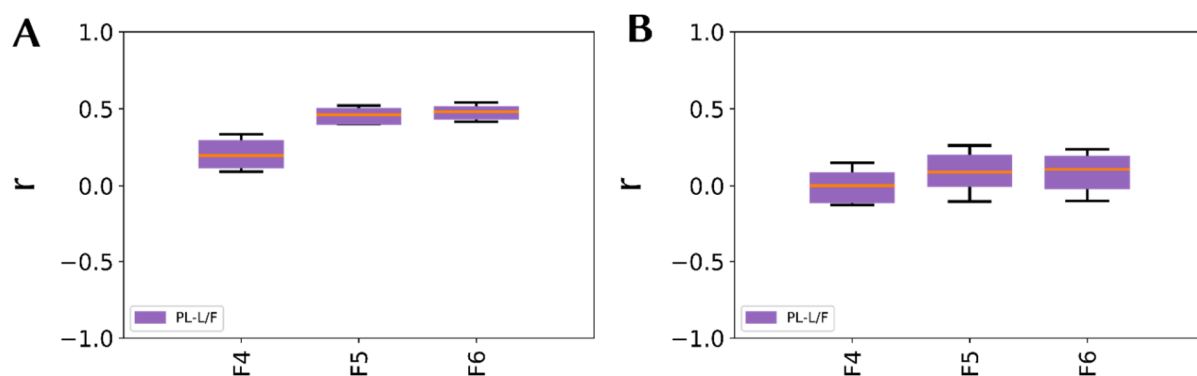


Figure S2-15: Boxplots showing correlation based on training data from five-fold cross validation for solvent functionals PL-L/F (with $F=\{F4,F5,F6\}$) using the actual datasets. **A:** Pearson r for the actual dataset; **B** Pearson r for the shuffled dataset.

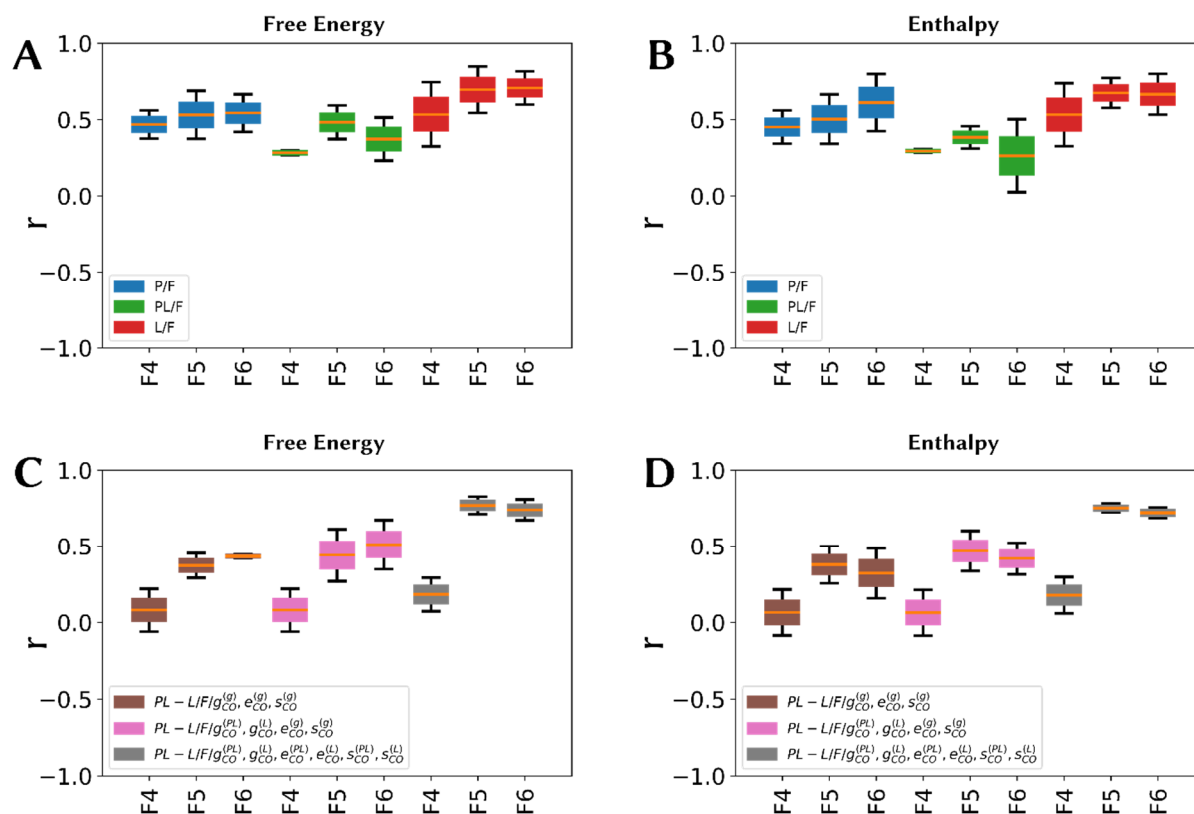


Figure S2-16: Boxplots showing the Pearson correlation coefficient based on training data from actual datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Free Energy and Enthalpy calculated with displaced-solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$); **C,D:** Free Energy and Enthalpy calculated with full binding-displacement treatment.

2.8.7 Correlation Statistics for Test Data Based on Shuffled Datasets

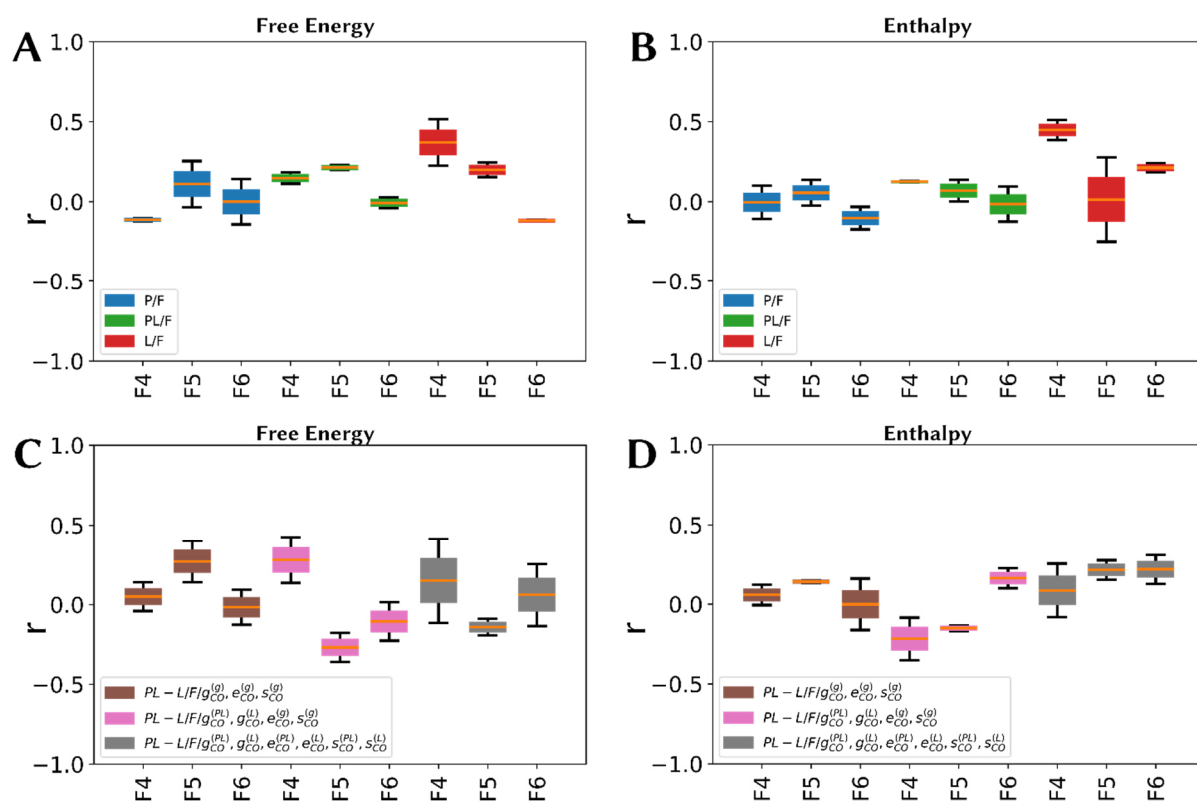


Figure S2-17: Boxplots showing the Pearson correlation coefficient based on test data from shuffled datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Free Energy and Enthalpy calculated with displaced-solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$); **C,D:** Free Energy and Enthalpy calculated with full binding-displacement treatment.

2.8.8 Correlation Statistics for Training Data Based on Shuffled Datasets

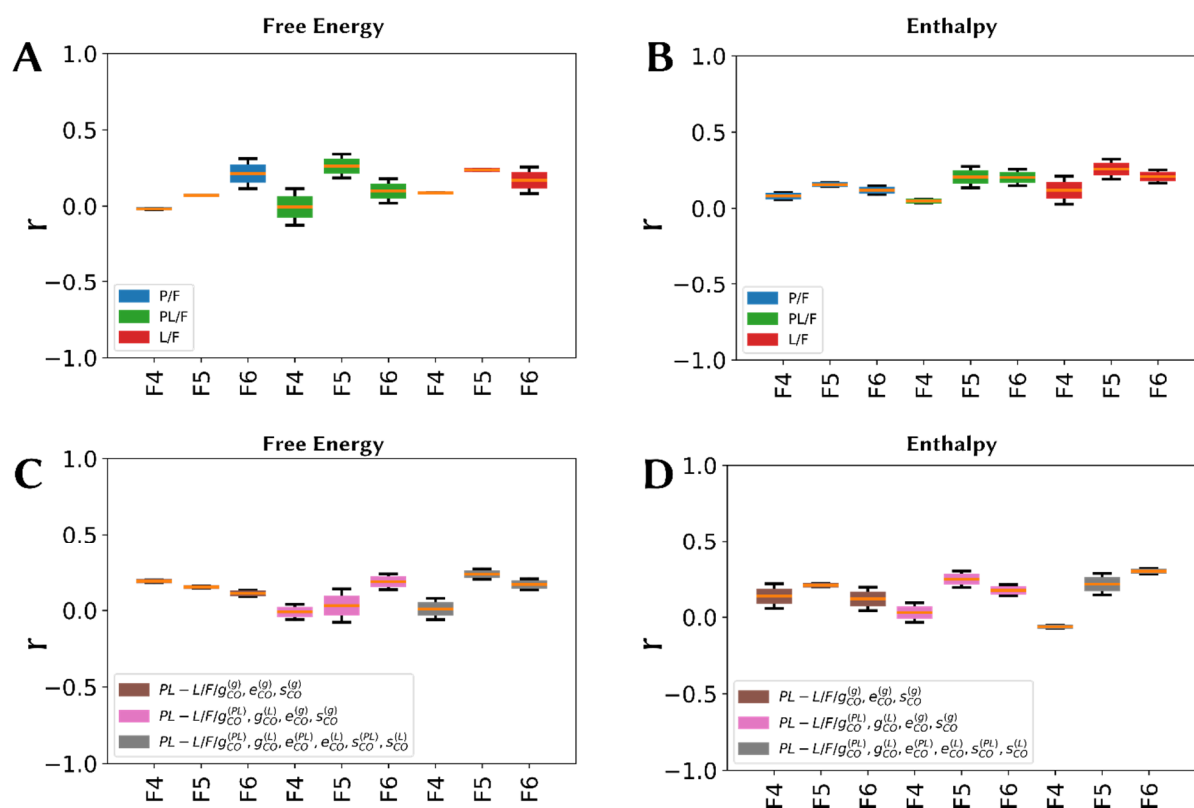


Figure S2-18: Boxplots showing the Pearson correlation coefficient based on training data from shuffled datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Free Energy and Enthalpy calculated with displaced-solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$); **C,D:** Free Energy and Enthalpy calculated with full binding-displacement treatment.

2.8.9 MUE statistics for actual and shuffled data for different solvent functionals

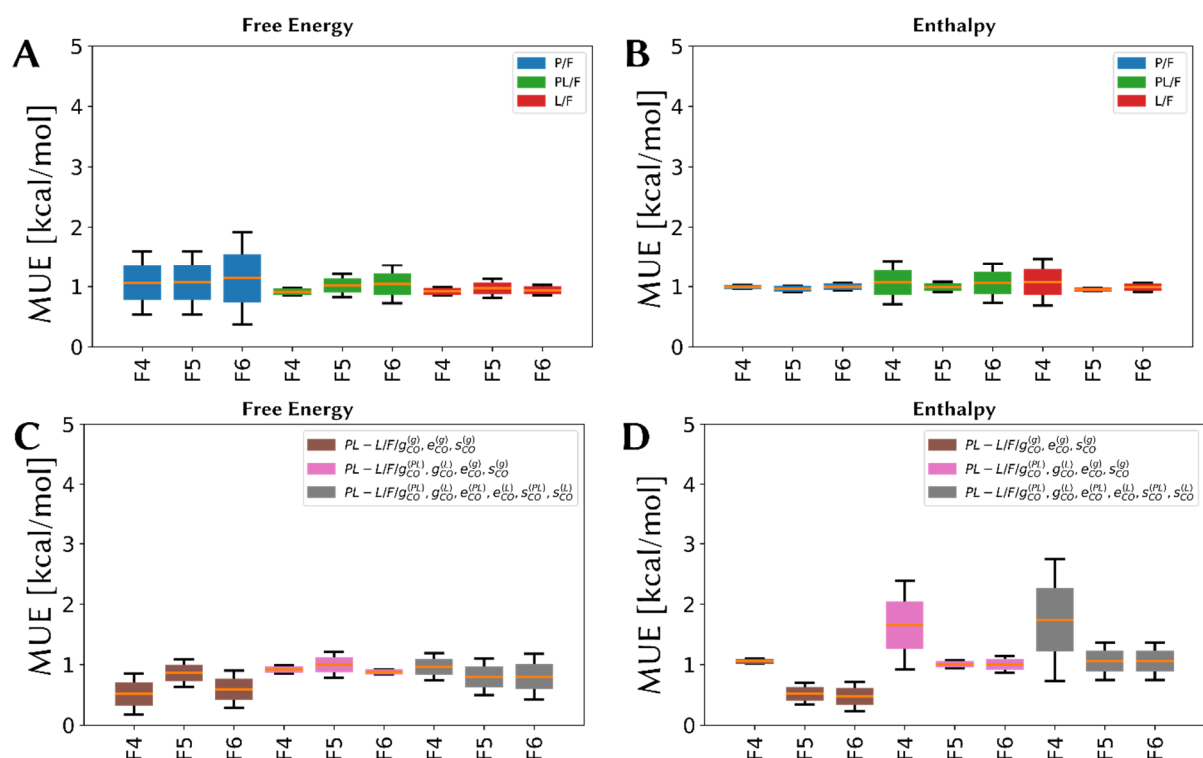


Figure S2-19: Boxplots showing the MUE based on test data from actual datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Free Energy and Enthalpy calculated with displaced-solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$); **C,D:** Free Energy and Enthalpy calculated with full binding-displacement treatment.

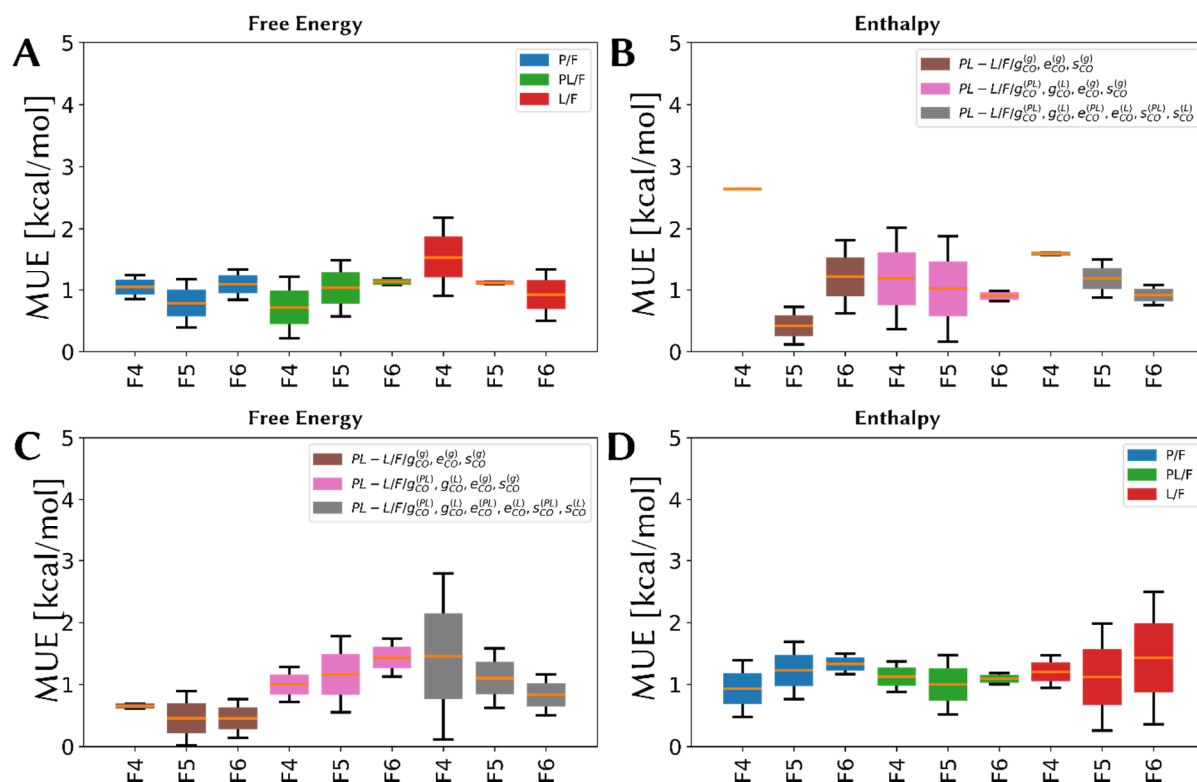


Figure S2-20: Boxplots showing the MUE based on test data from shuffled datasets. The inner box indicates the upper to lower quartile range, the whiskers indicate the lowest and highest datum that is still within 1.5 IQR. **A,B:** Free Energy and Enthalpy calculated with displaced-solvent functionals P/F, PL/F and L/F (with $F=\{F4,F5,F6\}$); **C,D:** Free Energy and Enthalpy calculated with full binding-displacement treatment.

3 Mapping Solvation Thermodynamics on Building Blocks: A Strategy to Design Better Binders

3.1 Abstract

The previously developed approach is applied in order to analyze the solvation thermodynamics of thrombin inhibitors with respect to individual building blocks. The building blocks are obtained by performing a virtual decomposition of the series of thrombin ligands that were already investigated in the previous chapter. For each of these building blocks, solvation thermodynamics are computed using molecular dynamics simulations, GIST and *Gips*. We find remote solvent structuring effects on the surface of an unbound ligand, which explains the experimentally determined differences in binding free energy. Furthermore, we demonstrate that fluorination of the building blocks has a huge influence on the desolvation energy of an unbound ligand molecule and thus explains an increased binding enthalpy value.

3.2 Introduction

During the binding of a ligand to its receptor, a complex process involving multiple intermediate steps is passed. Usually, most of these steps are hardly accessible and cannot be explored with sufficient detail by experimental techniques. In order to shed light on some of these hardly accessible steps, computer simulations have proven to be a valuable tool to enhance our understanding of association processes on the atomistic level. One of these, admittedly poorly understood steps during association are molecular solvation and desolvation processes of protein and ligand molecules. During the binding of a ligand to a receptor molecule, the ligand molecule sheds several layers of water molecules (see schematic depiction in Figure 3-1). Also, the protein binding pocket gets, depending on the situation before binding, partly or fully, dried upon the association of the ligand. Once the ligand molecule is accommodated in the binding pocket, water molecules are allowed to spatially rearrange in the binding pocket and finally, a new solvation shell around the formed complex is assembled. By that, the water molecules can interact with the protein and/or ligand molecules, but they can also oppose the binding process by adopting a less favorable arrangement in the formed protein-ligand complex. All these individual steps are associated with a contribution to the solvation free energy, and therefore also impact the enthalpy and entropy contributions to binding. Since these steps are determined by individual structural properties of the interacting species, they can be optimized and exploited to improve binding of a ligand to its receptor, in terms of affinity as well as with respect to selectivity. However, the molecular interactions established by water molecules are often poorly understood or difficult to visualize intuitively and therefore hard to predict. In this regard, local solvation effects on the surface of a formed protein-ligand complex as well as on the unbound ligand molecule prior to binding are often treated implicitly as local modulation of the dielectric constant, instead of considering explicit distributions of water molecules. From the perspective of high-resolution crystallography, it is well known that the spatial arrangement of water molecules on the surface of the formed protein-ligand complex has a clear impact on the thermodynamics of binding.^{48,96} By that, the contributions to solvation thermodynamics arising from the interactions of water molecules with the protein-ligand complex can be used to enhance the binding affinity of ligands to its receptor. Moreover, the explicit treatment of water molecules using molecular dynamics (MD) simulations allows to explicitly estimate the solvation thermodynamics from the ensemble of water molecules surrounding the ligand molecule in the complex as well as in solution prior to binding.

interactions formed to solvent molecules in aqueous solution prior to binding. Nevertheless, they can be determinant for the thermodynamic binding profile.⁴³ Interactions of the protein-solvent-ligand type can readily be optimized, since the routine usage of synchrotron radiation enables the exploration of high-resolution protein crystal structures. But also, in-depth understanding of these interactions using experimental techniques requires exhaustive coverage of chemical variations within a congeneric ligand series. Interactions of the ligand and water molecules prior to binding are usually not explored sufficiently - mainly because not many methods are accurate enough to capture solution ensembles. Nuclear magnetic resonance (NMR) spectroscopy techniques are usually the first choice in this context, however the locations of water molecules and their binding properties across the surface of an unbound ligand molecule are often not possible to explore on the NMR time-scale.

For the reasons mentioned above, and probably many more could be listed, drug discovery efforts do not consider structure-thermodynamics relationships simply as they are not easy to translate into design parameters that highlight contributions of water molecules in the binding process. Due to these missing considerations in the search for alternative scaffolds or the decoration of existing ones, we supposedly miss putative drug candidates during pre-clinical studies, which bind favorably due to their solvation and desolvation properties. Here, we propose a new strategy that is based on computer simulations and combines solvation thermodynamics with a kind of fragment-based drug discovery strategy. We will use our previously introduced solvent functionals in order to describe the solvation thermodynamics of the ligand in its protein-bound state as well as in its pre-bound state in aqueous solution. The spatial contributions found by these solvent functionals are readily decomposed by dismantling the original ligand into fragment-like substructures, called in the following building blocks (BB). They are generated by using chemically intuitive decomposition rules (Figure 3-2). In the current case, the peptidomimetic ligand scaffolds is cleaved along the various peptide bonds. The solvation properties of the BBs within the molecule are compared with respect to the properties of the entire ligand, as well as with the properties of a minimal representation of the building blocks (MRBB, here capping the splitted amide bonds) in aqueous solution (see Figure 3-2). As a system to test our model, we investigated ligands for the serine protease thrombin - a key factor in the human blood coagulation cascade for which are high-resolution crystal structures and experimentally determined thermodynamic profiles from isothermal titration calorimetry (ITC) or surface plasmon resonance (SPR) are available. In the present study, we highlight an unconsidered aspect of drug discovery, namely the correlation of thermodynamic

and structural data with solvation and desolvation properties of BBs and their mutually enhancement or loss in binding contributions due to solvation features.

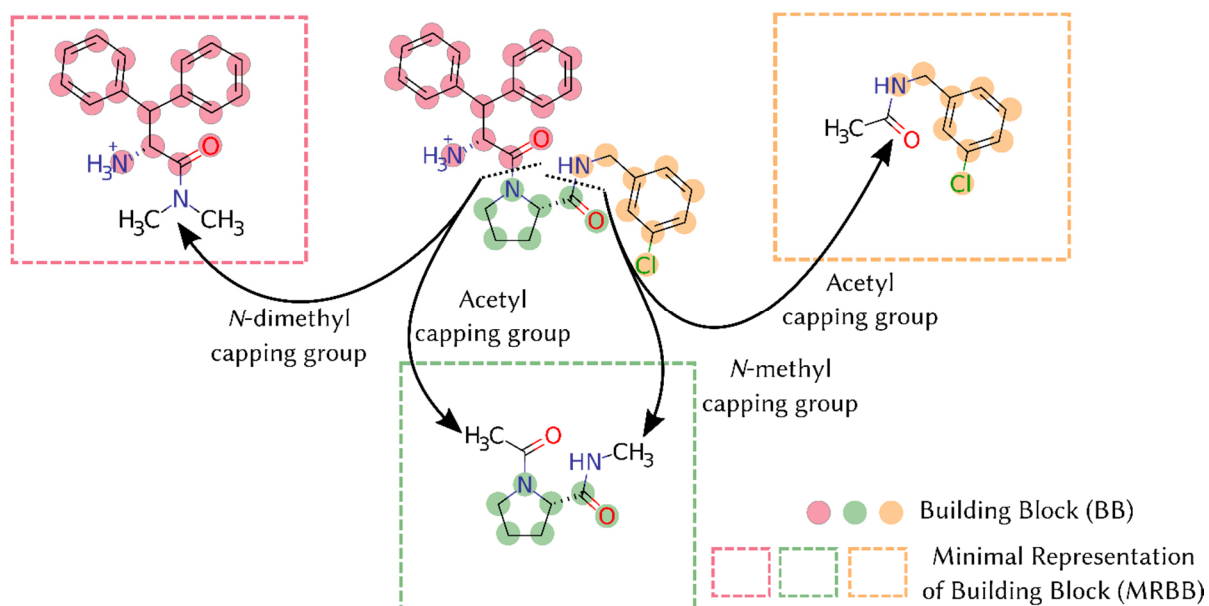


Figure 3-2: Overview of the building block decomposition strategy for a typical thrombin ligand (in this case, PDB code 3DHK¹⁰⁵). The ligand molecule is cleaved at the amide bonds in order to obtain BBs, which are essentially instructions for a topology-based decomposition of the molecules. The BBs can readily be capped with N-dimethyl (NDME), N-methyl (NME) or acetyl (ACE) capping groups in order to obtain MRBBs.

3.3 Results

In this study, we used a set of 53 thrombin ligands, which were determined by crystal structure analysis and corresponding binding thermodynamic data were available from ITC or SPR (48 with ITC data and 5 with SPR data). The ligands were combined into 186 matching pairs, such that the binding affinity between the ligand molecules in the pair is attributed to a difference in solvation or desolvation. This dataset was already used for the derivation of parameters for solvent functionals based on grid inhomogeneous solvation theory^{80,83,101} (GIST) and MD simulations in our previous contribution.

In this section, we will shortly summarize our previously presented GIST-based solvent functionals. Then, we will analyze the distribution of BBs across the dataset. In the subsequent main part of this results section, we will analyze the spatial decomposition of the solvent functionals based on the substructural BBs derived from the ligand molecules. The individual

contributions of the different BBs are compared across different ligands as well as with the corresponding capped BBs in aqueous solution.

3.3.1 GIST-based solvent functionals

The term *solvent functional* refers simply to a mathematical formulation of a function that uses the three-dimensional distribution of solvent free energy, entropy and solvent density relative to the bulk phase as independent variables. A solvent functional employs different parameters that are required to transform these distributions into scalar values describing the solvent free energy, enthalpy or entropy. The parameters are obtained by fitting the solvent functionals to experimental data. The three-dimensional distributions of solvent free energy, entropy and solvent density (see GIST^{80,83,101}) are calculated from the energies and spatial coordinates of water molecules found by molecular dynamics simulations of the protein-ligand complexes as well as the ligand molecules alone in aqueous solution. The solvent energy and entropy values are always reported as the difference to bulk water phase. Similarly, the solvent density evaluated across volume elements (so-called voxels) of a grid embedding the studied molecules is reported in terms of multiples of the mean bulk solvent density ρ^0 . Such derived solvent free energy, entropy and density distributions are discretized on the embedding grids that were centered on the studied molecules. Across these grids, only those grid voxels were considered showing a value that exceeded a predefined solvent energy and entropy threshold (cutoff parameters in the calculation of solvation energies and entropies). In addition, a solvent density cutoff value had to be defined, which permits only those grid voxels to be considered in the calculations that exceed a parameterized solvent density cutoff value. In other words, only grid voxel that are occupied by more than a previously-defined number of water molecules, corresponding to the solvent density cutoff value, are considered in the calculation of the solvent thermodynamics. The energy, entropy and density cutoff parameters are derived separately for the protein-ligand complex and the ligand molecule in aqueous solution. Details can be found in our previous contribution.

As defined by our solvent functional, only grid voxels around the protein-ligand complexes contribute to the protein-ligand solvation, if they are highly occupied ($>9.97 \rho^0$) by water molecules and exhibit unfavorable in solvent energy contributions ($>8.03 \text{ kcal} \cdot \text{mol}^{-1}$) compared to the mean energy value in bulk water phase. The solvent functional scores the contribution of these water molecules effectively as an energetically unfavorable quantity in the calculation of

the total solvation enthalpy. Consequently, these water molecules also contribute unfavorably to the solvation free energy of the formed protein-ligand complex. Furthermore, only those grid voxel around the protein-ligand complex that are considered to contain entropically unfavorable water molecules ($>7.83 \text{ kcal}\cdot\text{mol}^{-1}$) are evaluated as entropy contribution to the free energy of the protein-ligand complex desolvation. The solvent functional scores the placement of these water molecules into the protein-ligand complex as an unfavorable contribution to solvation free energy.

Only grid voxels around the ligand molecule found by the MD simulations of the unbound state that are highly populated ($>6.93 \rho^0$) by water molecules have been considered in the calculation. The energy of these water molecules is allowed to be slightly favorable ($>-0.95 \text{ kcal}\cdot\text{mol}^{-1}$), compared to their mean energy value in bulk water phase. However, most of the water molecules considered by this cutoff criterion will have an unfavorable ($>0 \text{ kcal}\cdot\text{mol}^{-1}$) energy contribution. From the perspective of desolvation entropy, only those grid voxel around the ligand in the case of the unbound situation, which contain entropically unfavorable water molecules ($>3.95 \text{ kcal}\cdot\text{mol}^{-1}$) are effectively considered. According to our solvent functional, the desolvation entropy of the ligand is scored as favorable contribution to the total free energy of the binding process. Consequently, this gain in binding free energy due to ligand desolvation is due to water molecules associated with the surface of the unbound ligand that are firmly fixed in terms of their translational and orientational degrees of freedom compared to bulk water phase.

Since the total contribution to the free energy of binding resulting from the protein-ligand complex solvation and the ligand desolvation is calculated as $\Delta G = \Delta G^{(PL)} - \Delta G^{(L)}$, the desolvation of the unbound ligand molecule is considered as a favorable contribution to solvation free energy (both in terms of energy and entropy) and the solvation of the protein-bound ligand molecule is considered as an unfavorable contribution to solvation free energy (also, both in terms of energy and entropy). For a comprehensive interpretation of the parameters of the solvent functional, please see the Supporting Information.

3.3.2 Distribution of Building Blocks across the Dataset

The BBs are generated by splitting the amide as well as sulfonamide bonds in the ligand molecules (see Figure 3-2; for a more comprehensive description, see Methods section). Subsequently, a total of 58 unique BBs and MRBBs was obtained. This number is further

reduced by keeping only those matching pairs of ligand molecules, and accordingly their corresponding BBs, that mutually differ by only one single BB. The resulting final library contained 44 BBs distributed over 125 pairs of ligand molecules. The BBs are devised into different groups according to their location in the protein binding pocket. Since the target protein thrombin is a trypsin-like protease, we applied the *Schechter* and *Berger* nomenclature¹³⁸ in order to classify each BB with respect to its sub-pocket occupancy. Accordingly, the sub-pockets are assigned as S_1 , S_2 , S_3 and S_A (A =Aryl, s. below) and the occupying portions are designated as P_1 , P_2 , P_3 and P_A (Figure 3-3). The nomenclature for the sub-pocket occupancy of S_1 - S_3 is derived from the positions at which the natural substrate (fibrinogen) accommodates its amino acid side chains next to the cleavage site. We named the so-called aryl binding pocket “ S_A ”, which is the fourth pocket, often also designated as $S_{3/4}$ pocket. It is populated by the distal Phe and Leu side chains (P_8/P_9) of the natural substrate.¹⁴⁹

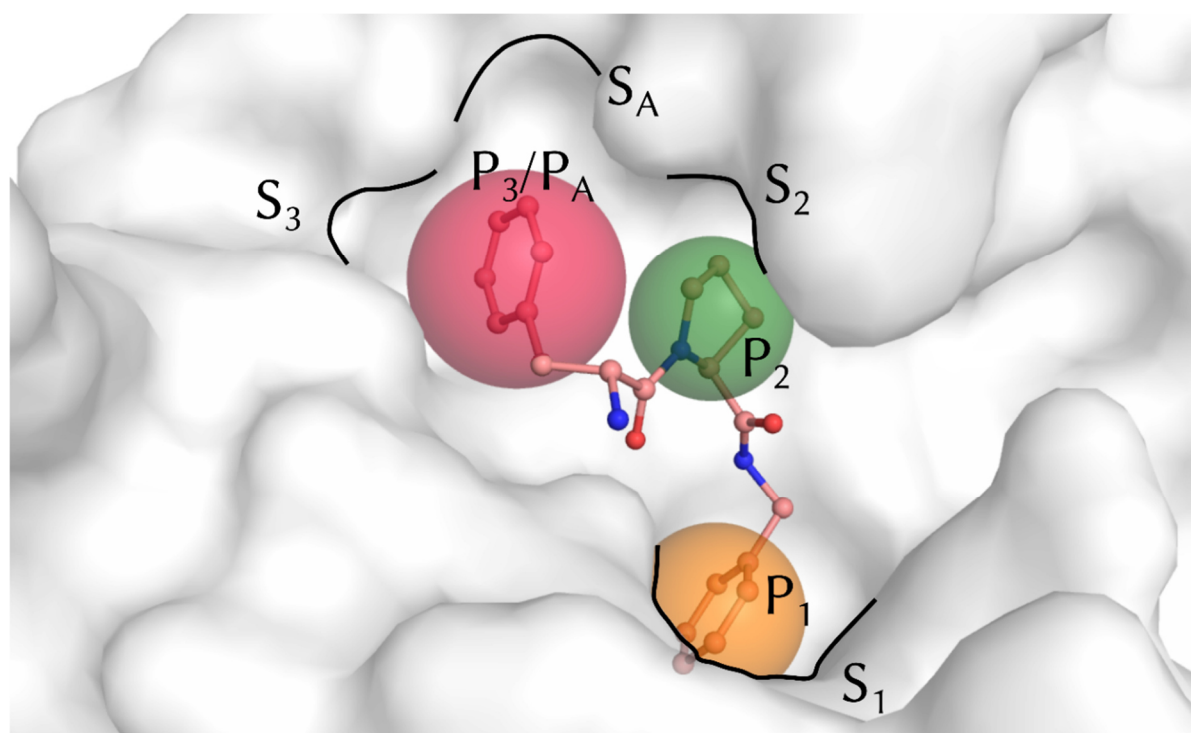


Figure 3-3: Binding pocket of thrombin (PDB 2ZFF¹⁰⁴) with sub-pocket annotation according to *Schechter* and *Berger*¹³⁸.

As can be seen from the overview of all BBs in our dataset (Figure 3-4), most of the variations in the ligand series was introduced by varying the P_1 head groups occupying the S_1 pocket. The P_2 , P_3 and P_A portions have been modified less throughout the dataset. Most notably, **B1** (*L*-Proline) is the most widely used P_2 portion for thrombin inhibitors. From the investigated 53

ligand molecules, 47 contained this BB at position P₂. The only other P₂-type BB is **B29** with five occurrences. From all BBs at positions P₃/P_A, **B0** (*D*-Phenylalanine) is the one which is most frequently used with 22 occurrences in the 53 ligand set.

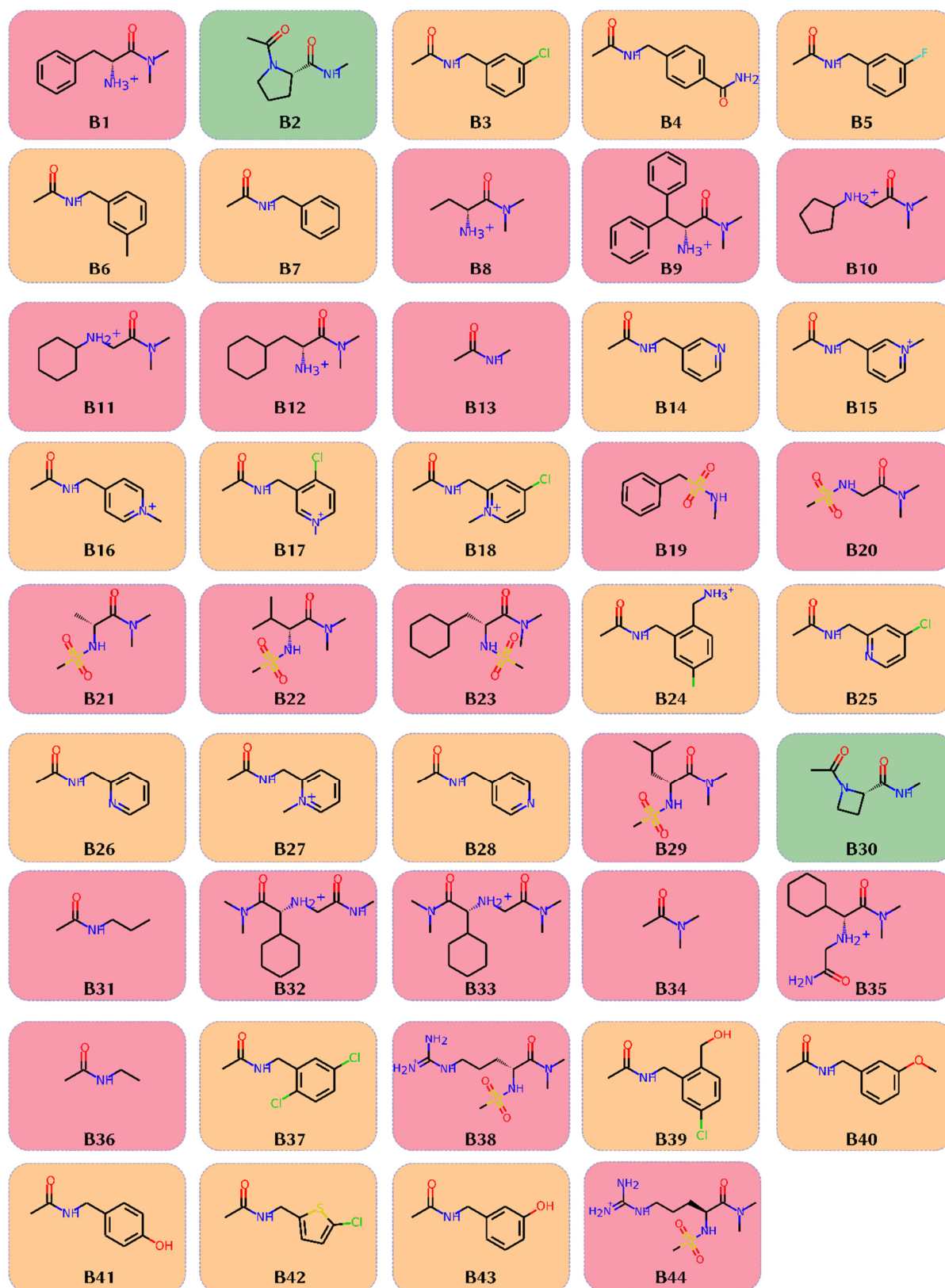


Figure 3-4: Overview of all BBs with capping groups attached. The coloring is accordance with the sub-pocket annotation: P₁ portion, orange; P₂ portion, green; P₃/P_A-portion, red.

3.3.3 Building Block Solvent Thermodynamics

As can be seen in Figure 3-5, the average thermodynamic contributions of the BBs in the bound as well as the unbound ligands vary greatly. Some tend to have strong solvent free energy signatures in the unbound state of the ligand, but do not contribute to a gain in solvent free energy in the bound complex (e.g. **B9**). Some others have almost similar contributions in the bound as well as the unbound state of the ligand and therefore have compensating thermodynamic signatures (e.g. **B7**).

One example for a BB with a deviating thermodynamic signature in the bound and unbound state is **B1**. The calculated average solvation free energy for this BB is $2.0 \text{ kcal}\cdot\text{mol}^{-1}$ in the unbound state (Figure 3-5B), but only $0.2 \text{ kcal}\cdot\text{mol}^{-1}$ in the bound state (Figure 3-5A). Interestingly, the MRBB of **B1** (Figure 3-5C) reveals a solvation free energy of $0.7 \text{ kcal}\cdot\text{mol}^{-1}$, which is inbetween the bound and the unbound form. Thus, cooperative effects resulting from other BBs in the unbound state have a large influence on this apolar BB. But also in the bound state when this BB is embedded in the ligand, interactions with the protein likely compensate for the difference between the MRBB and the bound ligand. Another related BB is **B9**, which is similar to **B1**, except that it has one additional phenyl group attached to the benzylic methylene group in P_A . The average solvation free energy of **B9** is $3.3 \text{ kcal}\cdot\text{mol}^{-1}$ in the unbound state and $0.2 \text{ kcal}\cdot\text{mol}^{-1}$ in the bound state. Remarkably, the calculated solvation free energy of the MRBB of **B9** is $1.0 \text{ kcal}\cdot\text{mol}^{-1}$ and therefore close to the value found for the related **B1**. These two BBs are found in thrombin ligands **1** and **2** (see Figure 3-6), which have an experimentally measured difference in binding free energy ($\Delta G(\mathbf{1}\rightarrow\mathbf{2}) = \Delta G(\mathbf{2}) - \Delta G(\mathbf{1})$) of $-1.0 \text{ kcal}\cdot\text{mol}^{-1}$ (see also Table 3-2 for a comprehensive overview). The calculated relative free energy difference for this pair is $-1.7 \text{ kcal}\cdot\text{mol}^{-1}$. In **1**, the value of the solvation free energy for **B1** in the unbound state was $2.3 \text{ kcal}\cdot\text{mol}^{-1}$ and the calculated value for the solvation free energy of **B9** in the unbound state of ligand **2** was $3.1 \text{ kcal}\cdot\text{mol}^{-1}$.

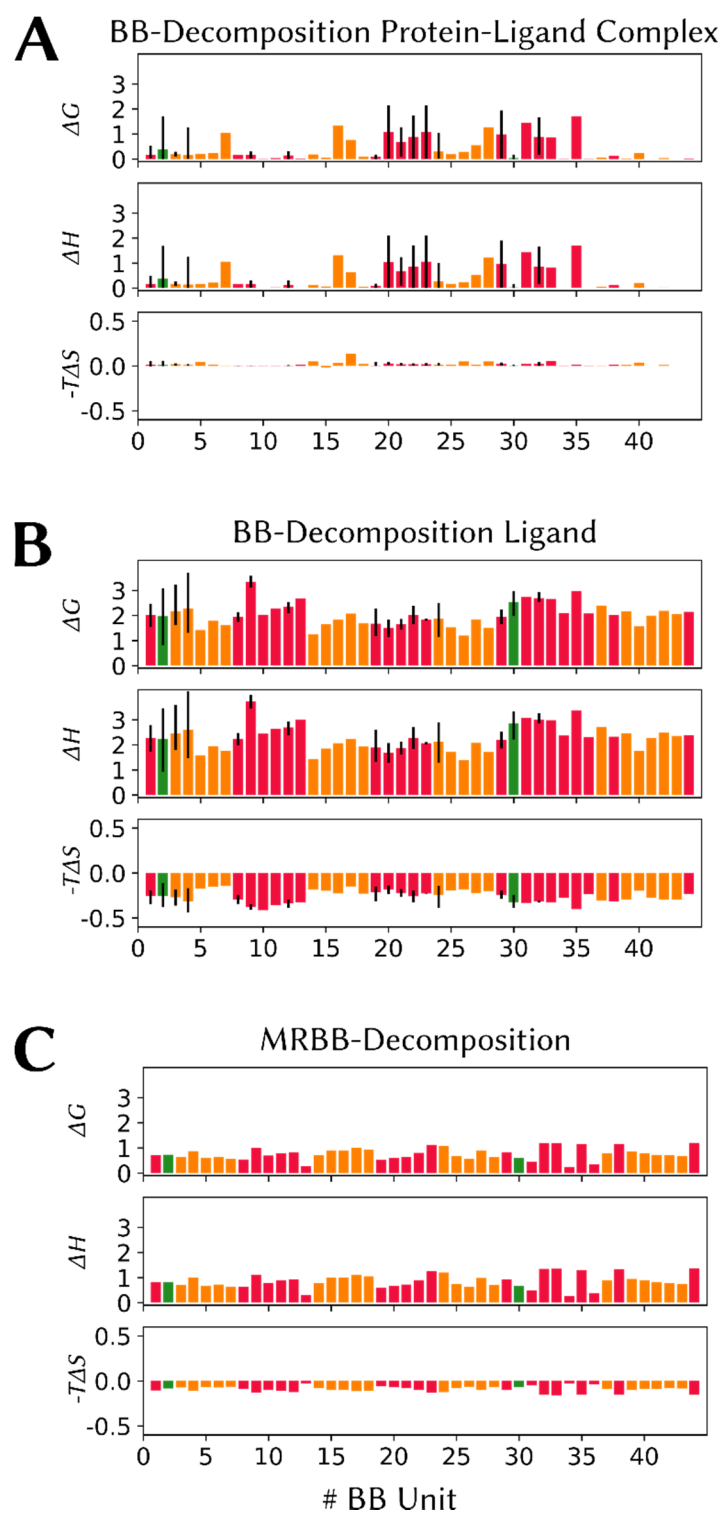
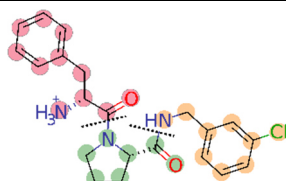
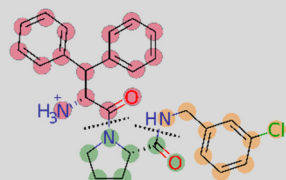
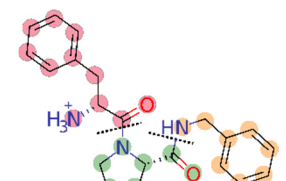
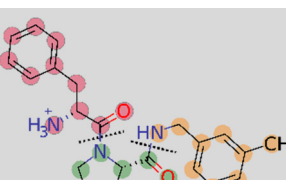
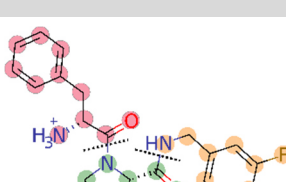


Figure 3-5: Overview of the average BB contributions to the protein-ligand complex solvation (A), ligand solvation (B) and MRBB molecule (C) in aqueous solution. The lines assigned to the bars indicate the observed value range of the BB contributions across all ligand molecules. The color of the bar encodes the assignment of the BBs to their position in the ligand (P₁, P₂ and P₃/P_A) and is in accordance with the previous figures. The units on the y-axis are in kcal·mol⁻¹.

Table 3-1: BB free energy decomposition for the ligands discussed in this work.

Ligand ^{a)}	BB P ₃ /P _A ^{b)}		BB P ₂ ^{b)}		BB P ₁ ^{b)}		
	ΔG	ΔH	ΔG	ΔH	ΔG	ΔH	
1 (2ZC9) 	B1		B2		B3		
	^{c)} bound	0.2±0.1	0.2±0.1	0.2±0.1	0.2±0.1	0.2±0.1	0.2±0.1
	^{d)} unbound	2.3±0.7	2.6±0.9	2.3±0.7	2.6±0.8	1.6±0.5	1.8±0.5
2 (3DHK) 	B9		B2		B3		
	bound	0.3±0.2	0.3±0.2	0.3±0.2	0.3±0.2	0.3±0.1	0.3±0.1
	unbound	3.1±0.9	3.5±1.0	1.9±0.5	2.1±0.6	3.2±0.9	3.6±1.0
3 (2ZFF) 	B1		B2		B7		
	bound	0.1±0.0	0.1±0.0	0.1±0.1	0.1±0.0	1.1±0.4	1.1±0.4
	unbound	1.9±0.5	2.1±0.6	2.0±0.6	2.2±0.6	1.6±0.5	1.8±0.5
4 (2ZF0) 	B1		B2		B6		
	bound	0.3±0.2	0.3±0.2	0.2±0.2	0.2±0.2	0.2±0.2	0.2±0.2
	unbound	2.3±0.7	2.6±0.8	2.4±0.7	2.6±0.8	1.8±0.6	1.9±0.6
5 (2ZDV) 	B1		B2		B5		
	bound	0.2±0.1	0.2±0.1	0.2±0.1	0.2±0.1	0.2±0.2	0.2±0.2
	unbound	1.9±0.6	2.1±0.7	1.9±0.6	2.1±0.7	1.4±0.4	1.6±0.5

- a) Ligand molecule and corresponding pdb code in parenthesis. The coloring of the BBs in the 2d depictions are in accordance with the color code from the previous figures (P₁ portion, orange; P₂ portion, green; P₃/P_A-portion, red).
- b) All units are in kcal·mol⁻¹. Error indicates 1 standard deviation from the mean estimated from the test set results of 10 random repetitions of 5-fold cross-validation (see our previous contribution).
- c) Free energy contribution of this BB in the *bound* state of the ligand.
- d) Free energy contribution of this BB in the *unbound* state of the ligand.

As can be seen from the solvation free energy maps for the MRBB of **B1** and **B9** (see Figure 3-6 C and F, respectively), the additional phenyl moiety in **B9** leads to further, highly populated, water positions on top of both phenyl moieties. Furthermore, populated regions close to the protonated amino group are observed. Both solvation features are perfectly mirrored in the unbound state of ligand molecules **1** and **2** (see Figure 3-6 B and E, respectively) but they get

lost upon protein binding (see Figure 3-6 A and D). However, as the difference between **B1** and **B9** in the unbound state amounts only to $0.8 \text{ kcal}\cdot\text{mol}^{-1}$ and only to $0.1 \text{ kcal}\cdot\text{mol}^{-1}$ in the bound state, these BBs alone cannot constitute the major contribution to the total calculated difference in binding free energy of $-1.7 \text{ kcal}\cdot\text{mol}^{-1}$ between **1** and **2** (see Table 3-2). Quite unexpectedly, it is **B3** (the P₁ portion), the BB that is found identically in both ligand molecules, that contributes major part of the difference in solvation free energy. The calculated free energy contribution of **B3** in the unbound state of **1** is $1.6 \text{ kcal}\cdot\text{mol}^{-1}$ whereas in the unbound state of **2** it contributes $3.2 \text{ kcal}\cdot\text{mol}^{-1}$. Thus, the contribution of **B3** is twice as large in the unbound state of **2** than in the unbound state of **1**. It becomes greatly enhanced compared to its corresponding MRBB ($0.7 \text{ kcal}\cdot\text{mol}^{-1}$). In this MRBB, the BB accommodates water molecules on top of both faces of the *m*-chlorophenyl ring (see Figure 3-7). The accommodation of these water molecules is further enhanced by the presence of the amide group which promotes the formation of an open ring-like solvent density distribution encompassing the N-H and the C=O groups. This solvent density is partly retained once **B3** is embedded into ligand **1** or **2**. In the bound state, the calculated free energy contribution of **B3** is similar for both ligands ($0.1 \text{ kcal}\cdot\text{mol}^{-1}$) due to their similar water structure in the S1 sub-pocket.¹⁰⁵ The difference between **1** and **2** in the unbound state is caused by the enhanced stability of water molecules on top of **B3** in ligand **2**. These water molecules are more efficiently entrapped by **2** due to the presence of the second phenyl ring in **B9**, which is not present in **B1** in case of **1** (Figure 6B and E). Thus, the contribution of **B3** in P₁ position is dominated by the remote solvent structuring induced by **B1** vs. **B9** in the P₃/P_A position.

Furthermore, in the bound state of **1** and **2**, additional water molecules become energetically entrapped unfavorably (compared to bulk water phase) beneath the side chain of Trp60D from the 60s loop (Figure 3-6A, D). However, the contribution of waters at this site are similar for both ligands and thus cannot contribute to the difference in affinity. Interestingly, in the crystal structure of the *apo* form of the protein a water molecule is observed at the position below Trp60D (see Figure 3-8A), which, however, is absent in the crystal structure with **2** (see Figure 3-8B). In this structure, no direct interactions of this water molecule or Trp60D with other symmetry-related crystal mates in the solid state packing were observed. Nonetheless, our observation might still indicate that the water molecule is missing in the crystal structure of bound **2** due to the crystal environment or the cryogenic conditions during the experiment.

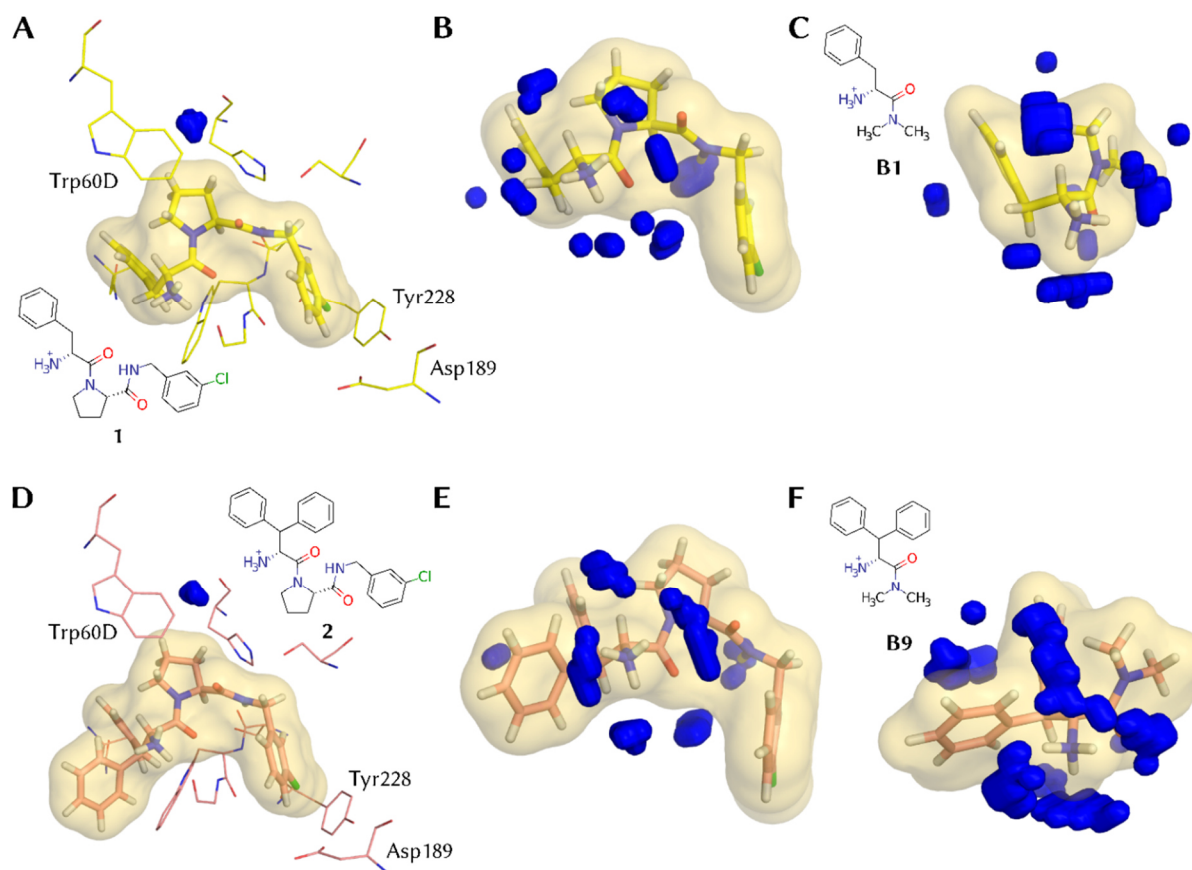


Figure 3-6: Solvent free energy maps for the thrombin complexes of **1** (PDB 2ZC9)¹⁰⁴ and **2** (PDB 3DHK)¹⁰⁵, the unbound ligands and MRBBs **B1** and **B9**. **A, D:** Protein-ligand complex of ligands **1** and **2** with corresponding solvent free energy maps, which were generated with energy and density cutoff values of $e_{CO}^{(PL)} = 8.03 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(PL)} = 9.97 \rho^0$, respectively. **B, E:** Ligand molecules **1** and **2** with corresponding solvent free energy maps, generated with energy and density cutoff values $e_{CO}^{(L)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(L)} = 6.93 \rho^0$, respectively. **C, F:** solvent free energy maps for MRBB **B1** and **B9**, generated with energy and density cutoff values of $e_{CO}^{(MRBB)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(MRBB)} = 3.0 \rho^0$. The displayed conformations are cluster representatives of the most populated cluster for the conformational ensemble of **B1** and **B9** (71.0% and 99.0% occupancy, respectively).

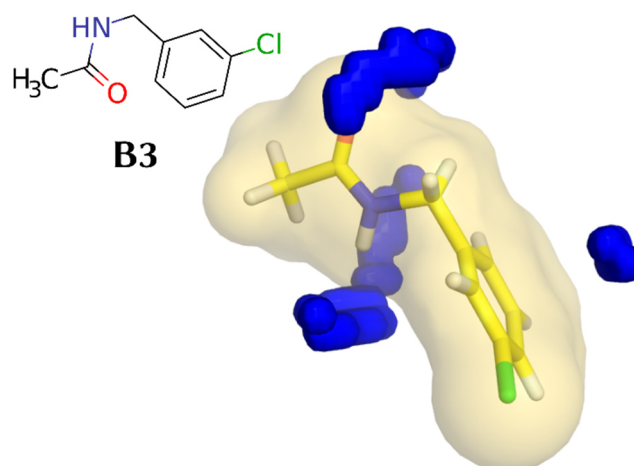


Figure 3-7: Solvent free energy map for MRBB **B3** generated with cutoff values $e_{CO}^{(MRBB)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(MRBB)} = 3.0 \rho^0$. The displayed conformer is the cluster representative of the most populated cluster (51.6% occupancy) for the conformational ensemble of MRBB **B3**.

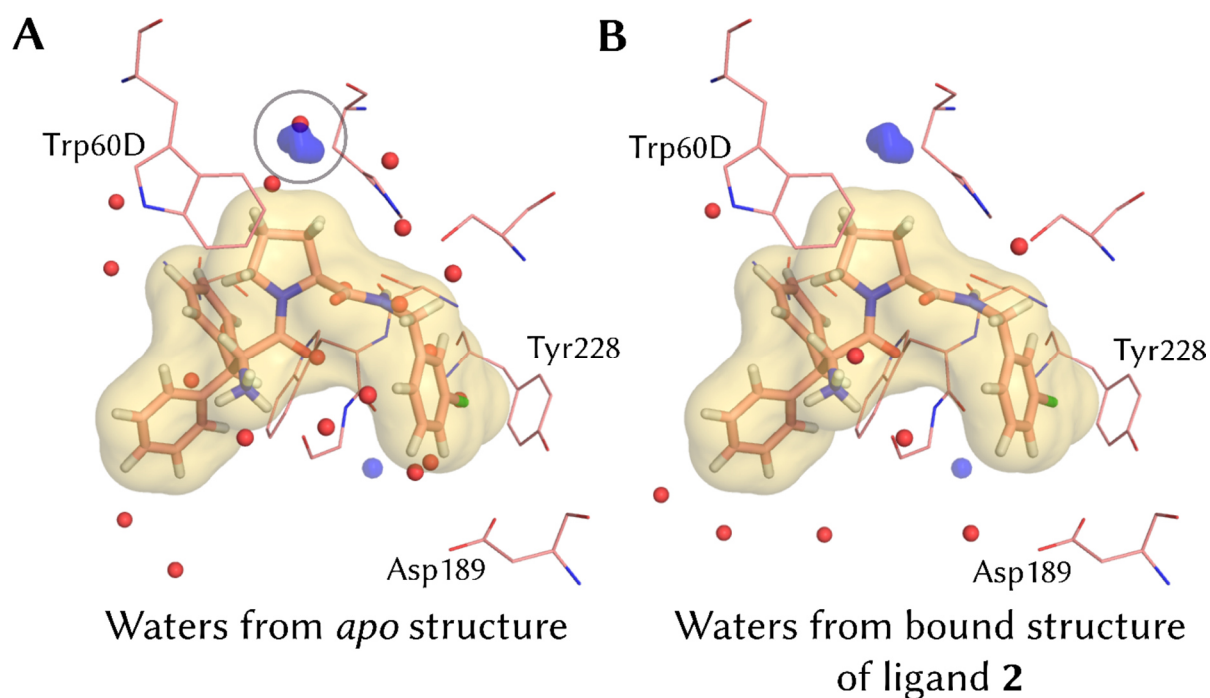


Figure 3-8: Solvent free energy map of the complex with ligand **2** superimposed with the water molecules (red spheres) found in the crystal structure of the *apo* form of thrombin (**A**) and found in the crystal structure with **2** (**B**). The grey circle (left) highlights the water molecule beneath Trp60D from the *apo* structure, which perfectly matches with the computed solvent free energy map. The maps were generated with energy and density cutoff values for the ligand molecule $e_{CO}^{(L)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(L)} = 6.93 \rho^0$, respectively.

In a second example, the difference in solvation free energy is dominated by the difference in solvation that is attributed to the P₁ portion. This portion was subject of many drug optimization efforts in the context of trypsin-like proteases, since the P₁ portion occupies the S₁ sub-pocket, which is responsible for selectivity discrimination either of substrates but also of developed inhibitors. As can be seen from Figure 3-5A and Table 1 (last column), the P₁ building block **B7** is one of the few BBs which has compensating free energy contributions in the bound and unbound state. In contrast, the structurally related P₁ building blocks **B5** and **B6** have negligible contributions in the bound state of the ligand (both 0.2 kcal·mol⁻¹) and mainly contribute through the unbound state of the ligands (1.4 kcal·mol⁻¹ and 1.8 kcal·mol⁻¹, respectively) to the value of the total free energy. The calculated differences in solvent thermodynamics between the BBs cannot be attributed to unique properties of the BBs themselves, as the solvation free energy of the MRBBs are quite similar in these cases (**B5** 0.6 kcal·mol⁻¹ ; **B6** 0.7 kcal·mol⁻¹; **B7** 0.6 kcal·mol⁻¹). The calculated solvation free energy for **B7** in the bound state of **3** is 1.1 kcal·mol⁻¹, which is close to the calculated value of 1.6 kcal·mol⁻¹ for the unbound state of this BB. The unfavorable solvation free energy of ligand **3** in the unbound state is due to the energetically unfavorable interaction of a water molecule trapped between Tyr228 and the phenyl moiety (see Figure 3-9A). Energetically frustrated water molecules in the vicinity of the phenyl ring of **B7** are also found in the simulations of the unbound state of **3** (see Figure 3-9B), thus we experience almost compensating contributions for both states. The free energy contribution of the related **B3**, the *m*-chloro derivative, has the same free energy contribution, 1.6 kcal·mol⁻¹, as **B7** in the unbound state but only 0.2 kcal·mol⁻¹ in the bound state. Thus, the calculated free energy difference for **3**→**1** is -1.4 kcal·mol⁻¹. Although the calculated value is not within the experimental error range (-0.9±0.2 kcal·mol⁻¹), our model successfully identifies **1** as the more affine ligand in this comparison.

The related ligand **4** bears a methyl group at *meta* position of the phenyl ring as part of the P₁ portion. The corresponding **B6** behaves quite similar to its *meta*-chloro analogue **B3** with respect to the solvation free energy of the MRBB (0.7 kcal·mol⁻¹ and 0.6 kcal·mol⁻¹, respectively), but also with respect to the contributions to the bound and unbound state of the ligands. The calculated contributions of **B6** to the free energy of **4** are 0.2 kcal·mol⁻¹ and 1.8 kcal·mol⁻¹ in the bound and unbound state, respectively. The missing unfavorable contribution in the bound state of **4** (as compared to **3**) is due to the lack of (trapped) water molecules between Tyr228 and the phenyl portion of **B6** (see Figure 3-9D). The calculated difference for the comparison **3**→**4** is -1.6 kcal·mol⁻¹ and thus quite similar to the one calculated

for the comparison **3**→**1** (-1.4 kcal·mol⁻¹). Within the error range of the calculation it would not be possible to decide which of the ligands, **1** or **4**, is more potent than **1**. However, the experimental uncertainties of the relative differences for **3**→**1** and **3**→**4** are also too high in order to effectively discriminate between ligands **1** and **4**.

Table 3-2: Relative free energies for some protein-ligand binding reactions.

Ligand comparison	$\Delta\Delta G_{\text{Calc}}^{\text{a) b)}$	$\Delta\Delta H_{\text{Calc}}^{\text{a) b)}$	$\Delta\Delta G_{\text{Exp}}^{\text{a) c)}$	$\Delta\Delta H_{\text{Exp}}^{\text{a) c)}$
1 ^{d)} → 2 ^{d)}	-1.7±0.5	1.9±0.5	-1.0±0.2	-2.0±0.4
3 ^{e)} → 1 ^{d)}	-1.4±0.7	-1.7±0.8	-0.9±0.2 ^{d)}	-5.6±0.4
3 ^{e)} → 4 ^{e)}	-1.6±0.4	-1.6±0.5	-0.7±0.2	-3.6±0.3
3 ^{e)} → 5 ^{e)}	-0.4±0.5	-0.6±0.4	+0.1±0.1	+0.1±0.3
1 ^{d)} → 5 ^{e)}	+1.0±0.4	+1.1±0.5	+1.0±0.2	+5.7±0.4

a) All units are in kcal·mol⁻¹. Error given as 1 standard deviation from the mean.

b) Standard deviation estimated from the test set results of 10 random repetitions of 5-fold cross-validation (see our previous contribution).

c) Standard deviation estimated from triplicate ITC measurements and error propagation. However, for **3** the standard error for the free energy and enthalpy was estimated to 0.12 (0.5 kJ·mol⁻¹) and 0.24 kcal·mol⁻¹ (1 kJ·mol⁻¹), respectively, since here no standard error from triplicate measurements was available.

d) Reference¹⁰⁵

e) Reference¹⁰⁴

The solvation free energy of the *m*-fluoro substituted MRBB **B5** is 0.6 kcal·mol⁻¹ and thus similar to the value found for the *m*-chloro substituted BB **B3**. Furthermore, the shape of the solvent density of the two MRBBs is virtually identical (cf. Figure 3-7 and Figure 3-9 I). Building block **B5** is embedded into **5**, which has an experimental free energy of binding that is indistinguishable from the one measured for **3** ($\Delta\Delta G_{\text{Exp}}(\mathbf{3}\rightarrow\mathbf{5}) = 0.1\pm 0.1$ kcal·mol⁻¹). Our calculations also confirm this observation ($\Delta\Delta G_{\text{Calc}}(\mathbf{3}\rightarrow\mathbf{5}) = 0.4\pm 0.5$ kcal·mol⁻¹), albeit with a greater range of error compared to the experiment. Furthermore, our model accurately calculates the experimental difference between **5** and **1** ($\Delta\Delta G_{\text{Exp}}(\mathbf{1}\rightarrow\mathbf{5}) = 1.0\pm 0.2$ kcal·mol⁻¹) as $\Delta\Delta G_{\text{Calc}}(\mathbf{1}\rightarrow\mathbf{5}) = 1.0\pm 0.4$ kcal·mol⁻¹. The reason for the low calculated binding affinity of **5** is its high desolvation penalty: It lacks energetically unstable water molecules in its unbound state and thus experiences a loss in solvation free energy of 0.4 kcal·mol⁻¹, 0.4 kcal·mol⁻¹ and 0.2 kcal·mol⁻¹ at the P₃/P_A (**B1**), P₂ (**B2**) and P₁ (**B7**→**B3**) portions compared to ligand **1**. Finally, the solvation pattern of the unbound ligand **5** (see Figure 3-9H) suggests that fewer energetically unstable water molecules seem to occupy the region between the ammonium group and the P₁ portion compared to **3** and **4**. Thus, the fluorinated ligand **5** reduces the number

of energetically unfavorable water molecules (compared to bulk water phase) on the surface of the unbound ligand molecule and is therefore more expansive to desolvate.

The less favorable desolvation of the fluorinated ligand **5** is most likely due to an enhanced bond dipole of the C-F bond compared to the C-Cl bond (see Figure 3-10C and D) and the enhanced bond dipole of the neighboring C-H bonds (cf. Figure 3-10A, C and D). According to our calculation of partial charges based on the RESP method, the carbon atom attached to the fluorine atom has a charge of +0.20 charge units, whereas the carbon atom attached to the chlorine atom has a charge of -0.04 charge units. For the halogen atoms, fluorine has a charge of -0.19 and chlorine atom -0.12. Thus, the fluorinated ligand **5** is expected to be engaged in more stable (hydrogen bond-like) interactions with the surrounding solvent molecules than the chlorinated derivative **1**. This is also emphasized by energetically more favorable ($-2.2 \text{ kcal}\cdot\text{mol}^{-1}$) solute-water interactions (based on the raw interaction energies extracted from the force field) in the first solvation layer of **B3** compared to **B5**. In addition, the raw water-water interactions (based on the raw interaction energies extracted from the force field) in the first solvation layer of **B5** are more favorable ($-3.3 \text{ kcal}\cdot\text{mol}^{-1}$) compared the ones of **B3**. This effect is likely not only due to the enhanced electrostatic interactions of the fluorinated species, but also due to the smaller atomic volume of the fluorine atom compared to chlorine. This difference opposes more favorable water-water interactions experienced by the water molecules in the first hydration layer of the fluorinated **B5** compared to the chlorinated **B3**. The increase in solvation free energy for fluorinated phenyl moieties compared to their non-fluorinated analogues was already studied using quantum chemical calculations.¹⁵⁰

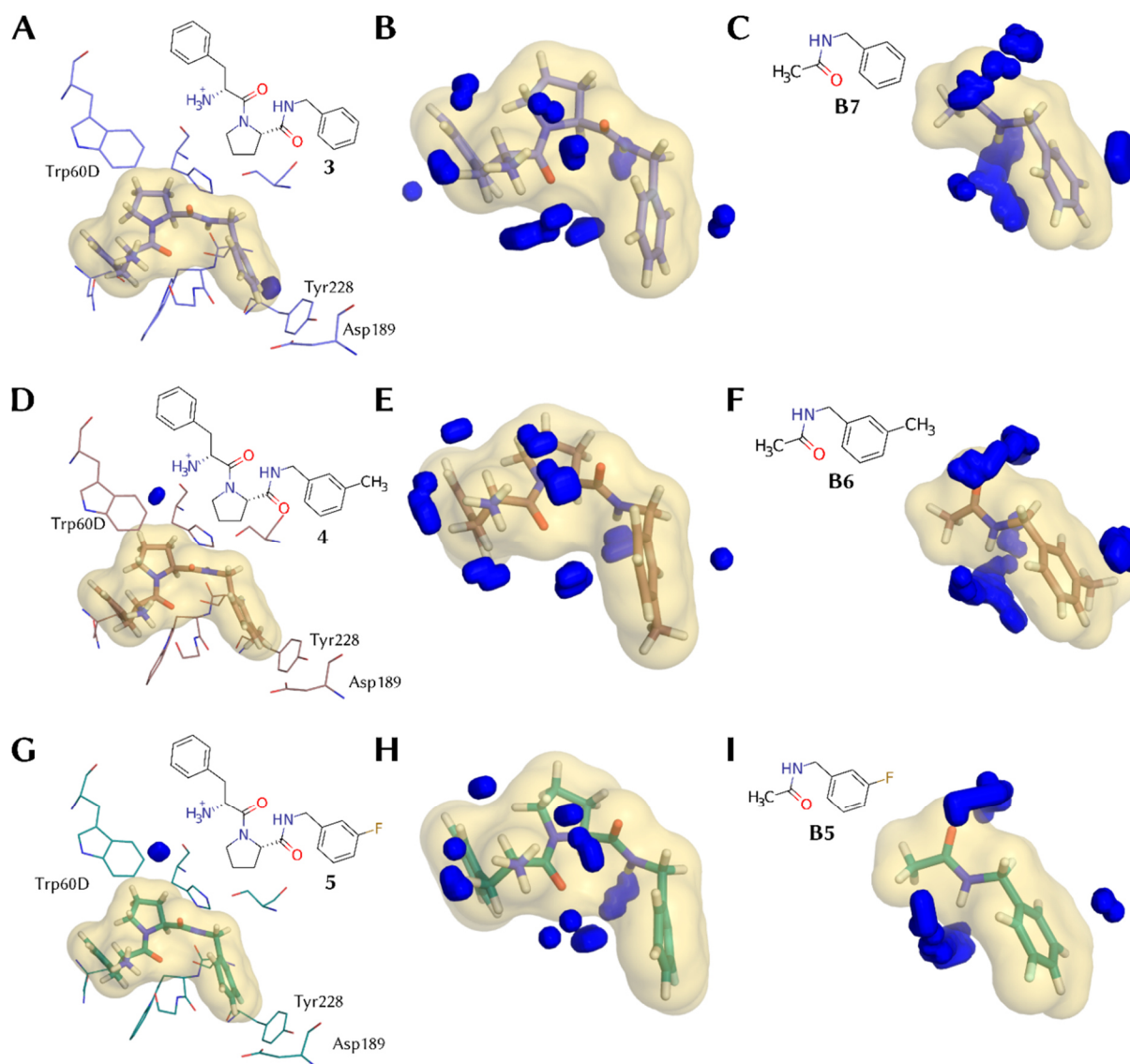


Figure 3-9: Solvent free energy maps for MRBBs of **B7**, **B6** and **B5** together with the maps found for the thrombin complexes with **3** (PDB 2ZFF), **4** (PDB 2ZF0) and **5** (PDB 2ZDV). **A**, **D**, **G**: Protein-ligand complex of ligands **3**, **4** and **5** with corresponding solvent free energy maps. The maps were generated with energy and density cutoff values of $e_{CO}^{(PL)} = 8.03 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(PL)} = 9.97 \rho^0$; **B**, **E**, **H**: Ligand molecules **3**, **4** and **5** with corresponding solvent free energy maps. The maps were generated with energy and density cutoff values of $e_{CO}^{(L)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(L)} = 6.93 \rho^0$; **C**, **F**, **I**: MRBBs and solvent free energy maps for **B7**, **B6** and **B5**. The maps were generated with energy and density cutoff values for the MRBB molecules of $e_{CO}^{(MRBB)} = -0.95 \text{ kcal} \cdot \text{mol}^{-1}$ and $g_{CO}^{(MRBB)} = 3.0 \rho^0$. The displayed conformations are the cluster representatives of the most populated cluster for the conformational ensembles of **B7**, **B6** and **B5** (40.0%, 54.0% and 51.0% occupancy, respectively).

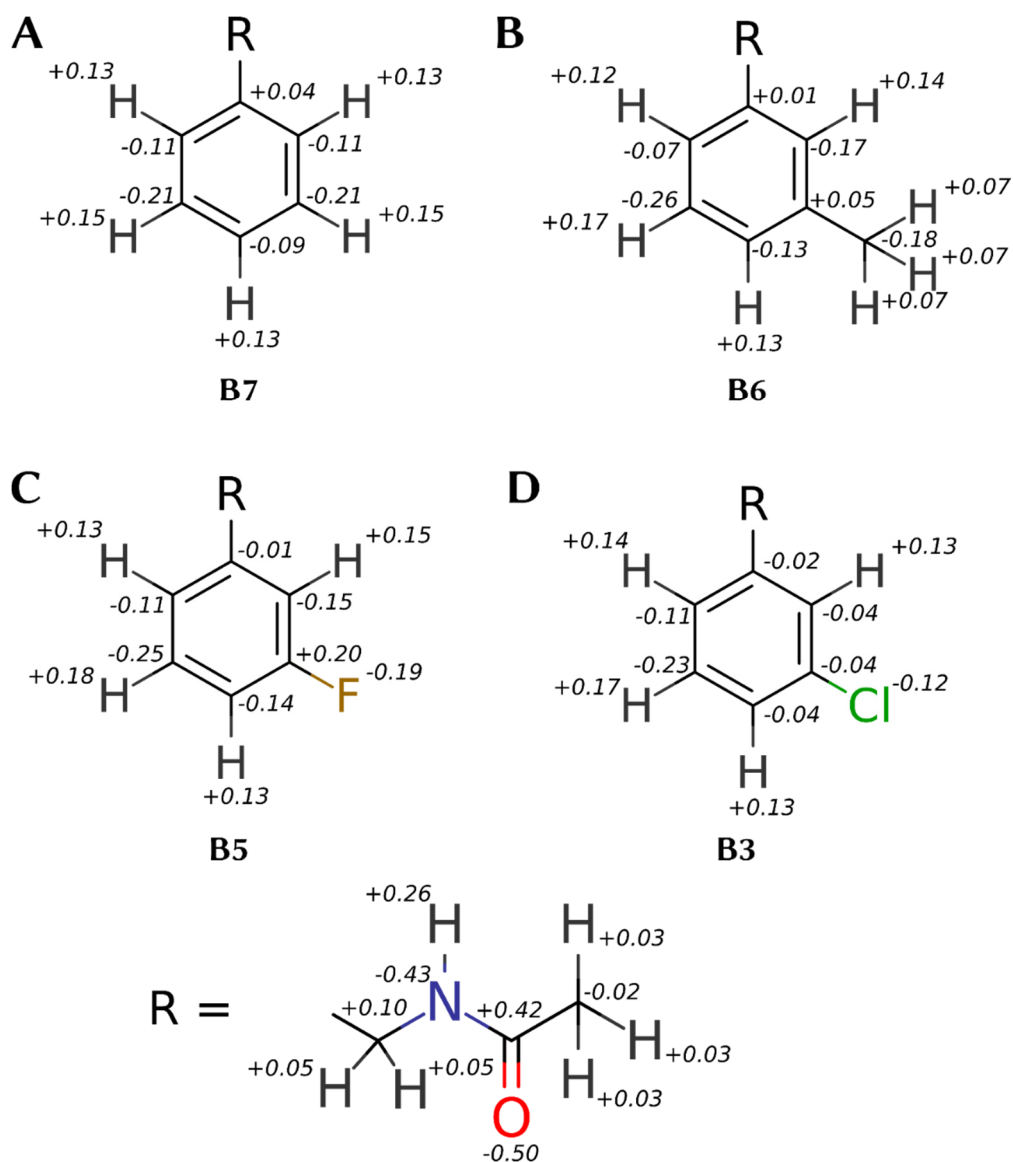


Figure 3-10: Partial charge distribution for MRBB **B7**, **B6**, **B5** and **B3** as obtained by the RESP charge calculation. The same charges as in the BBs were also used in the ligands. Small deviations from the expected total sum of zero for the charges in this depiction are due to round-offs. The actual charges that were used in the force field have a precision of 10^{-8} and have a total sum of charges equal to zero.

3.4 Discussion

We presented a novel strategy to partition and spatially map contributions of molecular solvation thermodynamics onto protein-ligand binding following chemically intuitive decomposition rules that split given ligands of a dataset in reoccurring building blocks (BBs). As a set of peptidomimetic ligands was investigated, the splitting into smaller sub-structural BBs occurred at primary and secondary amide groups or primary sulfonamide groups. A virtual library of 44 BBs was annotated with solvation thermodynamic properties using capped analogs (MRBBs) of the BBs and the whole ligand molecules in the unbound and protein-bound state. The thermodynamic properties were calculated by analyzing molecular dynamics trajectories with a GIST-based solvation functional that was specifically optimized for this dataset. Our solvation functional suggests, as already mentioned in our previous contribution, that the desolvation of unfavorably bound water molecules on the surface of the unbound ligand accounts for the main contribution to the free energy of binding. In contrast, the differences in the solvation of the protein-ligand complexes are determined particularly in regions that contain in some complexes energetically very unfavorable water molecules.

The BBs have greatly varying free energy values, depending on the ligand scaffold in which they are embedded. Thus, strong cooperative effects between the individual BBs forming the entire ligand are observed. In the unbound state of the ligands, the solvation free energy of a BB can be greatly enhanced by another BB, even when it is located at a distal site. In the intriguing case of the congeneric pair **1**→**2**, we have found enhanced solvent structuring around the P₁ site in the unbound state of **2** which bears, compared to **1**, an additional phenyl ring at the remote P₃ site. In a previous contribution,¹⁰⁵ the difference in binding free energy could not be unambiguously explained by the crystal structures alone, since the gain in hydrophobic contact area of only 10 Å² of **2** over **1** was too small to explain the trend in binding free energy. Furthermore, **2** binds with a stronger enthalpic signal than **1**, which somewhat contradicts (according to the classical hydrophobic effect) the observation that the additional phenyl group of **2** displaces more water molecules from the binding pocket than **1**. Our explanation based on the remote solvent stabilization for the observed difference in binding affinity would most likely not be considered in a drug optimization process, since cooperative solvation effects present in the unbound state prior to protein binding are usually not investigated. Instead, it appears rather tempting to attribute the difference in binding affinity directly to the interactions of the phenyl group with the protein or the individual physicochemical properties attributed to the phenyl

group. Most likely however, the marked cooperative influence of a phenyl group attached to a remote portion at the ligand scaffold would not be assumed as affinity enhancing factor. In a putative next step of optimization, the design strategy would try to keep the di-phenyl group at the P₃/P_A site and optimize the P₁ occupant while retaining the unfavorable solute-water interactions in the unbound state of the ligand induced by the remote solvent stabilization between the P₃/P_A and P₁ portions.

In another congeneric series, we demonstrated how varying decorations (-H (**3**), -CH₃ (**4**), -Cl (**1**), -F (**5**)) of the P₁ phenyl portion affect the solvation free energy of the protein-ligand complex and the unbound ligand in solution. Ligand **3** with an unsubstituted P₁ phenyl ring entraps a water molecule at an energetically unfavorable site between its P₁ portion and Tyr228 of the protein. Due to this energetically unfavorable situation, **3** is less potent than its *meta*-methyl and *meta*-chloro analogs **4** and **1**, respectively. This interpretation differs from a previous study, in which the differences in binding affinity were attributed to distinct contributions attributed to the displacement of solvent molecules from the binding site.⁷⁹ In this previous study, the difference between the fluorinated **5** and its chlorinated derivative **1** was mainly based on the different volume of the fluorine and chlorine atoms and their resulting difference in solvent displacement volume. Whereas in our work, the differing water interactions in the vicinity of a fluorine-substituted phenyl moiety and a chlorine-substituted phenyl moiety are considered. Our solvation functional suggests a decrease in solvation free energy for unbound **5** due to a more tightly binding of water molecules which makes accordingly the desolvation of **5** less favorable than of **1**.

Due to the fact that in our approach, the contributions of the water molecules in the unbound and bound state are effectively considered, renders our model a physically more realistic picture of the formed protein-ligand complex. A caveat is however, that the bioactive conformation, or a reasonable estimate of it, must be known *a priori* to the calculation. However, this is a common problem in any free energy calculation method and, in particular, in molecular field-based 3D-QSAR approaches.

Particular with respect to the latter 3D-QSAR approaches (e.g. CoMFA¹⁴⁴ and CoMSIA¹⁴⁵), methods still very popular in medicinal chemistry and drug design, our study might suggest some intriguing insights. In 3D-QSAR, a set of ligands is mutually aligned in their (assumed) bioactive conformations and embedded into an equally spaced 3D grid. Subsequently, by means of a molecular probe placed at the intersections of the grid, the exposed properties and spatial differences of the ligands are scanned using some kind of molecular interaction potential (in

the simplest case Lennard-Jones and Coulomb potentials, but more sophisticated potentials have been applied). Overall, the generated input data for the relative comparison of the ligands reminds about our spatial maps generated by exploring the solvation properties around our molecules using MD trajectory data generated with water molecules and analyzed with the GIST method.

Major criticism of the 3D-QSAR approaches related to the lack of consideration of the protein environment that definitely provides a much stronger differentiated interaction pattern than an encompassing grid scanned with a uniform molecular probe. Furthermore, the 3D-QSAR methods seem to fully ignore the entropic contributions of the free energy of binding. Therefore, it always appeared as a miracle that 3D-QSAR methods performed so well in relating structural ligand data with binding affinities.

Our GIST analysis using a novel functional for evaluation, admittedly collected at one data set, suggests that the desolvation of unfavorably bound water molecules on the surface of the unbound ligands accounts for the main contribution to the free energy of binding. This, as in 3D-QSAR, requires an alignment of the ligands and a mapping of the ligand properties across their surfaces by a force-field implemented in the applied MD simulation. The subsequent analysis by our GIST functional reminds about the data evaluations used in 3D-QSAR. Possibly, a significant portion of the binding properties are already encoded in the desolvation properties of the ligands. Contributions arising from features in the protein relate to more special situations involving water molecules that significantly deviate from their properties in the bulk phase and become entrapped at energetically unfavorable sites. Obviously, a large part of the intuitively assumed modulations of the distinct interactions formed within the highly structured environment of a binding pocket and which can be exploited to bind a ligand are compensated by the individual desolvation costs required for the displacement of water molecules from the binding site. Therefore, scanning the ligands only with a simple probe provides already a relevant picture to reasonably predict affinity data. Perhaps these considerations explain to some degree why 3D-QSAR performs so surprisingly well.

3.5 Conclusion

In this work, we demonstrated how the solvation thermodynamic properties obtained from GIST-based solvent functionals can be readily decomposed into individual contributions from chemically meaningful BBs. With our approach, drug candidates can be optimized using

solvation as an active and intuitively accessible design parameter. The decomposition into BBs is effectively a mean to navigate through chemical space using mapped solvation properties obtained from a physically meaningful model. In the next step, our approach must be evaluated experimentally by linking it to a generative method in order to foster its full potential. Further testing against different target proteins is also needed, however the training with reasonable structural and thermodynamic data is of great importance.

Our approach is implemented in the latest version of *Gips*. It is available from the GitHub page of the first author (github.com/wutobias) accompanied with a tutorial on how to derive solvation properties based on BBs. The BB decomposition can be carried out automatically using the *Recap*¹⁵¹ algorithm as implemented in *RDKit*¹³⁹, or using a custom BB definition (as used in this work).

3.6 Methods

In this section, we describe the procedure for decomposing a set of 53 thrombin ligands into 44 unique BBs. In the following, these are used to calculate the spatial decomposition of GIST-based solvent functionals. As a point of reference, we also carried out MD simulations and GIST calculations of the BBs. The calculation of the GIST grids that are the input for our solvation functionals, was already introduced in our last contribution using the same dataset of thrombin ligands. For this reason, the structure preparation procedure as well as the molecular dynamics protocol applied to all protein-ligand complexes and the ligands separately will not be described here. The MRBB molecules (see Figure 3-11) required a different treatment than the entire ligand and protein structures regarding the structure preparation and simulation protocol. For this reason, the structure preparation and simulation protocol of the MRBBs is described in the following section.

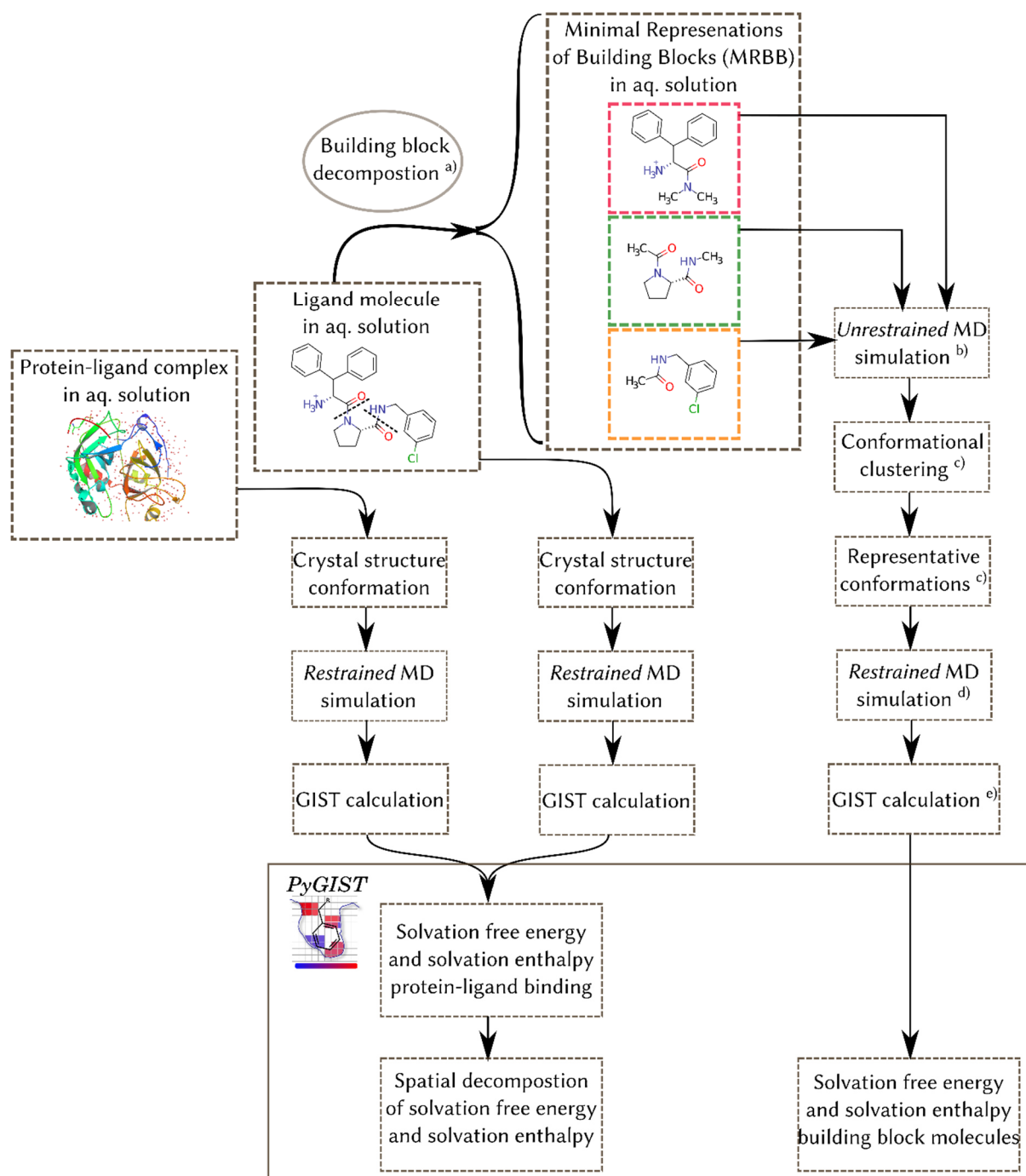


Figure 3-11: Overview of the workflow employed in this study. The terms unrestrained and restraint MD simulations refer to simulations without and with positional restraints on the non-hydrogen atoms, respectively. The lower-case letters a)-e) refer to different steps in the workflow as referenced in the following.

3.6.1 The Dataset

The dataset that is investigated in this study consists of 53 thrombin ligands characterized by crystal structures and thermodynamic profiles using ITC and SPR measurements^{49,103–111}. These ligands were mutually paired such that any difference in binding thermodynamics can, most likely, be attributed predominantly to changes in the solvation/desolvation properties. This dataset was already introduced in our previous contribution.

3.6.2 Decomposition of the Ligands into BBs and MRBBs

For the generation of a virtual BB library (see a) in Figure 3-11), we searched for all primary and secondary amide groups, as well as all primary sulfonamide groups in the set of ligand molecules. For the amide groups, the bond between the carbonyl carbon atom and the adjacent nitrogen atom was cleaved. The formally created *C*- and *N*-terminal ends of the cleaved bond were capped using NME (*N*-methyl) and ACE (acetyl) capping groups, respectively. If the cleaved bond was part of a secondary amide group, the *N*-terminal ends were capped using NDME (*N*-dimethyl). For the sulfonamide groups, the bond between the sulfur and nitrogen atom was cleaved. Here, an NME group was attached to the *S*-terminal end and a methylsulfonate ($-\text{SO}_2\text{CH}_3$) group was attached to the *N*-terminal end. After the BB decomposition and capping procedure, a BB has effectively become an MRBB molecule without any open valences. Finally, all redundant entries in the resulting set of MRBBs and BBs are eliminated, resulting in a library of 44 unique BBs.

3.6.3 Structure Preparation

For each entry in the BB library, a conformational ensemble of at most three conformers per BB was generated using *Omega*^{152,153} from the *OpenEye* suite of programs. For each conformer, a geometry optimization at the b3lyp/6-31G* level was carried out, followed by the calculation of the ESP at the HF/6-31G* level using the Gaussian09 program.¹¹⁵ Partial atomic charges were calculated from the ESP by a multimolecule and multiconformational *RESP* fitting^{55,114} using the *resp* program from the AmberTools17 program package.¹¹⁷ The restraints on the partial charges were applied in accordance with the original work published on the derivation of partial charges for the Amber force field. The complete procedure was carried out using an in-house workflow.

For each entry in the BB library, *GAFF* atom types and force field parameters¹¹⁶ were assigned using *parmchk2* and *tLeAP*. The simulation boxes with the shape of a truncated octahedron

were filled with TIP4P-Ew water molecules¹¹⁸, such that the distance between any solute atom and the box edges is no longer than 16 Å. Then, sodium or chlorine counter ions were added at random positions in order to ensure net neutrality using the *addIonsRand* utility of *tLeAP*. From the resulting parameter and structure files, water molecules were removed by random (approximately 1% of the initially placed water molecules), such that each system contained exactly 2000 water molecules in total. The energetically most favorable geometry from each conformer ensemble that was generated for the partial charge calculation, was used as the starting structure for the subsequent molecular dynamics runs.

3.6.4 Unrestraint MD Simulations

We initially performed MD simulations without any positional restraints in order to get an ensemble of conformations for the MRBB molecules (see b) in Figure 3-11). All minimization manipulations were carried out using the *pmemd* program from Amber16 and all molecular dynamics runs were carried out using the GPU accelerated *pmemd.cuda*¹²⁰⁻¹²². During all following operations, periodic boundary conditions were applied using a 9.0 Å cutoff for the direct space sum. The SHAKE algorithm¹¹⁹ was used on all bonds involving hydrogen atoms during the molecular dynamics runs. All simulations parameters were kept at their default values except stated otherwise. Each simulation was carried out in triplicates.

Initially, each MRBB had positional restraints on all non-hydrogen atoms of the starting structure using a harmonic force constant of 25 kcal·mol⁻¹·Å⁻². In the first step, the potential energy of the system was minimized with 250 steps of steepest descent and 250 steps of conjugate gradient optimization. Then, the system was heated gradually to 300 K within 25 ps using an integration time-step of 1 fs. At this temperature and with an integration time-step of 2 fs, the system was equilibrated to a target pressure of 1 bar using the Berendsen barostat¹⁵⁴, while gradually lowering the positional restraints within 100 ps. In a last step, the system was equilibrated under NVT conditions for 1 ns. Final production MD runs were carried out for 50 ns for each MRBB. Coordinates were saved to disk every 10 ps.

3.6.5 Conformational Clustering

The conformations of the MRBB molecules (see c) in Figure 3-11) were clustered based on the local symmetry-corrected RMSDs of all non-hydrogen atoms using average linkage, single linkage and complete linkage clustering as implemented in *cpptraj*¹²³ (V17). For the clustering,

every second frame from the triplicate MD runs was sieved off using the *sievetoframe* utility. In order to keep the computational effort in a reasonable range, we only considered the clustering solutions for two and three clusters for each clustering algorithm. This strategy resulted in a maximum number of 396 MD simulations and GIST calculations to run (3 replica * 3 clusters * 44 MRBB molecules). From each of the clustering algorithms, the conformational ensemble was clustered into two and three clusters. Then, from the three clustering algorithms and two different clustering ($N = 2,3$) solutions for each clustering algorithm, the clustering solution that had the lowest Davies–Bouldin index was chosen. In the Davies-Bouldin index, the ratio between within-cluster scatter and between-cluster separation is considered. This index is a common measure to identify a cluster solution that has compact clusters well separated from each other. The number of conformational clusters for each MRBB as well as their population statistics can be found in the Supporting Information.

3.6.6 Restraint MD Simulations

For each cluster from the optimal clustering solution, the most representative conformation (i.e. the frame from the MD trajectory that is closest to the cluster centroid) was selected as the starting structure for restraint MD simulations (see d) in Figure 3-11). The structure preparation for these MD simulations was analogous to the one for the unrestraint MD simulations (see step b)). During the following energy minimization and MD runs, all non-hydrogen solute atoms were fixed to the coordinates from their starting structure using a harmonic potential. Langevin dynamics ($\gamma = 2$ ps) were applied to keep the system at constant temperature. All parameters were kept similar to the protocol that was used during the unrestraint simulations (see step b)), except when stated otherwise. Each simulation was carried out in triplicates.

In the first step, the energy of the system was minimized using 2500 steps of steepest descent and 2500 steps of conjugate gradient optimization. All non-hydrogen solute atoms were positionally restraint to their initial coordinates (i.e. to the cluster representative structure) using a harmonic potential with a force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$. In a second minimization, 2500 steps of steepest descent and 2500 steps of conjugate gradient optimization were carried out, while using a force constant of $2 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$ for the positional restraints. Then, the system was heated to 300 K within 25 ps using an integration time step of 1 fs and a harmonic force constant of $25 \text{ kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^{-2}$ to keep the positions of the solute atoms fixed. At this temperature the system was equilibrated to a target pressure of 1 bar within 5 ns using the Berendsen

barostat¹⁵⁴. The integration time step is now switched to 2 fs. Finally, the system is equilibrated for 5 ns under NVT conditions. Final production runs were carried out for 30 ns and coordinates were saved to disk every 2 ps.

3.6.7 GIST Calculations

The solvent energies and entropies were calculated and mapped on a three-dimensional rectangular grid using the *GIST*^{79,80} (see e) in Figure 3-11) implementation of *cpptraj* (V17). For each MRBB, the grid box was centered at the center-of-mass of the MRBB molecule. The dimensions of the grid box were chosen such, that the distance of every edge to its closest atom of the MRBB was 3 Å. Each grid voxel had dimensions 0.5x0.5x0.5 Å.

3.6.8 GIST-based Solvent Functionals

The GIST-based solvent functionals were used as introduced in our previous contribution. For the MRBB molecules, we used the ligand-bound density, entropy and energy cutoff parameters ($g_{CO}^{(L)}, s_{CO}^{(L)}$ and $s_{CO}^{(L)}$, respectively) from the PL-L/F6($g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}$) functional. However, this required the correction of the density cutoff parameter, $g_{CO}^{(L)}$, in order to reflect the difference in the molecular volume of the entire ligand molecules and the MRBB molecules. This parameter correction procedure is outlined in the Supporting Information.

3.6.9 Spatial Decomposition of Solvent Functionals

GIST is a spatially resolved approach to solvation thermodynamics. Within this approach, the spatial distribution of solvent molecules and their corresponding thermodynamic properties are obtained from spatial integrals over a grid that is superimposed onto the solute molecule of interest (or parts of it). The spatial integrals can be readily decomposed into sub-integrals, which reflect the topology of the solute molecule. These sub-integrals allow one to rewrite any integral over enthalpies or entropies from *GIST* as follows (expressed as sums, instead of integrals):

$$A_L = \sum_k^{G_L} v(\vec{r}_k) g(\vec{r}_k) A(\vec{r}_k) = \sum_i^B \sum_k^{G_L} b_i(\vec{r}_k) v(\vec{r}_k) g(\vec{r}_k) A(\vec{r}_k) \quad (3-1)$$

$$\mathbf{A}_L^{(B_i)} = \sum_k^{G_L} b_i(\vec{r}_k) v(\vec{r}_k) g(\vec{r}_k) A(\vec{r}_k) \quad (3-2)$$

$$\mathbf{A}_L = \sum_i^B \mathbf{A}_L^{(B_i)} \quad (3-3)$$

In Eq. (3-1), \mathbf{A}_L is either the enthalpy or the entropy calculated by using a grid G_L obtained from a GIST calculation of the ligand molecule L (either in solution or in the protein-bound state). The value of \mathbf{A}_L at grid voxel \vec{r}_k is denoted as $A(\vec{r}_k)$, the volume indicator function is $v(\vec{r}_k)$ and evaluates to 1, if the grid voxel is within the molecular volume of the ligand molecule and to 0 otherwise. The normalized density is given by $g(\vec{r}_k)$ and can be interpreted as a weighting function for $A(\vec{r}_k)$. The most right side of the equation contains the binary BB indicator function b_i , which is assigned a value of 1, if grid voxel k is inside the molecular volume of BB i , and a value of 0 otherwise (for a graphical depiction of the spatial decomposition approach, see Figure 3-12). The index i runs over all B BBs that are contained in the ligand molecule. Thus, for each BB, B_i , its fractional contribution of the total value of \mathbf{A}_L , can be expressed as $\mathbf{A}_L^{(B_i)}$ from eq. (3-2). Consequently, the value of \mathbf{A}_L , can be expressed as a sum over the contributions from all BBs that are contained in molecule L , $\mathbf{A}_L^{(B_i)}$, using eq. (3-3).

In addition to the MD simulations and GIST calculations that were carried out for the ligand molecule, we carried out MD simulations and GIST calculations for MRBB molecules in aqueous solution (see Figure 3-11). From these, a similar spatial decomposition of the GIST grids has been carried out as in the case of the entire ligand molecules (see Figure 3-12). The spatial decomposition is carried out for the same atoms as in the case of the BBs in the entire ligand molecule. A similar approach as in the case of the entire ligand molecule has been used to calculate the thermodynamic solvation quantities (i.e. solvation energy and solvation entropy), \mathbf{A}_{B_i} , from a MRBB molecule B_i as:

$$\mathbf{A}_{B_i} = \sum_k^{G_{B_i}} b_i(\vec{r}_k) v(\vec{r}_k) g(\vec{r}_k) A(\vec{r}_k) \quad (3-4)$$

In eq. (3-4), the quantities $b_i(\vec{r}_k)v(\vec{r}_k)g(\vec{r}_k)$ and $A(\vec{r}_k)$ have the same meaning as in eq. (3-1), however here they are based on the grids, G_{B_i} , obtained from a GIST calculations of the MRBB

molecule (see Figure 3-12). Thus, the results obtained from the analysis of the BB in the entire ligand molecule and the MRBB are readily compared to each other in order to quantify the perturbation of an individual BB upon assembling into the entire ligand molecule.

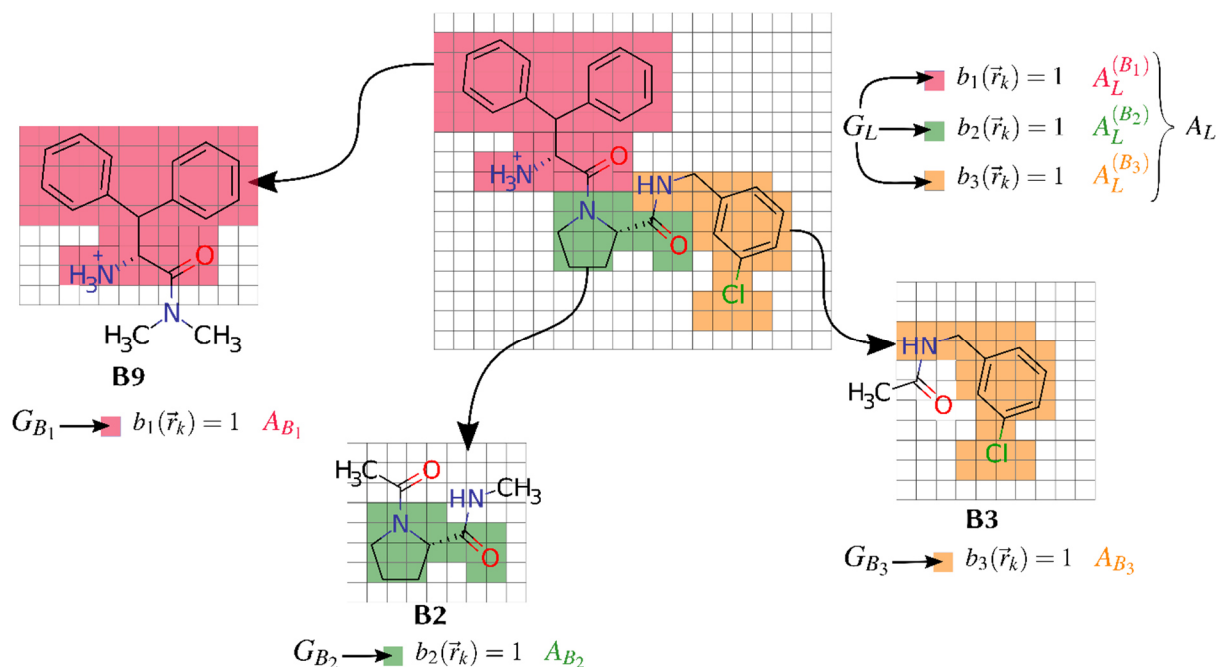


Figure 3-12: Schematic two-dimensional illustration of the spatial decomposition of the GIST grids. The colored grid voxel show the BB indicator functions b_1 , b_2 and b_3 as outlined in eq. (3-1). Any BB indicator function is zero at white grid voxels (i.e. without any assigned color). The molecule in this example (cf. Figure 3-11) is a thrombin inhibitor taken from PDB code 3DHK¹⁰⁵ and its corresponding MRBBs **B9**, **B2** and **B3**.

3.7 Supporting Material

3.7.1 PDB Accession Codes

Ligand bound structures:

Reference [¹⁰⁴] 2ZC9, 2ZDA, 2ZDV, 2ZF0, 2ZFF.

Reference [¹⁰⁵] 2ZFP, 3DHK, 2ZGX, 2ZO3, 3DUX.

Reference [¹¹⁰] 3BIU, 3BIV.

Reference [¹⁰³] 3P17, 3QTO, 3SI3, 3SI4, 3SV2, 3QTV, 3SHC, 3QWC, 3QX5.

Reference [⁴⁹] 3RLW, 3RLY, 3RM0, 3RM2, 3RML, 3RMM, 3RMN, 3RMO, 3T5F, 3UWJ.

Reference [¹¹¹] 3UTU.

Reference [¹⁰⁸] 4BAK, 4BAM, 4BAN, 4BAO, 4BAQ.

Reference [¹⁰⁹] 4UD9, 4UDW, 4UE7, 5AF9, 5AFZ.

Reference [^{107,148}] 6GBW, 5JFD, 5LCE, 5JZY, 5LPD

Reference [¹⁰⁶] CC01, CC04, CC05, CC08, CC10, CC11.

3.7.2 List of ligand pairs

Table S3-3: List with PDB codes of all ligand pairs that differed only by a single BB.

PDB-code ligand 1	PDB-code ligand 2	PDB-code ligand 1	PDB-code ligand 2	PDB-code ligand 1	PDB-code ligand 2
3QWC	3QTV	3DUX	2ZC9	3RM0	3RLY
3SI4	3QTV	CC05	2ZC9	3UWJ	3RLY
3QTV	3QTO	CC08	2ZC9	3RM2	3RLY
3QX5	3QTV	CC10	2ZC9	3UWJ	3RM0
3SI4	3QWC	CC11	2ZC9	3RM2	3RM0
3QWC	3QTO	3P17	2ZFF	3UWJ	3RM2
3QX5	3QWC	3SV2	2ZFF	3DHK	2ZC9
3SI3	2ZF0	4UDW	2ZFF	3RM2	3RLW
2ZF0	2ZDV	CC05	2ZFF		
3SHC	2ZF0	CC08	2ZFF		
2ZF0	2ZC9	CC10	2ZFF		
2ZFF	2ZF0	CC11	2ZFF		
3P17	2ZF0	3SV2	3P17		
3SV2	2ZF0	4UDW	3P17		
4UDW	2ZF0	CC05	3P17		
CC05	2ZF0	CC08	3P17		
CC08	2ZF0	CC10	3P17		
CC10	2ZF0	CC11	3P17		
CC11	2ZF0	3QX5	3QTO		
3SI3	2ZDV	4UDW	3SV2		
3SI3	3SHC	CC05	3SV2		
3SI3	2ZC9	CC08	3SV2		
3SI3	2ZFF	CC10	3SV2		
3SI3	3P17	CC11	3SV2		
3SV2	3SI3	CC05	4UDW		
4UDW	3SI3	CC08	4UDW		
CC05	3SI3	CC10	4UDW		
CC08	3SI3	CC11	4UDW		
CC10	3SI3	3DHK	2ZFP		
CC11	3SI3	3DUX	2ZFP		
3SHC	2ZDV	2ZO3	2ZGX		
2ZDV	2ZC9	5JZY	2ZGX		
2ZFF	2ZDV	5LCE	3DUX		
3P17	2ZDV	4BAO	4BAN		

3SV2	2ZDV	4BAO	4BAM	
4UDW	2ZDV	4BAQ	4BAO	
CC05	2ZDV	4BAN	4BAM	
CC08	2ZDV	4BAQ	4BAN	
CC10	2ZDV	4BAQ	4BAM	
CC11	2ZDV	3BIV	3BIU	
3SHC	2ZC9	CC08	CC05	
3SHC	2ZFF	CC10	CC05	
3SHC	3P17	CC11	CC05	
3SV2	3SHC	CC10	CC08	
4UDW	3SHC	CC11	CC08	
CC05	3SHC	CC11	CC10	
CC08	3SHC	3RMM	3RML	
CC10	3SHC	3RMN	3RML	
CC11	3SHC	3T5F	3RML	
2ZGX	2ZDA	3RMO	3RML	
2ZO3	2ZDA	3RMN	3RMM	
5JZY	2ZDA	3T5F	3RMM	
3SI4	3QTO	3RMO	3RMM	
3SI4	3QX5	3T5F	3RMN	
2ZFF	2ZC9	3RMO	3RMN	
3P17	2ZC9	3T5F	3RMO	
3SV2	2ZC9	3RLY	3RLW	
4UDW	2ZC9	3RM0	3RLW	
2ZFP	2ZC9	3UWJ	3RLW	

3.7.3 The GIST-based Solvent Functional and its Parameters

The GIST-based solvent functional and the corresponding parameter were already introduced in our previous work. We will make use of the PL-L/F6/ $(g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)})$ functional. The characteristic feature of this solvent functional is the use separate energy, entropy and density parameters for the protein-ligand complex (PL) as well as the ligand (L) in aqueous solution. The grids from the GIST calculation are processed with the F6 base functional, which employs a set of energy, entropy and density cutoff values (e_{CO} , s_{CO} and g_{CO}) in order to filter the grid voxels for appropriate values of energy, entropy and density. These grid voxels are then employed for the calculation of solvation energies and entropies by assigning an energy and an entropy weighting factor (E_{aff} and S_{aff}) to each grid voxel that passes the filter criteria. The sum of all these weighted grid voxel that are within the first solvation layer of a ligand give the solvation energy and entropy of a ligand (or a BB of it). The values found for the individual parameters that were used in this study are listed in Table S3-4. These parameter values were obtained from 10 randomized attempts of five-fold cross-validation and was carefully evaluated against shuffled data generated with the same dataset.

The positive value for E_{aff} results in a solvation energy contribution from the protein-ligand complex, $E_{Solv}^{(PL)}$, which opposes binding. However the solvation energy contribution from the ligand molecule, $E_{Solv}^{(L)}$, in aqueous solution actually favors binding, since the solvation energy of the binding reaction is calculated as $\Delta E_{Solv} = E_{Solv}^{(PL)} - E_{Solv}^{(L)}$. The high value for the energy

cutoff parameter for the protein-ligand complex and the corresponding density cutoff parameter, $8.13 \text{ kcal}\cdot\text{mol}^{-1}$ and $8.31 \rho^0$, respectively, allow only for grid voxel that are very unfavorable in solvation energy and are highly populated (approximately 8 times higher than bulk water phase). As a result, only in few regions in the pocket grid voxel are found that actually exceed these cutoff parameter values. These correspond to water molecules that are found on the surface of the protein-ligand complex and are not placed favorably with respect to energy. In the context of structure-based ligand design, one would want to replace this water molecule with an apolar moiety or modify the ligand such that it interacts energetically favorable with this water molecule. The energy cutoff parameter for the ligand molecule is close to zero ($-0.95 \text{ kcal}\cdot\text{mol}^{-1}$) and its density cutoff parameter ($6.93 \rho^0$) is close to the density parameter of the protein-ligand complex. This combination of energy and density cutoff parameters for the ligand effectively identifies water molecules in high density regions on the surface of the unbound ligand molecule with a total energy that is close to (or higher) than their in energy bulk water. By that, the solvation energy of the ligand is dominated by high density regions with unfavorable energy.

The entropy weighting factor, S_{aff} , is positive and close to the value found for the energy E_{aff} . The positive sign of this factor is anticipated, since it indicates that the binding of water molecules in the protein-ligand complex is entropically unfavorable and the desolvation of the ligand molecule is entropically favorable since water molecules are released into bulk (note, that the GIST functionals use the negative the entropy term, $-T\Delta S$). The entropy cutoff parameter for the protein-ligand complex is very high, $7.83 \text{ kcal}\cdot\text{mol}^{-1}$, and therefore only is fulfilled in regions that have very tightly bound water molecules, like structural water molecules. The cutoff parameter for the ligand molecules is lower, $3.95 \text{ kcal}\cdot\text{mol}^{-1}$ and identifies bound water molecules on the surface of ligand.

Table S3-4. Parameters for the PL-L/F6/ $(g_{CO}^{(PL)}, g_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)}, e_{CO}^{(PL)}, s_{CO}^{(L)})$ solvent functional.

$E_{aff}^{a) e)}$	$e_{CO}^{(PL) b) e)}$	$e_{CO}^{(L) b) e)}$	$S_{aff}^{a) e)}$	$s_{CO}^{(PL) c) e)}$	$s_{CO}^{(L) c) e)}$	$g_{CO}^{(PL) d) e)}$	$g_{CO}^{(L) d) e)}$
0.17/ 0.16/ 0.29	8.03/ 7.43/ 8.56	-0.95/ -0.96/ -0.65	0.01/ -0.14/ 0.61	7.83/ 6.83/ 7.89	3.95/ -0.96/ 5.15	9.97/ 5.63/ 9.99	6.93/ 6.88/ 6.95

- a) Weighting factors for energy, E_{aff} , and entropy, S_{aff} in kcal·mol⁻¹.
- b) Energy cutoff parameters for the protein-ligand complex (PL), $e_{CO}^{(PL)}$, and the ligand molecule in aqueous solution, $e_{CO}^{(L)}$, in kcal·mol⁻¹.
- c) Entropy cutoff parameters for the protein-ligand complex (PL), $s_{CO}^{(PL)}$, and the ligand molecule in aqueous solution, $s_{CO}^{(L)}$, in kcal·mol⁻¹.
- d) Density cutoff parameters for the protein-ligand complex (PL), $g_{CO}^{(PL)}$, and the ligand molecule in aqueous solution, $g_{CO}^{(L)}$. This quantity is given in multiples of ρ^0 .
- e) The first value indicates the median of the parameters obtained from all training/testing attempts with this functional. The second and third values represent the upper to lower quartile range of the parameters obtained all training/testing attempts.

3.7.4 Correction of the Density Cutoff Parameter for the MRBB

In order to be able to compare solvation thermodynamic properties from the MRBBs with the BBs in the ligands, it is important compare the same amount of water molecules in both environments of the respective BB. The amount of water molecules, which are considered in the calculation of solvation energy and entropy is controlled by the density cutoff parameter, g_{CO} . As outlined in our previous work on GIST functionals, this parameter controls whether or not a grid voxel k must be considered in the calculation or not. If the normalized water density at grid voxel k , $g(\vec{r}_k)$, exceeds the density cutoff value, g_{CO} , then this grid voxel is considered in the calculation. If the normalized water density at this grid voxel is lower than the cutoff value, then the grid voxel is not considered in the calculation.

The BB in the ligand is able to accommodate a higher number of water molecules, than in an isolated environment. This is due to cooperative effects between the BB and the other BBs in the molecule, which enhance the solute-water interactions and thereby increase the probability to find a water molecule in the vicinity of the solute surface. Therefore, we searched for the density cutoff value, which results in the approximate same number of water molecules for the BB in MRBB and the BB embedded in the ligand. This corresponds to finding the density cutoff parameter value of the MRBB, $g_{CO}^{(MRBB)}$, which minimizes the difference in the number of water molecules between the BB embedded in the ligand and the BB mapped to the MRBB molecule.

$$\Delta N = \rho^0 \sum_i^B \left(\sum_k^{G_L} V_k b_i(\vec{r}_k) g(\vec{r}_k) g_s(\vec{r}_k, g_{CO}^{(L)}) \right) - \left(\sum_l^{G_B} V_l g(\vec{r}_l) g_s(\vec{r}_l, g_{CO}^{(MRBB)}) \right) \quad (\text{S3-5})$$

$$g_s(\vec{r}, g_{CO}) = \begin{cases} 1, & \text{if } g(\vec{r}) > g_{CO} \\ 0, & \text{otherwise} \end{cases} \quad (\text{S3-6})$$

In eq. (S3-5), the difference in the number of water molecules is denoted as ΔN , the grid that covers the ligand molecule is called G_L and the grid that covers the MRBB molecule is called G_B . The volume of a grid voxel from G_L is denoted V_k and from the grid G_B , it is called V_l . The density function is called g and the corresponding density step function is g_s . We scanned the average difference in the number of water molecules calculated for all ligand molecules in the dataset, $\langle \Delta N \rangle$, against different density cutoff parameter values for the MRBB molecule, $g_{CO}^{(MRBB)}$. Furthermore, we carried out this scan for different values of the additive parameter, R_a , that is added to the radius of each atom during the molecular volume calculation using a water probe. The greater this parameter is, the greater also will be the molecular volume and by that, it effectively controls the size of the molecular volume. As can be seen from see Figure S3-13, $\langle \Delta N \rangle$, drops to a minimum at a cutoff parameter value of $3 \rho^0$. The same behavior is observed for different values of the additive parameter, R_a , which indicates a consistent number of water molecules in the different radial increments within the first solvation layer at this level of the density.

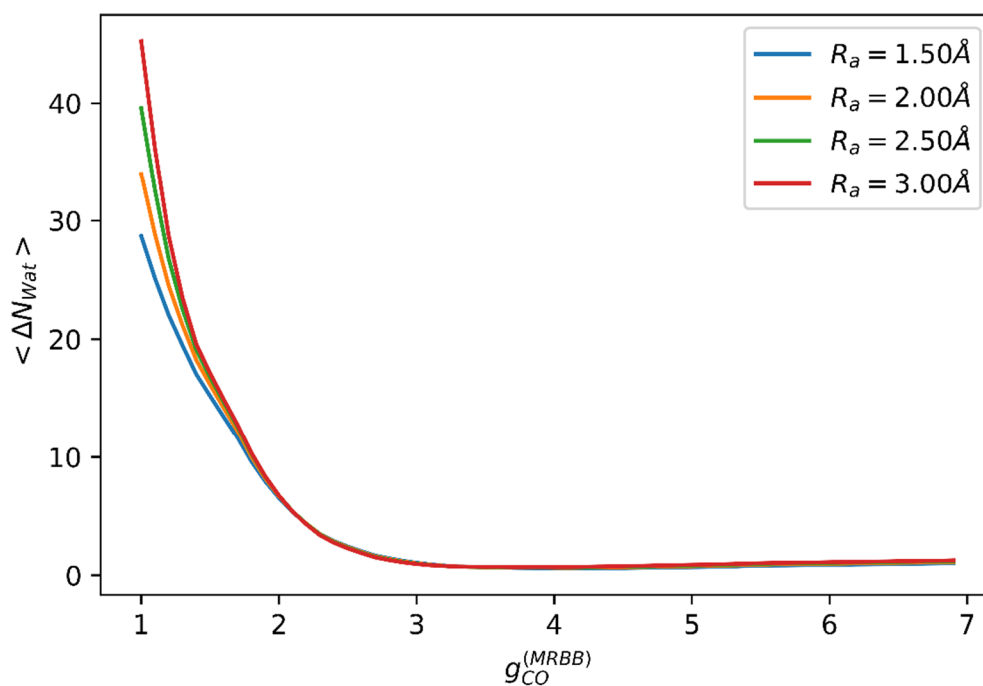


Figure S3-13: Average difference in the number of water molecules, $\langle \Delta N_{Wat} \rangle$, between the MRBB and the BB embedded into the molecule calculated with different density cutoff values g_{CO} . Each line represents a different volume definition per atom.

3.7.5 Building Block Thermodynamics for each Ligand.

Table S3-5: Thermodynamic contributions for each BB in each molecule.

PDB	BB	$\Delta G^{(PL)}$ +/- s.d.				$\Delta H^{(PL)}$ +/- s.d.				$T\Delta S^{(PL)}$ +/- s.d.			
3QTV	1	0.1	0.1	2.5	0.7	0.1	0.1	2.8	0.9	0.0	0.0	-0.3	0.7
3QTV	2	0.1	0.1	2.5	0.7	0.1	0.1	2.8	0.9	0.0	0.0	-0.3	0.7
3QTV	16	1.3	0.7	1.8	0.5	1.3	0.7	2.1	0.6	0.0	0.2	-0.2	0.5
3QWC	1	0.1	0.1	2.4	0.7	0.1	0.1	2.6	0.8	0.1	0.1	-0.2	0.7
3QWC	2	0.0	0.0	2.4	0.7	0.0	0.0	2.6	0.8	0.0	0.0	-0.2	0.7
3QWC	17	0.8	0.5	2.1	0.6	0.6	0.4	2.2	0.7	0.1	0.2	-0.1	0.6
3SI4	1	0.1	0.1	2.1	0.6	0.1	0.1	2.4	0.8	0.0	0.0	-0.3	0.6
3SI4	2	0.1	0.1	2.1	0.6	0.1	0.1	2.4	0.7	0.0	0.0	-0.3	0.6
3SI4	27	0.6	0.5	1.8	0.5	0.5	0.5	2.1	0.6	0.0	0.2	-0.2	0.5
3QTO	1	0.0	0.1	1.9	0.6	0.0	0.0	2.2	0.7	0.0	0.1	-0.3	0.6
3QTO	2	0.0	0.0	1.9	0.6	0.0	0.0	2.2	0.7	0.0	0.0	-0.3	0.6
3QTO	15	0.0	0.3	1.7	0.5	0.1	0.1	1.8	0.6	-0.0	0.3	-0.2	0.5
3QX5	1	0.1	0.0	2.2	0.6	0.1	0.0	2.4	0.7	0.0	0.0	-0.3	0.6
3QX5	2	0.1	0.0	2.1	0.6	0.1	0.0	2.4	0.7	0.0	0.0	-0.3	0.6
3QX5	18	0.1	0.1	1.7	0.5	0.1	0.0	1.9	0.6	0.0	0.2	-0.2	0.5
2ZF0	1	0.3	0.2	2.3	0.7	0.3	0.2	2.6	0.8	0.0	0.0	-0.2	0.7
2ZF0	2	0.2	0.2	2.4	0.7	0.2	0.2	2.6	0.8	0.0	0.0	-0.2	0.7
2ZF0	6	0.2	0.2	1.8	0.6	0.2	0.2	1.9	0.6	0.0	0.2	-0.2	0.5
3SI3	1	0.3	0.2	1.6	0.5	0.3	0.2	1.9	0.5	0.0	0.1	-0.3	0.6
3SI3	2	0.2	0.1	1.6	0.5	0.2	0.1	1.8	0.5	0.0	0.0	-0.3	0.6
3SI3	26	0.3	0.2	1.2	0.4	0.2	0.1	1.4	0.4	0.1	0.2	-0.2	0.4
2ZDV	1	0.2	0.1	1.9	0.6	0.2	0.1	2.1	0.7	0.0	0.1	-0.2	0.6
2ZDV	2	0.2	0.1	1.9	0.6	0.2	0.1	2.1	0.7	0.0	0.0	-0.2	0.6
2ZDV	5	0.2	0.2	1.4	0.4	0.2	0.1	1.6	0.5	0.0	0.2	-0.2	0.4
3SHC	1	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.7	0.0	0.0	-0.3	0.6
3SHC	2	0.0	0.0	2.1	0.6	0.0	0.0	2.3	0.7	0.0	0.0	-0.3	0.6
3SHC	25	0.2	0.2	1.5	0.5	0.2	0.2	1.7	0.5	0.0	0.1	-0.2	0.4
2ZC9	1	0.2	0.1	2.3	0.7	0.2	0.1	2.6	0.9	0.0	0.0	-0.3	0.7
2ZC9	2	0.2	0.1	2.3	0.7	0.2	0.1	2.6	0.8	0.0	0.0	-0.3	0.7
2ZC9	3	0.2	0.1	1.6	0.5	0.2	0.1	1.8	0.5	0.0	0.1	-0.2	0.5
2ZFF	1	0.1	0.0	1.9	0.5	0.1	0.0	2.1	0.6	0.0	0.0	-0.2	0.6
2ZFF	2	0.1	0.0	2.0	0.6	0.1	0.0	2.2	0.6	0.0	0.0	-0.2	0.6
2ZFF	7	1.1	0.4	1.6	0.5	1.1	0.4	1.8	0.5	0.0	0.1	-0.1	0.5
3P17	1	0.1	0.1	1.6	0.5	0.1	0.1	1.8	0.6	0.0	0.1	-0.3	0.6
3P17	2	0.1	0.1	1.6	0.5	0.1	0.1	1.8	0.6	0.0	0.0	-0.3	0.6
3P17	14	0.2	0.2	1.2	0.4	0.1	0.1	1.4	0.4	0.0	0.2	-0.2	0.4
3SV2	1	0.4	0.2	1.9	0.6	0.3	0.2	2.1	0.7	0.0	0.1	-0.3	0.6
3SV2	2	0.3	0.2	1.9	0.6	0.3	0.2	2.1	0.7	0.0	0.0	-0.3	0.6
3SV2	28	1.3	0.5	1.5	0.5	1.2	0.6	1.7	0.5	0.0	0.2	-0.2	0.5
4UDW	37	0.1	0.1	2.4	0.7	0.0	0.1	2.7	0.8	0.0	0.1	-0.3	0.7
4UDW	2	0.0	0.0	2.4	0.7	0.0	0.0	2.7	0.8	0.0	0.0	-0.3	0.7
4UDW	1	0.0	0.2	1.8	0.5	0.0	0.0	2.0	0.6	-0.0	0.2	-0.2	0.6
2ZFP	8	0.2	0.1	1.7	0.5	0.2	0.1	2.0	0.6	0.0	0.0	-0.2	0.5
2ZFP	2	0.2	0.1	1.4	0.4	0.2	0.1	1.6	0.5	-0.0	0.1	-0.2	0.4
2ZFP	3	0.1	0.1	1.7	0.5	0.1	0.1	2.0	0.6	0.0	0.0	-0.2	0.5
3DHK	9	0.3	0.2	3.1	0.9	0.3	0.2	3.5	1.0	0.0	0.0	-0.3	0.7
3DHK	2	0.3	0.2	1.9	0.5	0.3	0.2	2.1	0.6	0.0	0.1	-0.2	0.5
3DHK	3	0.3	0.1	3.2	0.9	0.3	0.1	3.6	1.0	0.0	0.1	-0.4	0.8
CC05	1	0.2	0.1	2.2	0.6	0.2	0.1	2.5	0.7	0.0	0.0	-0.3	0.6
CC05	2	0.2	0.1	2.2	0.6	0.2	0.1	2.5	0.7	0.0	0.0	-0.3	0.6
CC05	40	0.2	0.1	1.6	0.4	0.2	0.1	1.8	0.5	0.0	0.1	-0.2	0.4
CC08	2	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.8	0.0	0.0	-0.3	0.6
CC08	41	0.0	0.0	2.0	0.6	0.0	0.0	2.3	0.7	0.0	0.0	-0.3	0.6
CC08	1	0.5	0.4	1.9	0.6	0.5	0.4	2.2	0.7	0.0	0.1	-0.2	0.6
CC10	42	0.0	0.0	2.2	0.7	0.0	0.0	2.5	0.8	0.0	0.0	-0.3	0.6
CC10	2	0.0	0.0	2.2	0.7	0.0	0.0	2.5	0.8	0.0	0.0	-0.3	0.6
CC10	1	0.3	0.3	1.5	0.5	0.3	0.3	1.7	0.5	0.0	0.2	-0.2	0.4
CC11	2	0.0	0.0	2.2	0.7	0.0	0.0	2.5	0.8	0.0	0.0	-0.3	0.7
CC11	43	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.7	0.0	0.0	-0.3	0.6
CC11	1	0.0	0.1	1.8	0.6	0.0	0.0	2.0	0.6	0.0	0.1	-0.3	0.6
2ZDA	1	0.2	0.1	2.4	0.7	0.2	0.1	2.8	0.9	0.0	0.0	-0.3	0.7
2ZDA	2	0.2	0.1	2.4	0.7	0.2	0.1	2.7	0.8	0.0	0.0	-0.3	0.7
2ZDA	4	1.3	0.5	2.5	0.8	1.2	0.6	2.8	0.9	0.0	0.2	-0.3	0.7
2ZGX	8	0.2	0.2	2.1	0.7	0.2	0.2	2.5	0.9	0.0	0.0	-0.3	0.7
2ZGX	2	1.6	0.6	2.3	0.8	1.6	0.6	2.6	0.9	0.0	0.1	-0.3	0.7

Mapping Solvation Thermodynamics on Building Blocks

2ZGX	4	0.1	0.1	2.2	0.8	0.1	0.1	2.6	0.9	0.0	0.0	-0.4	0.7
2ZO3	9	0.0	0.0	3.6	1.0	0.0	0.0	4.0	1.2	0.0	0.0	-0.4	0.9
2ZO3	2	1.7	0.6	3.1	1.0	1.7	0.7	3.5	1.2	0.0	0.2	-0.4	0.8
2ZO3	4	0.0	0.1	3.7	1.1	0.0	0.0	4.1	1.3	0.0	0.1	-0.4	0.9
5JZY	12	0.3	0.1	2.5	0.8	0.3	0.1	2.9	1.0	0.0	0.0	-0.4	0.8
5JZY	2	1.6	0.6	2.7	0.9	1.5	0.6	3.1	1.0	0.1	0.2	-0.4	0.8
5JZY	4	0.3	0.1	2.7	0.9	0.3	0.1	3.1	1.0	0.0	0.1	-0.4	0.8
3DUX	12	0.1	0.1	2.1	0.6	0.1	0.1	2.4	0.7	0.0	0.0	-0.3	0.6
3DUX	2	0.1	0.1	1.7	0.5	0.1	0.1	1.9	0.6	0.0	0.1	-0.2	0.5
3DUX	3	0.1	0.1	2.1	0.6	0.1	0.1	2.4	0.7	0.0	0.0	-0.3	0.6
5LCE	12	0.0	0.0	2.4	0.7	0.0	0.0	2.8	0.8	0.0	0.0	-0.3	0.7
5LCE	2	0.0	0.0	2.4	0.7	0.0	0.0	2.8	0.8	0.0	0.0	-0.3	0.7
5LCE	39	0.0	0.1	2.2	0.7	0.0	0.0	2.5	0.7	0.0	0.1	-0.3	0.6
4BAO	35	1.7	0.9	3.0	0.9	1.7	0.9	3.4	1.1	0.0	0.1	-0.4	0.9
4BAO	30	0.0	0.0	2.5	0.8	0.0	0.0	2.9	0.9	0.0	0.0	-0.4	0.8
4BAO	4	0.0	0.0	2.8	0.9	0.0	0.0	3.3	1.0	0.0	0.0	-0.4	0.9
4BAN	32	1.7	0.9	2.6	0.7	1.7	0.9	2.9	0.9	0.0	0.2	-0.3	0.6
4BAN	30	0.0	0.0	2.0	0.6	0.0	0.0	2.2	0.7	0.0	0.0	-0.2	0.5
4BAN	4	0.0	0.0	2.4	0.7	0.0	0.0	2.7	0.8	0.0	0.0	-0.3	0.6
4BAN	13	0.0	0.0	2.7	0.8	0.0	0.0	3.0	0.9	0.0	0.0	-0.3	0.7
4BAM	33	0.9	0.4	2.7	0.8	0.8	0.4	3.0	0.9	0.1	0.1	-0.3	0.6
4BAM	30	0.0	0.0	2.5	0.7	0.0	0.0	2.8	0.8	0.0	0.0	-0.3	0.6
4BAM	4	0.0	0.0	2.8	0.8	0.0	0.0	3.2	1.0	0.0	0.0	-0.4	0.7
4BAM	34	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.7	0.0	0.0	-0.3	0.6
4BAQ	32	0.8	0.2	2.9	0.8	0.7	0.2	3.3	1.0	0.0	0.2	-0.3	0.8
4BAQ	30	0.0	0.0	2.7	0.8	0.0	0.0	3.0	0.9	0.0	0.0	-0.3	0.7
4BAQ	4	0.0	0.0	3.1	0.9	0.0	0.0	3.4	1.0	0.0	0.0	-0.4	0.8
4BAQ	36	0.0	0.0	2.1	0.6	0.0	0.0	2.3	0.7	0.0	0.0	-0.2	0.6
4BAK	31	1.5	0.4	2.7	0.8	1.4	0.4	3.1	0.9	0.0	0.1	-0.3	0.7
4BAK	32	0.2	0.1	2.5	0.7	0.2	0.1	2.9	0.8	0.0	0.0	-0.3	0.7
4BAK	30	0.2	0.1	3.0	0.9	0.2	0.1	3.3	1.0	0.0	0.0	-0.4	0.8
4BAK	4	0.0	0.0	2.1	0.6	0.0	0.0	2.3	0.7	0.0	0.0	-0.2	0.5
3BIU	10	0.0	0.0	2.0	0.7	0.0	0.0	2.4	0.8	0.0	0.0	-0.4	0.8
3BIU	2	1.6	0.6	2.5	0.8	1.5	0.6	2.9	0.9	0.0	0.1	-0.4	0.7
3BIU	4	0.0	0.0	2.1	0.8	0.0	0.0	2.5	0.8	0.0	0.0	-0.4	0.8
3BIV	11	0.0	0.0	2.3	0.7	0.0	0.0	2.6	0.9	0.0	0.0	-0.4	0.7
3BIV	2	1.6	0.8	2.4	0.7	1.6	0.8	2.7	0.9	0.0	0.2	-0.3	0.6
3BIV	4	0.0	0.0	2.2	0.7	0.0	0.0	2.6	0.8	0.0	0.0	-0.3	0.7
3RML	19	0.0	0.0	1.2	0.4	0.0	0.0	1.3	0.4	0.0	0.0	-0.1	0.3
3RML	20	0.0	0.0	1.1	0.3	0.0	0.0	1.3	0.4	0.0	0.0	-0.1	0.3
3RML	2	0.0	0.0	1.2	0.4	0.0	0.0	1.4	0.4	0.0	0.0	-0.2	0.3
3RML	24	0.0	0.1	1.1	0.3	0.0	0.0	1.3	0.4	0.0	0.1	-0.1	0.3
3RMM	19	0.2	0.1	1.7	0.5	0.2	0.1	1.9	0.6	0.0	0.0	-0.2	0.4
3RMM	21	0.1	0.0	1.4	0.4	0.1	0.0	1.6	0.5	0.0	0.0	-0.2	0.4
3RMM	2	1.0	0.4	1.7	0.5	1.0	0.4	2.0	0.6	0.0	0.1	-0.2	0.5
3RMM	24	0.2	0.1	1.7	0.5	0.2	0.1	2.0	0.6	0.0	0.0	-0.2	0.5
3RMN	19	0.0	0.0	1.7	0.5	0.0	0.0	2.0	0.6	0.0	0.0	-0.2	0.4
3RMN	22	0.0	0.0	1.7	0.5	0.0	0.0	1.8	0.5	0.0	0.0	-0.2	0.4
3RMN	2	1.0	0.5	1.6	0.5	0.9	0.5	1.9	0.5	0.0	0.1	-0.2	0.4
3RMN	24	0.0	0.0	1.8	0.5	0.0	0.0	2.1	0.6	0.0	0.0	-0.2	0.5
3T5F	19	0.2	0.1	1.9	0.6	0.2	0.1	2.1	0.6	0.0	0.0	-0.2	0.5
3T5F	29	0.0	0.0	1.7	0.5	0.0	0.0	1.9	0.5	0.0	0.0	-0.2	0.4
3T5F	2	1.2	0.6	1.8	0.5	1.1	0.6	2.0	0.6	0.0	0.1	-0.2	0.4
3T5F	24	0.6	0.3	1.9	0.6	0.6	0.3	2.2	0.7	0.0	0.1	-0.2	0.5
3RMO	19	0.0	0.0	2.0	0.6	0.0	0.0	2.2	0.7	0.0	0.0	-0.2	0.5
3RMO	23	0.0	0.0	1.9	0.5	0.0	0.0	2.1	0.6	0.0	0.0	-0.2	0.5
3RMO	2	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.7	0.0	0.0	-0.3	0.5
3RMO	24	1.0	0.6	1.8	0.5	1.0	0.6	2.1	0.6	0.0	0.1	-0.2	0.5
3RLW	19	0.1	0.0	1.2	0.4	0.1	0.0	1.3	0.4	0.0	0.0	-0.2	0.3
3RLW	20	2.1	0.9	1.8	0.6	2.1	0.9	2.1	0.6	0.0	0.2	-0.2	0.5
3RLW	2	0.0	0.0	0.8	0.3	0.0	0.0	0.9	0.3	0.0	0.0	-0.1	0.2
3RLW	4	0.1	0.0	1.3	0.4	0.1	0.0	1.5	0.5	0.0	0.0	-0.2	0.3
3RLY	19	0.2	0.1	1.5	0.5	0.2	0.1	1.7	0.6	0.0	0.0	-0.2	0.4
3RLY	21	1.3	0.6	1.9	0.6	1.2	0.6	2.1	0.7	0.0	0.1	-0.3	0.5
3RLY	2	0.1	0.0	1.3	0.4	0.1	0.0	1.5	0.5	0.0	0.0	-0.2	0.4
3RLY	4	0.2	0.1	1.5	0.5	0.2	0.1	1.7	0.6	0.0	0.0	-0.2	0.4
3RMO	19	0.1	0.1	1.6	0.5	0.1	0.1	1.9	0.6	0.0	0.0	-0.2	0.5
3RMO	22	1.7	0.8	2.4	0.7	1.7	0.8	2.7	0.8	0.0	0.1	-0.3	0.6
3RMO	2	0.0	0.0	1.5	0.5	0.0	0.0	1.7	0.5	0.0	0.0	-0.2	0.4

Chapter 3

3RM0	4	0.1	0.1	1.7	0.5	0.1	0.1	1.9	0.6	0.0	0.0	-0.2	0.5
3UWJ	19	0.1	0.1	1.5	0.4	0.1	0.1	1.7	0.5	0.0	0.0	-0.2	0.4
3UWJ	29	1.9	1.1	2.2	0.6	1.9	1.2	2.5	0.7	0.0	0.1	-0.3	0.6
3UWJ	2	0.0	0.0	1.3	0.4	0.0	0.0	1.5	0.5	0.0	0.0	-0.2	0.4
3UWJ	4	0.1	0.1	1.6	0.5	0.1	0.1	1.8	0.6	0.0	0.0	-0.2	0.5
3RM2	19	0.1	0.1	1.6	0.5	0.1	0.1	1.8	0.6	0.0	0.0	-0.2	0.5
3RM2	23	2.1	1.3	1.8	0.5	2.1	1.3	2.0	0.6	0.0	0.2	-0.2	0.5
3RM2	2	0.0	0.0	1.6	0.5	0.0	0.0	1.8	0.5	0.0	0.0	-0.2	0.5
3RM2	4	0.1	0.1	1.8	0.5	0.1	0.1	2.0	0.6	0.0	0.0	-0.2	0.5
5JFD	19	0.2	0.1	2.3	0.7	0.2	0.1	2.6	0.8	0.0	0.0	-0.3	0.6
5JFD	38	0.1	0.1	2.0	0.6	0.1	0.1	2.3	0.7	0.0	0.0	-0.3	0.6
5JFD	2	1.2	0.6	1.9	0.5	1.1	0.6	2.1	0.6	0.0	0.1	-0.2	0.5
5JFD	24	0.2	0.1	2.5	0.8	0.2	0.1	2.9	0.9	0.0	0.0	-0.4	0.8
6GBW	44	0.0	0.0	2.1	0.6	0.0	0.0	2.4	0.7	0.0	0.0	-0.2	0.6
6GBW	2	0.0	0.0	1.4	0.5	0.0	0.0	1.6	0.5	0.0	0.0	-0.2	0.4
6GBW	19	0.1	0.1	1.9	0.6	0.0	0.0	2.1	0.6	0.0	0.2	-0.2	0.5
6GBW	24	0.0	0.0	2.2	0.7	0.0	0.0	2.4	0.8	0.0	0.0	-0.3	0.6

All units are kcal·mol⁻¹. Standard deviations estimated from the test set results of 10 random repetitions of 5-fold cross-validation.

3.7.6 Building Block Thermodynamics for each MRBB.

Table S3-6: Thermodynamic contributions for each BB in each MRBB molecule.

BB	$\Delta G^{(PL)}$ +/- s.d.		$\Delta H^{(PL)}$ +/- s.d.		$T\Delta S^{(PL)}$ +/- s.d.		Cluster Population
1	0.7	0.2	0.8	0.2	-0.1	0.2	0.71 0.23 0.07
2	0.7	0.2	0.8	0.2	-0.1	0.2	0.55 0.36 0.10
3	0.6	0.2	0.7	0.2	-0.1	0.1	0.52 0.48
4	0.9	0.3	1.0	0.3	-0.1	0.2	0.60 0.40
5	0.6	0.2	0.7	0.2	-0.1	0.1	0.51 0.49
6	0.7	0.2	0.7	0.2	-0.1	0.1	0.54 0.46
7	0.6	0.2	0.6	0.2	-0.1	0.1	0.40 0.32 0.28
8	0.5	0.2	0.6	0.2	-0.1	0.2	0.92 0.08
9	1.0	0.3	1.1	0.3	-0.1	0.2	0.99 0.01
10	0.7	0.2	0.8	0.2	-0.1	0.2	0.94 0.04 0.02
11	0.8	0.3	0.9	0.3	-0.1	0.2	0.88 0.12
12	0.8	0.3	0.9	0.3	-0.1	0.2	0.83 0.10 0.08
13	0.3	0.1	0.3	0.1	-0.0	0.1	0.99 0.01
14	0.7	0.2	0.8	0.2	-0.1	0.2	0.52 0.48
15	0.9	0.3	1.0	0.3	-0.1	0.2	0.51 0.49
16	0.9	0.3	1.0	0.3	-0.1	0.2	0.51 0.49
17	1.0	0.3	1.1	0.3	-0.1	0.2	0.52 0.48
18	0.9	0.3	1.1	0.3	-0.1	0.2	0.50 0.50
19	0.5	0.2	0.6	0.2	-0.1	0.1	0.61 0.20 0.18
20	0.6	0.2	0.7	0.2	-0.1	0.1	0.47 0.30 0.23
21	0.6	0.2	0.7	0.2	-0.1	0.1	0.99 0.01
22	0.8	0.2	0.9	0.3	-0.1	0.2	0.82 0.17 0.01
23	1.1	0.3	1.3	0.4	-0.1	0.3	0.57 0.43
24	1.1	0.3	1.2	0.4	-0.1	0.2	0.52 0.48
25	0.7	0.2	0.8	0.2	-0.1	0.1	0.48 0.38 0.14
26	0.6	0.2	0.6	0.2	-0.1	0.1	0.48 0.34 0.18
27	0.9	0.3	1.0	0.3	-0.1	0.2	0.52 0.48
28	0.6	0.2	0.7	0.2	-0.1	0.1	0.52 0.29 0.18
29	0.8	0.3	0.9	0.3	-0.1	0.2	0.98 0.01 0.01
30	0.6	0.2	0.7	0.2	-0.1	0.1	0.64 0.36
31	0.4	0.1	0.5	0.1	-0.1	0.1	0.41 0.30 0.28
32	1.2	0.4	1.3	0.4	-0.2	0.3	0.95 0.04 0.01
33	1.2	0.4	1.4	0.4	-0.2	0.3	0.71 0.29
34	0.2	0.1	0.3	0.1	-0.0	0.1	0.52 0.48
35	1.2	0.4	1.3	0.4	-0.2	0.3	0.97 0.03
36	0.3	0.1	0.4	0.1	-0.0	0.1	0.60 0.20 0.19
37	0.8	0.2	0.9	0.3	-0.1	0.2	0.53 0.47
38	1.2	0.4	1.3	0.4	-0.2	0.3	0.97 0.03
39	0.8	0.3	0.9	0.3	-0.1	0.2	0.53 0.47
40	0.8	0.2	0.9	0.3	-0.1	0.2	0.52 0.48
41	0.7	0.2	0.8	0.2	-0.1	0.2	0.50 0.50
42	0.7	0.2	0.8	0.2	-0.1	0.2	0.50 0.50
43	0.7	0.2	0.7	0.2	-0.1	0.2	0.51 0.49
44	1.2	0.4	1.4	0.4	-0.2	0.3	0.49 0.44 0.08

All units are kcal·mol⁻¹. Standard deviations estimated from calculations with the sets of parameters that were obtained from 10 random repetitions of 5-fold cross-validation with the test/training sets from the actual ligand molecules.

4 The Role of Water Molecules in Protein-Ligand Dissociation: An Analysis of the Mechanisms and Kinetics of Biomolecular Solvation using Molecular Dynamics

4.1 Abstract

In the following chapter, the mechanism and time scale of desolvation is being analyzed for the protein-ligand dissociation reaction of trypsin and thrombin in complex with benzamidine and *N*-amidinopiperidine. The analysis is carried out using umbrella sampling free energy calculations and *LoCorA* calculations. The *LoCorA* approach is a method for the analysis of residence times of water molecules on the surface of amino acids. It was found that water molecules reside approximately 1.3 ns in the binding pocket of thrombin, whereas in trypsin they are residing one order of magnitude shorter (0.3 ns). This difference is explained with special solvent channels that connect the interior of the binding pocket to bulk solvent environment. The solvent channels are present in thrombin but not in trypsin. Furthermore, the selectivity profiles of benzamidine and *N*-amidinopiperidine are related to a solvent-mediated free energy barrier that is present in thrombin but not trypsin. Also due to the presence of the solvent channels, the water molecules show similar residence time for both complexes in the case of thrombin but differing residence times in the case of the two trypsin complexes.

4.2 Introduction

The study of drug-protein association kinetics is one of the most challenging, but at the same time, one of the most insightful aspects of early-stage drug discovery.^{155,156} It ultimately reveals insights into aspects of the binding mechanism, and in this context provides information about whether binding affinity is dominated by the association or the dissociation process. Nevertheless, the mechanism itself and its various intermediate steps are usually hardly understood and often not accessible on the atomistic level by experimental techniques alone. Often, the lifetime of several intermediate steps during association and dissociation remain hidden under the global binding event, but can be elucidated by computer simulations.^{157,158} These intermediate steps can occur on a time-scale which is too fast to be detected by experiments or cannot be sufficiently discriminated from other steps in the process. From all these intermediate steps, solvation and desolvation of drug molecules are one of the most intriguing yet unknown events. It was already noted earlier that they play a crucial role in the association process of G-protein-coupled receptors¹⁵⁹ or Hsp90¹⁵⁶. From an experimental perspective, several techniques have emerged for the investigation of hydration dynamics of biomolecules, such as terahertz spectroscopy¹⁶⁰, NMR¹⁶¹⁻¹⁶⁴ or femto-second infrared spectroscopy¹⁶⁵. In addition, computer simulations have been used to complemented experimental results and gain an in-depth understanding on the atomistic level.^{166,167}

During the protein-ligand association process, a ligand molecule (i.e. a drug or substrate molecule) undergoes desolvation, i.e. it loses its hydration layer, and binds to the protein binding pocket. Similarly, during the dissociation process, the ligand molecule as well as the binding site must both resolvate themselves by several layers of water molecules. However, not only the end-states of this process (i.e. the fully bound or fully unbound states) must be considered, but also intermediate steps along the association/dissociation path. Alongside these complex steps, other intermediate interactions are possible, such as the attachment of a ligand to apolar surface patches of the protein.¹⁵⁷

The acknowledgement of biomolecular solvation in the context of binding thermodynamics is contrasted by the lack of research that is devoted to mechanistic insights and kinetics of biomolecular solvation. Consequently, we likely miss a considerable portion of putative drug molecules exhibiting solvation directed selectivity profiles, due to our lack of understanding of solvation mechanistic features. Furthermore, many important endogenous substrate molecules (such as peptides), are likely tailored with respect to their solvation and desolvation mechanisms

in order to achieve an optimal selectivity profile. It is already known for DNA molecules that the time-scale of hydration processes is ultimately linked to their structure and function.¹⁶⁸

In the current contribution, we present a systematic study on the role of water molecules during the dissociation process of *N*-amidinopiperidine and benzamidine from the serine proteases trypsin and thrombin. The two proteins are both from the large family of serine proteases and are very similar in the structural arrangement next to the catalytic center. The studied ligands molecules are both fragment-like in size and reminiscent of drugs like Melagatran^{169,170} or the natural peptide substrates.¹⁴⁹ The two ligand molecules bind with opposing preference to both proteins (see Table 4-1). Furthermore, they share a very similar binding mode in the binding pockets of thrombin and trypsin (cf. Figure 4-1C/E and D/F). However, the *apo* forms of both proteins display two completely different water structures surrounding the charged side chain of Asp198: In *apo* thrombin, the carboxylate group of Asp198 is solvated by a network of three water molecules (see Figure 4-1A), whereas in trypsin, the same carboxylate group is solvated by only two water molecules (see Figure 4-1B). Since data from neutron diffraction are available for trypsin, disclosing details about the orientation of hydrogen atoms, the water molecules seem to be able to adopt two different configurations in which their orientations are mutually depended on each other. Most interestingly, a water inventory, called water reservoir, is found below Asp189 in the case of trypsin. In the case of thrombin, the water reservoir is replaced by a water channel, which facilitates the water exchange with bulk water molecules.

We will elucidate the mechanism of the binding process, by analysis of the Potential of Mean Force (PMF) along the reaction coordinate of the protein-ligand dissociation by means of Umbrella Sampling (US). For each individual window along the reaction coordinate, we will investigate the mean residence time (MRT) of translation of the water molecules that assemble around key residues in the binding site or around the ligand molecule. For this temporal characterization of the solvation mechanism, we will use the Local Correlation Analysis (*LoCorA*) approach, which was partly introduced in our previous contribution.⁵⁴ We will analyze the temporal properties of the water molecules qualitatively in order to understand the functional role of the water reservoir and water channels in trypsin and thrombin, respectively. We found that the solvation of the *apo* binding pocket of thrombin and trypsin occur on completely different time-scales. In thrombin, water molecules are seemingly stable in the *apo* binding pocket and do only exchange on the scale of nanoseconds. On the contrary, water molecules in the *apo* binding pocket of trypsin exchange approximately one order of magnitude faster than in thrombin. This difference in exchange rate is due to the presence of water channels

in the binding pocket of thrombin, which are lacking in trypsin. However, trypsin has a reservoir instead of water channels, which facilitate the unbinding of ligand molecules. Due to the fact that water molecules can readily exit the binding pocket of thrombin through a different path than the ligand molecules enter the binding pocket, the exchange rate of solvent molecules in the binding pocket does not vary between different protein-ligand complexes. However, in the case of trypsin the solvent exchange rate in the binding pocket greatly varies between the two complexes. Furthermore, the binding mechanism of the ligand molecules critically depends on the presence of water molecules in intermediate states. In these states, water molecules can intercalate between key residues of the protein and the ligand molecule. This intercalation behavior is also reflected by high water residence times in these states.

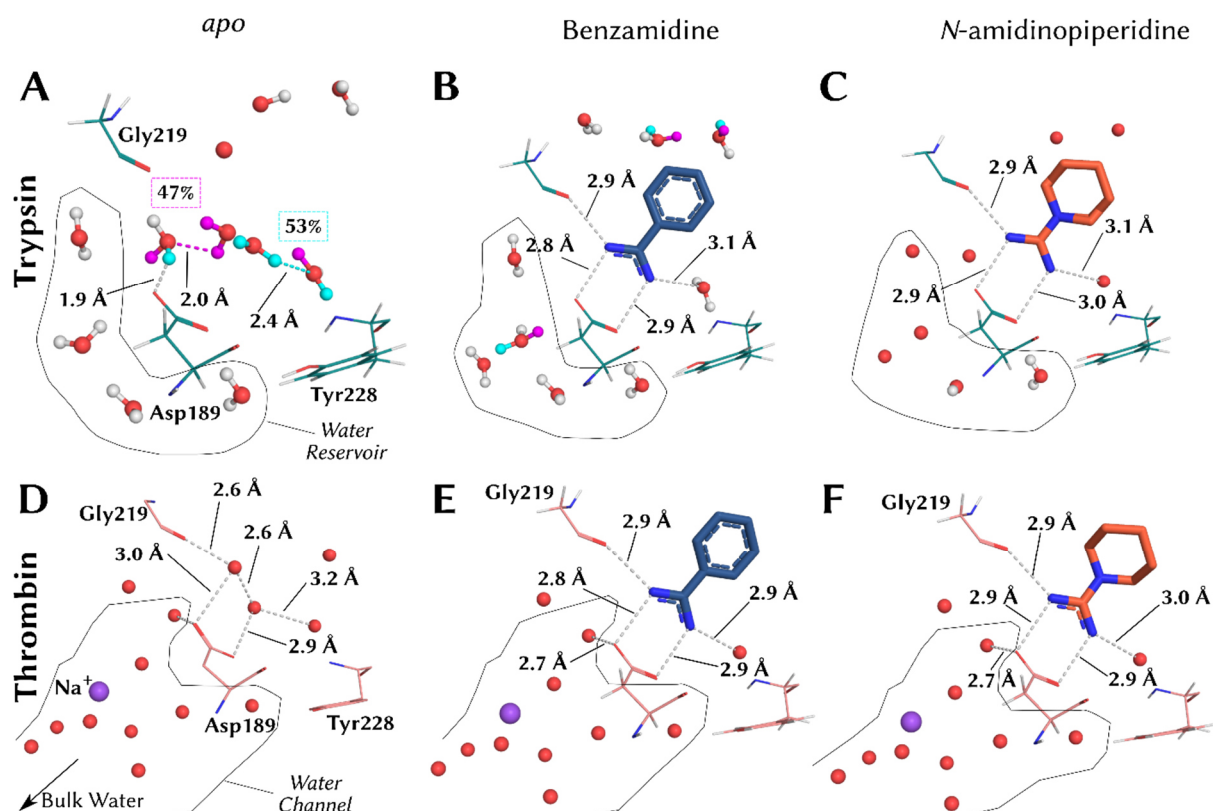
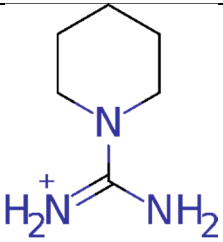
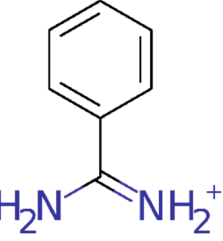


Figure 4-1: Experimentally determined structures of trypsin (top row, neutron structures) and thrombin (bottom row, X-ray structures) of the S₁ subpocket. **A,D:** thrombin and trypsin in their apo state; **B,E:** in complex with benzamidine; **C,F:** in complex with N-amidinopiperidine. Structures **A,B,C** are based on a joint refinement from neutron/X-ray scattering. The different colors (magenta, cyan) of apo trypsin (**A**) indicate two mutually exclusive water configurations (47% and 53% populated)⁵⁴.

Table 4-1: Experimental binding affinities for N-amidinopiperidine and Benzamidine.

Ligand (Selectivity Index)	Trypsin ^{a)}	Thrombin ^{b)}
 <p>N-amidinopiperidine (1.86)</p>	<p>$366 \pm 105 \mu\text{M}$ $-4.68 \pm 0.17 \text{ kcal} \cdot \text{mol}^{-1}$</p>	<p>$197 \pm 74 \mu\text{M}$ $-5.45 \pm 0.24 \text{ kcal} \cdot \text{mol}^{-1}$</p>
 <p>Benzamidine (0.05)</p>	<p>$23.8 \pm 5.3 \mu\text{M}$ $-6.31 \pm 0.12 \text{ kcal} \cdot \text{mol}^{-1}$</p>	<p>$455 \pm 109 \mu\text{M}$ $-4.57 \pm 0.17 \text{ kcal} \cdot \text{mol}^{-1}$</p>

The binding affinities of *N*-amidinopiperidine and benzamidine towards thrombin and trypsin are reported in terms of K_d (upper value) and ΔG^0 (lower value).

a) Reference Schiebel et al.⁵⁴

b) Reference Rühmann et al.¹⁷¹

4.3 Theoretical Background

In this section, we will elaborate on the theoretical background and underlying principles that are part of the *LoCorA* approach used in this study. In the first part, we will introduce the local coordinate systems aligned to the solute as well as the solvent molecules. We will describe, how our approach enabled us to obtain positions and orientations of solvent molecules with respect to the positions and orientations of solute molecules. In the second part of this Theoretical Background section, we will introduce the concept how to calculate the translational and orientational time correlation functions (TCF). Furthermore, we will introduce a weighted double-exponential decay function that we used to explain the computed TCF as the basis for all further temporal-mechanistic considerations.

4.3.1 Local Coordinate Systems

The acronym *LoCorA* stands for Local Correlation Aalysis, and is a approach to derive translational and orientational MRT of water molecules in the local coordinate system of solute molecules. Local coordinate systems are assigned to amino acid side chains and the ligand molecules by using individual subsets of atoms on the respective solute moiety S . For the x-axis ($\vec{R}^{(S,x)}$), a subset of two atoms is used in order to define a position vector. For the z-axis ($\vec{R}^{(S,z)}$), a subset of three or more atoms is used and each distinguishable combination of position vectors without consideration of order from this subset of atoms is used to build a set of planes. The mean orientation vector of these planes, i.e. the vector perpendicular to the plane, then gives the z-axis of the local coordinate system. Finally, the y-axis ($\vec{R}^{(S,y)}$) is calculated as the cross product of the x-axis and z-axis. The origin ($\vec{R}^{(S,0)}$) is calculated as the mean position vector from the atoms used to define the three coordinate axis.

For instance, in the local coordinate system assigned to the tyrosine side chain, the x-axis was defined by a vector connecting the C_γ and C_ζ atoms, whereas the z-axis was defined by the plane spanned by the C_γ , C_δ , C_ϵ and C_ζ atoms (see Figure 4-2D). The origin was placed in the center of the aromatic ring.

For each water molecule j , an internal coordinate system with axis vectors $\vec{W}_j^{(S,x)}$, $\vec{W}_j^{(S,y)}$, $\vec{W}_j^{(S,z)}$ and origin $\vec{W}_j^{(S,0)}$ with respect to the local coordinate system of solute S (defined by $\vec{R}^{(S,x)}$, $\vec{R}^{(S,y)}$, $\vec{R}^{(S,z)}$ and origin $\vec{R}^{(S,0)}$) is defined using the following axis definitions (see also Figure 4-2E): The x-axis of the water molecule is defined as the O-H bond vector (based on

hydrogen atom H1), whereas the z-axis is defined as the vector perpendicular to the plane spanned by the two O-H bond vectors. The y-axis is calculated as the cross-product of x-axis and z-axis. For any S , only the water molecules within the first hydration shell of the atom subset used for the definition of S are considered (see Results section).

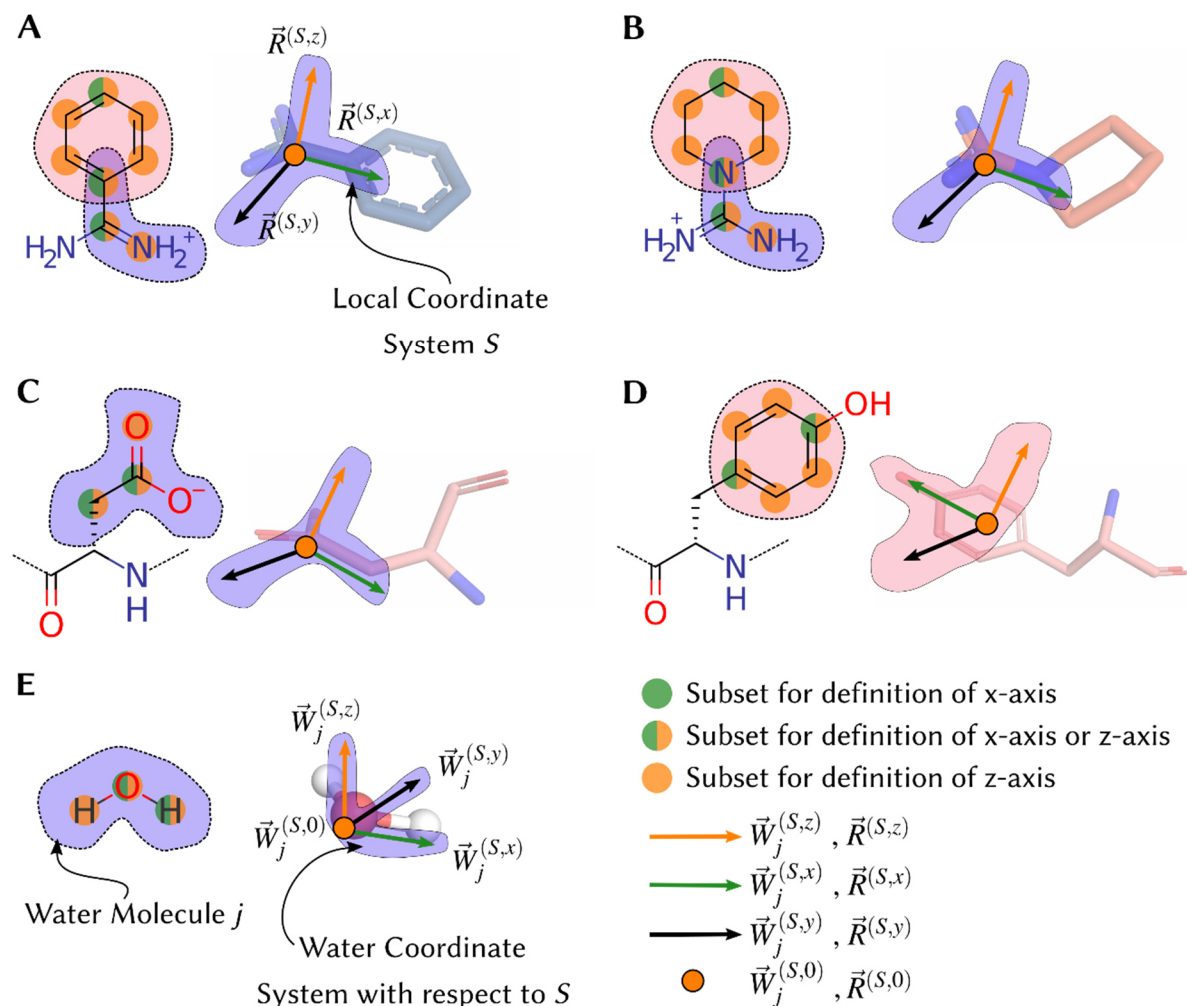


Figure 4-2: Definition of the local coordinate systems for (A) benzamidine, (B) N-amidinopiperidine, (C) aspartate side chain, (D) tyrosine side chain, (E) water.

4.3.2 Calculation of Mean Residence Times

In the previous paragraph, we have defined an internal coordinate system for the water molecules in the reference frame of a solute molecule (with corresponding coordinates for the origin). From that, we will now derive expressions that allow for the calculation of water MRT based on a bimodal process that comprises a slow and a fast relaxation component.

For each water molecule j , we define a survival function $B_j^{(S)}(t)$ that indicates, if at time t the origin of the water coordinate system, $\vec{W}_j^{(S,0)}$, assigned to water molecule j is part of the first hydration shell of the solute atom subset ($b_t = 1$) or not ($b_t = 0$):

$$B_j^{(S)}(t) = \left\{ b_0, b_1, \dots, b_t, \dots, b_{N_f} \right\}_j^{(S)}$$

$$b_t \in [0,1]$$

(4-1)

From the survival function, $B_j^{(S)}$ (eq. (4-1)), the time-correlation function (TCF) for the translation, $C_{trans}^{(S)}$, and orientation, $C_{orient}^{(S,x)}$, $C_{orient}^{(S,y)}$, $C_{orient}^{(S,z)}$ for all water molecules, N_f , is calculated from the temporal evolution of a molecular system. The TCF describes the self-correlation of the (binary) water population between different points in time of the system. The time between two points in time is called the lag time t' . The translational and orientational TCFs are defined as follows (for the orientation only the TCF for x is shown, but the TCF for the y and z components are defined analogously):

$$C_{trans}^{(S)}(t') = \sum_{j=0}^{N_w} \sum_{t=0}^{N_f-t'} B_j^{(S)}(t') \prod_{k=t}^{t'+t} B_j^{(S)}(k)$$

(4-2)

$$C_{rot}^{(S,x)}(t') = \sum_{j=0}^{N_w} \sum_{t=0}^{N_f-t'} P_n \left(\vec{W}_{j,t'}^{(S,x)} \cdot \vec{W}_{j,(t+t')}^{(S,x)} \right) B_j^{(S)}(t') \prod_{k=t}^{t'+t} B_j^{(S)}(k)$$

(4-3)

In eq. (4-3), the function $P_n \left(\vec{W}_{j,t'}^{(S,x)} \cdot \vec{W}_{j,(t+t')}^{(S,x)} \right)$ represents the n -th order Legendre Polynomial of the scalar product of the axis-vectors $\vec{W}_{j,t'}^{(S,x)}$ and $\vec{W}_{j,(t+t')}^{(S,x)}$. In the present work, we use the

1st order Legendre Polynomial, which is simply $P_1(x) = x$. It must be noted that all TCFs were normalized, such that $C(0) = 1$.

From the definition of the TCF in eqs. (4-2) and (4-3), one cannot directly obtain a quantitative estimate of the MRT of the water molecules. Therefore, we follow an approach of Pettit *et al.*¹⁷² and fitted the TCF from eqs. (4-2) and (4-3) to a double-exponential decay that reflects the bimodal behavior of hydration water:

$$C_{trans}^{(S)}(t') = w_{trans}^{(S)} \cdot \exp\left(-\frac{t'}{\tau_{trans,1}^{(S)}}\right) + (1 - w_{trans}^{(S)}) \cdot \exp\left(-\frac{t'}{\tau_{trans,2}^{(S)}}\right) \quad (4-4)$$

$$C_{orient}^{(S,x)}(t') = w_{rot}^{(S,x)} \cdot \exp\left(-\frac{t'}{\tau_{rot,1}^{(S,x)}}\right) + (1 - w_{rot}^{(S,x)}) \cdot \exp\left(-\frac{t'}{\tau_{rot,2}^{(S,x)}}\right) \quad (4-5)$$

In eqs. (4-4) and (4-5), τ_1 and τ_2 are the MRT for the slow and the fast component of the TCF, respectively. The MRT can also be interpreted in terms of a rate constant via the expression $k = \frac{1}{\tau}$, and thus reveals the number of water molecules per unit time that undergo diffusion away from the first hydration shell of the solute site. The weighting factor w , is constrained to be on the interval [0,1] and is effectively proportional to the number of water molecules that undergo slow (τ_1) or fast (τ_2) exchange with the environment beyond the first hydration shell. Note that eq (4-5) contains the TCF of the x component of the local frame of the water molecules. It is needless to say that the same equation will be used analogously for the calculation of the MRTs of the y and z components.

In the work of Pettitt *et al.*, eqs. (4-4) and (4-5) were applied in the calculation of rate constants for the MRT of water molecules in spherical hydration sites and included the prefactor W_0 , which accounts for the average number of water molecules occupying a spherical hydration site. In our work, we did not include this prefactor, as our TCF were normalized. Nonetheless, we will report the average number of water molecules that populate a solute.

Our approach is distinct from the formulation of the stable state picture (SSP) of Laage and Hynes,¹⁷³ which was also employed in the calculation of MRT around DNA base pairs.¹⁷⁴ In one very popular approach, first introduced by Impey, Madden, and McDonald (referred to as the IMM approach),¹⁷⁵ a transient recrossing time (also referred to as tolerance time), t^* , is applied in the calculation of $B_j^{(S)}(t)$ in order to account for unsuccessful exchange attempts.

These events typically occur in cases of a low energy barrier between the first and second hydration layer. Once a water molecule has left the first hydration shell and cannot stabilize itself in the second hydration shell within time t^* , it will have to return (recross) into the first hydration shell. This event is treated as if this water would have never left the first hydration layer. As noted elsewhere,¹⁷³ this approach has several caveats, therefore we did not employ it in our studies. This is also justified, because the MRTs of water molecules in our study are mostly far beyond typical values of $t^*=2.0$ ps. Nonetheless, for the purpose of benchmarking we implemented the IMM approach in our program *LoCorA*.

4.4 Results

In the first part of this section, we will investigate the proportions of water molecules assembled around single amino acids and ligand molecules in bulk solvent, which we will use as a point of reference when computing water MRTs and occupancies in proteins. This is followed by a brief analysis of the spatial structure and MRTs of water molecules in the binding pocket of uncomplexed thrombin and trypsin. In the second part of this section, we will elucidate the mechanism of drug dissociation in protein-ligand complexes formed by benzamidine and *N*-amidinopiperidine with trypsin and thrombin. In the last part, we will focus specifically on the role of water molecules and will compute the MRTs of water molecules assembling at key residues along the dissociation path of ligand molecules from the binding site.

We validated our approach by analyzing the translational and orientational MRT of an individual water molecule in pure bulk water. We compared these computed values with the ones from other water models reported in literature as well as with experimental values. We found that our calculated translational MRTs are in agreement with computed values reported in literature as well as with experimental values. The orientational lifetimes differ slightly from the ones reported in literature, which is explained by the different definitions of orientational states. Since the temporal analysis of bulk solvent has already been studied extensively, and here serves purely as a benchmark, we will not discuss it in the main text but provide a detailed analysis in the Supporting Material.

4.4.1 Residence Times of Water Molecules Assembling next to Reference Solute Molecules

The length of the MRT is generally very sensitive to the definition of the physical states that they are supposed to characterize during the MD simulation. In our case, we investigated the lifetime of water molecules residing at amino acids in protein binding pockets or adjacent to ligand molecules in the bulk phase. During these MD simulations, all molecules were completely unrestrained and were allowed to move freely. We defined (though quite arbitrarily) that a water molecule resides next to an amino acid (or ligand molecule), if it populates the first hydration layer of this amino acid (or ligand molecule). Since the “thickness” and “roughness” (roughness in terms of different intermediate polyhedron geometries) of a hydration layer depends on the environment of the solute, we chose capped amino acids in the pure bulk solvent as reference point. Under these conditions, the amino acids are maximally solvent-exposed and

have minimal influence by other amino acids. As capping groups, we selected acetyl (ACE) for the *N*-terminus and *N*-methyl (NME) for the *C*-terminus in order to mimic the backbone sequence with adjacent amino acid residues (see Figure 4-3A for the 2D-depiction of ACE-*Asp*-NME). Note that we did not use any tolerance time to allow for transient recrossing (i.e. we set $t^*=0$ ps) during the following calculations.

For capped aspartate (ACE-*Asp*-NME), we found the first hydration shell of the (deprotonated) carboxylate group to be best described by water molecules up to 4.1 Å, as indicated by the first local minimum of the radial distribution function (RDF, see purple line in Figure 4-3A). This corresponds to approximately six water molecules in the semi-spherical region around the oxygen atoms of the carboxylate group, as evident from a plot of the number of water molecules ($n_W(r)$, see dashed purple line in Figure 4-3A). Water molecules in this region were previously identified with strong solute-solvent interactions, however they also showed depleted solvent-solvent interactions due to their unfavorable arrangement with respect to each other.¹⁷⁶ The first hydration layer of the tyrosine side chain corresponds to approximately two water molecules (see cyan dashed line in Figure 4-3B) on top of the aromatic portion and exceeds up to 3.5 Å as indicated by the corresponding RDF plot (see cyan line in Figure 4-3A). Note that the second hydration layer of the tyrosine side chain is bigger than the first one (at approximately 6.0 Å), but also contains water molecules coordinating the hydroxyl group (not shown).

In the case of the amidino moiety in *N*-amidinopiperidine and benzamidine, we found that the first hydration shell is confined in a region up to 4.8 Å for both ligands (see Figure 4-3B). This region comprises approximately 10 water molecules (see dashed lines in Figure 4-3B), which are mostly assembling around the amidine hydrogen atoms. Thereby, these water molecules act as hydrogen bond acceptors with respect to their interactions with the positively charged amidino moiety.

For the water MRTs around the charged side chain of the aspartic acid in ACE-*Asp*-NME, we found a slow time component of $\tau_l = 9.7$ ps for the translation of the water molecules (see Table 4-2). For the orientational relaxation of the water molecules, we found $t_1^x = 5.9$ ps for the slow component of the water x-axis with respect to the solute frame of reference defined by the carboxylate group. By that, the relaxation time value of the water x-axis is about 1.5 ps higher than the corresponding relaxation of the z-axis ($t_1^z = 4.3$ ps). This difference in relaxation time behavior is due to the fact that the x-axis of the water molecules corresponds to the O-H bond vector, which is spatially restricted more firmly than the z-axis (perpendicular to the H-O-H

plane), due to hydrogen bonding interactions between the water molecules and the carboxylate group. We assume that while a water molecule interacts with the carboxylate group, one of its O-H bond vectors remains rather fixed and the other O-H bond vector tangles in space. The y-axis of the water molecules shows similar relaxation time for the slow component ($t_1^y = 5.7 \text{ ps}$). Since this y-axis bisects the two O-H bond vectors of the water molecules, we assume that water molecules potentially also undergo interactions with the carboxylate group in which both hydrogen atoms are involved in a hydrogen bond.

In the case of the apolar tyrosine side chain in ACE-Tyr-NME, an MRT of $\tau_l = 3.3 \text{ ps}$ was computed. This value indicates a much faster exchange rate (approx. one third faster) of water molecules from the first hydration shell of the apolar tyrosine side chain compared to the negatively charged aspartate side chain. Also, the orientational relaxation behavior of the water molecules on top of the apolar aromatic side chain seems to be rather isotropic, as all axis from the water coordinate system show quite similar relaxation times ($t_1^z = 2.1 \text{ ps}$, $t_1^y = 2.4 \text{ ps}$, $t_1^x = 2.3 \text{ ps}$) in the solute frame of reference. The relaxation time of the y- and x-axis are slightly elevated, which indicates a weak influence of the hydroxyl group on the orientational behavior of the water molecules.

For the two small molecule ligands, benzamidine and *N*-amidinopiperidine, we computed quite comparable values for the slow components of the τ_l MRT of 4.6 ps and 4.1 ps, respectively. Thus, water molecules assembling at the amidino group show a higher exchange rate compared to the corresponding value found for the carboxylate group of the aspartic acid side chain. The individual components of the orientational relaxation times are quite similar for both molecules, benzamidine and *N*-amidinopiperidine. This is anticipated, as water molecules act as hydrogen bond acceptor towards the amidino group and no preferred orientation of the O-H bond vector, corresponding to the x-axis of the water coordinate system, is expected. This results in a slightly reduced directionality of the three coordinate axis of these water molecules accompanied with an enhanced isotropic orientational behavior.

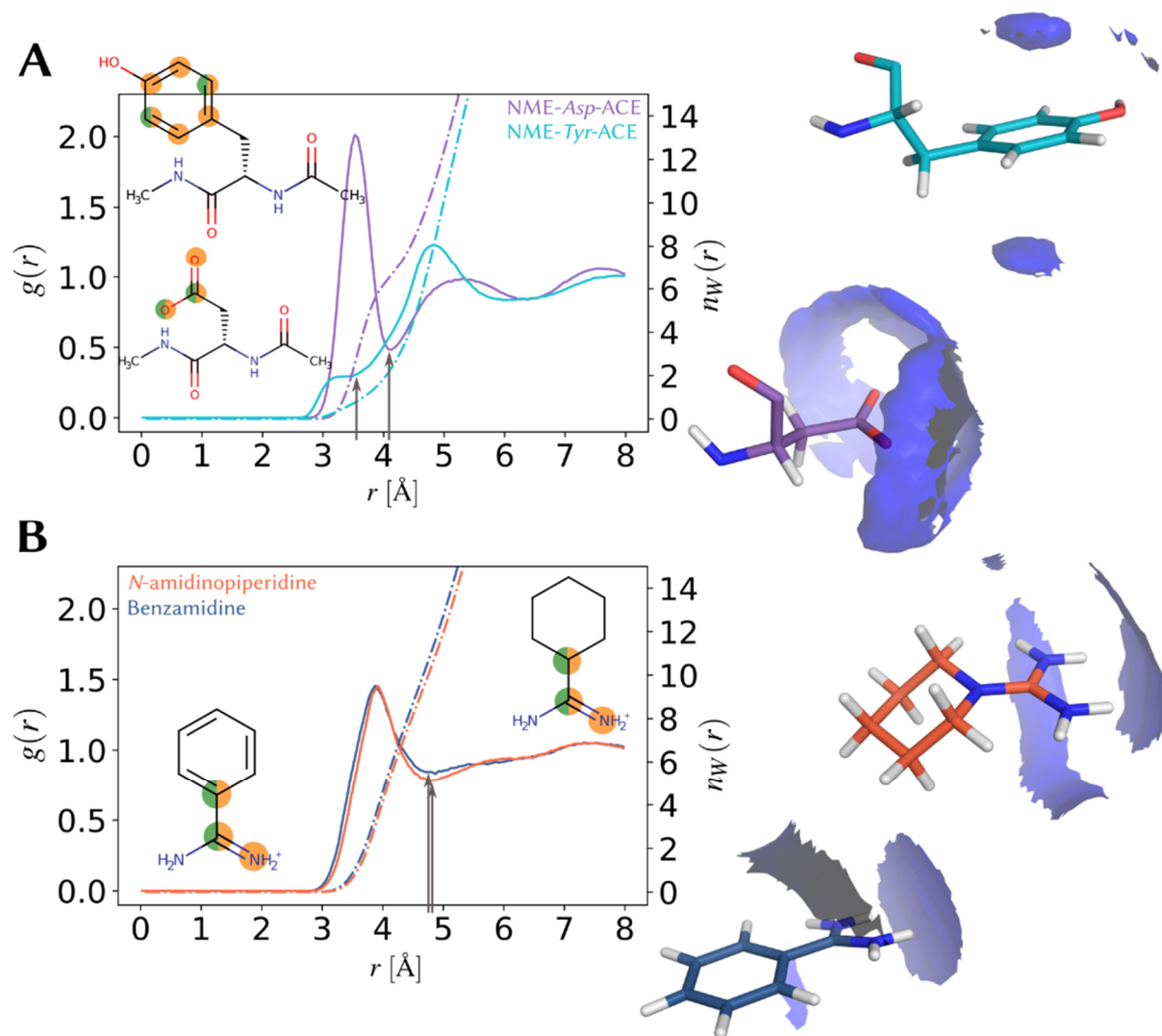


Figure 4-3: Radial distribution functions $g(r)$ (solid lines, plots to the left) and coordination number $n(r)$ (dashed lines, plots to the left) with respect to water oxygen atoms around defined subsets of atoms. The definition of the atom subsets (colored circles imposed to the 2D-depictions) is in accordance with Figure 4-2. **(A)** Amino acid side chain of a capped aspartate residue (purple) and capped tyrosine residue (cyan), **(B)** amidine portion of benzamidine (blue) and *N*-amidinopiperidine (orange). The blue isosurfaces on the right display the distribution of water oxygen atoms countered at $1.5 \rho^0$ (ρ^0 : bulk water density, 0.0332 \AA^{-3}) around the respective solute atom subset. The vertical arrows assigned to the RDF plots indicate the positions of the boundary of the first hydration layer. The coordination number $n(r)$ is calculated from the RDF integral: $n(r) = 4\pi\rho^0 \int_0^r r'^2 g(r') dr'$.

Table 4-2: Overview of water MRTs at reference solute molecules.

Residue	Component	w	τ_1 [ps]	τ_2 [ps]
<i>NME-Asp-ACE</i>	$\langle \bar{W}^{(S,0)} \rangle_{N_f}$	0.4 ± 0.4	9.7 ± 0.2	0.7 ± 0.1
	$\langle \bar{W}^{(S,z)} \rangle_{N_f}$	0.5 ± 0.2	4.3 ± 0.2	0.3 ± 0.1
	$\langle \bar{W}^{(S,y)} \rangle_{N_f}$	0.5 ± 0.2	5.7 ± 0.2	0.4 ± 0.1
	$\langle \bar{W}^{(S,x)} \rangle_{N_f}$	0.5 ± 0.1	5.9 ± 0.1	0.4 ± 0.1
<i>NME-Tyr-ACE</i>	$\langle \bar{W}^{(S,0)} \rangle_{N_f}$	0.5 ± 0.1	3.3 ± 0.2	0.5 ± 0.1
	$\langle \bar{W}^{(S,z)} \rangle_{N_f}$	0.5 ± 0.1	2.1 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,y)} \rangle_{N_f}$	0.5 ± 0.1	2.4 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,x)} \rangle_{N_f}$	0.5 ± 0.1	2.3 ± 0.1	0.3 ± 0.1
<i>Benzamidine</i> (amidine)	$\langle \bar{W}^{(S,0)} \rangle_{N_f}$	0.5 ± 0.2	4.6 ± 0.1	0.7 ± 0.1
	$\langle \bar{W}^{(S,z)} \rangle_{N_f}$	0.5 ± 0.1	2.2 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,y)} \rangle_{N_f}$	0.5 ± 0.1	2.5 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,x)} \rangle_{N_f}$	0.5 ± 0.1	2.4 ± 0.1	0.3 ± 0.1
<i>N-Amidinopiperidine</i> (amidine)	$\langle \bar{W}^{(S,0)} \rangle_{N_f}$	0.4 ± 0.3	4.1 ± 0.1	0.6 ± 0.1
	$\langle \bar{W}^{(S,z)} \rangle_{N_f}$	0.5 ± 0.1	2.1 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,y)} \rangle_{N_f}$	0.5 ± 0.1	2.3 ± 0.1	0.3 ± 0.1
	$\langle \bar{W}^{(S,x)} \rangle_{N_f}$	0.5 ± 0.1	2.3 ± 0.1	0.3 ± 0.1

4.4.2 Mean Residence Time of Water Molecules in the *apo* Protein Binding Pocket

Before any ligand molecule is accommodated in the binding pocket of a protein, the binding pocket is filled with water molecules. These water molecules are, unless they are structurally tightly bound to the protein, able to exchange with other regions on the (solvent accessible) surface of the protein or the bulk water environment. The time-scale of this exchange is largely dependent on the environment of the water molecule in the binding pocket. This dependency is due to interactions of the water molecule and amino acids that form the binding pocket, but also due to the other water molecules that are accommodated in the binding pocket. The shape of the binding pocket as well as the electrostatic properties of the amino acids effectively determine the frequency by which water molecules enter or leave the protein binding pocket. We investigated these time-dependent processes for key amino acids in the protein binding pocket of thrombin and trypsin. Most important are the amino acid residues Asp189 and Tyr228 and are located at the bottom of the S1 specificity pocket (c.f. Figure 4-1). The first hydration shell around these amino acids, which confines the water molecules considered in the MRT calculation, was defined according to our analysis of the RDF computed on the basis of the capped amino acids (i.e. ACE-*Asp*-NME and ACE-*Tyr*-NME) in pure bulk water as introduced in the previous subsection.

The distribution of water molecules in the S1 binding pocket of thrombin and trypsin generally matches well with the positions of the water molecules as found in the crystal structures, which were refined to a resolution of 1.26 Å and 0.99 Å for thrombin¹¹² and trypsin,⁵⁴ respectively. In trypsin, two major solvent sites, W1 and W2, are found adjacent to the carboxylate group of Asp189, even though for solvent site W2 water molecules also populate positions in between the carboxylate group of Asp189 and W3 topping Tyr228 (see Figure 4-4A). In the *apo* binding pocket of thrombin, W1 is located more distal to Asp189 as compared to trypsin, which agrees well with the crystal structure (see Figure 4-4B). Another solvent site, W4, is topping solvent sites W1 and W2 in both *apo* pockets and is located close to the exit of the S1 subpocket. In trypsin, this solvent site is heavily populated, as indicated by the pronounced solvent density distribution of this site. Another solvent site, W3, is found on top of the aromatic portion of the Tyr228 side chain in both proteins. In both proteins, the computed density distribution at this position agrees fairly well with the experimentally determined water molecule. An important structural feature of the two serine proteases thrombin and trypsin, is the water reservoir located below Asp189 (see Figure 4-4A and B). As already noted in our previous contribution,⁵⁴ this reservoir provides water molecules that are needed for the association and dissociation process.

The MRTs of the water molecules in the binding pocket of thrombin are considerably longer than the corresponding MRT values in trypsin. This observation generally holds true for both key amino acids, Asp198 and Tyr225 (see Table 4-3). The slow MRT for Asp198 in thrombin is approximately 1.2 ns on average, which is about 10 times the value (0.15 ns) computed for the same residue in trypsin. However, the consideration of average values is somewhat misleading in the case of thrombin, since MRT results from a bimodal distribution of two separate Gaussian distributions with mean values at 1.2 ns and 2.3 ns (see Figure 4-4D). This bimodal Gaussian distribution clearly explains the high standard deviation from the mean found for this MRT. Furthermore, it indicates the occurrence of two distinct solvation mechanisms present in the case of thrombin, which is not the case in trypsin as evidenced by a uniform monomodal distribution (see Figure 4-4C). In the following, we will only consider the broad distribution at lower τ_l values (about 60%) in order to only capture the slower of the two MRTs in thrombin. In order to remain consistent and consider comparable MRTs for all residues, we performed a similar analysis in all other cases for both, the *apo* proteins and all protein-ligand complexes.

In all cases, the value of the orientation time constant of the water z-axis was lower than for the other two remaining axis. The same observation was made already for the reference solute molecules (see Table 4-2). As noted above, this indicates that once a water molecule establishes a hydrogen bond along its y- or x-axis, it tumbles (i.e. the orientation decays) slower along the axis of this hydrogen bond. Notably, the opposite was found for the water molecules assembling at Asp189 in thrombin. Here, the water z-axis decays with a τ_l time constant of 1153 ps, whereas the y- and x-axis decay at 1030 ps and 1012 ps, respectively. Although the standard deviation of the values is quite large, this notable exception may indicate completely different water rearrangement mechanisms in thrombin compared to trypsin. These water rearrangements may include a dominant pendulum-like movement of the water molecules around their z-axis, which allows for a mutual hydrogen-bond switching between the two hydrogen bonds of a single water molecule.

Interestingly, the MRT value of the fast component, τ_2 , for water molecules residing at Asp189 is lower in case of thrombin compared to trypsin (12.0 ps for thrombin and 15.5 ps for trypsin). In the case of Tyr228, the fast component is extremely low (< 1.5 ps) for trypsin and thus likely corresponds to fast recrossing events between the first hydration shell of the apolar tyrosine side chain and its second hydration layer. However, the τ_l component of the water molecules

at Tyr228 in the case of thrombin (8.6 ps) is more than twice the value of trypsin (3.4 ps). The value for trypsin is very close to the one calculated for capped tyrosine amino acid in bulk solvent ACE-Tyr-NME (3.3 ps), which indicates that the protein environment in trypsin does facilitate fast exchange between water molecules in the first hydration shell of Tyr228. Thus, the protein environment does not perturb the solvation dynamics of Tyr228 in case of trypsin, whereas it clearly perturbs the solvation dynamics in thrombin resulting in an enhanced MRT.

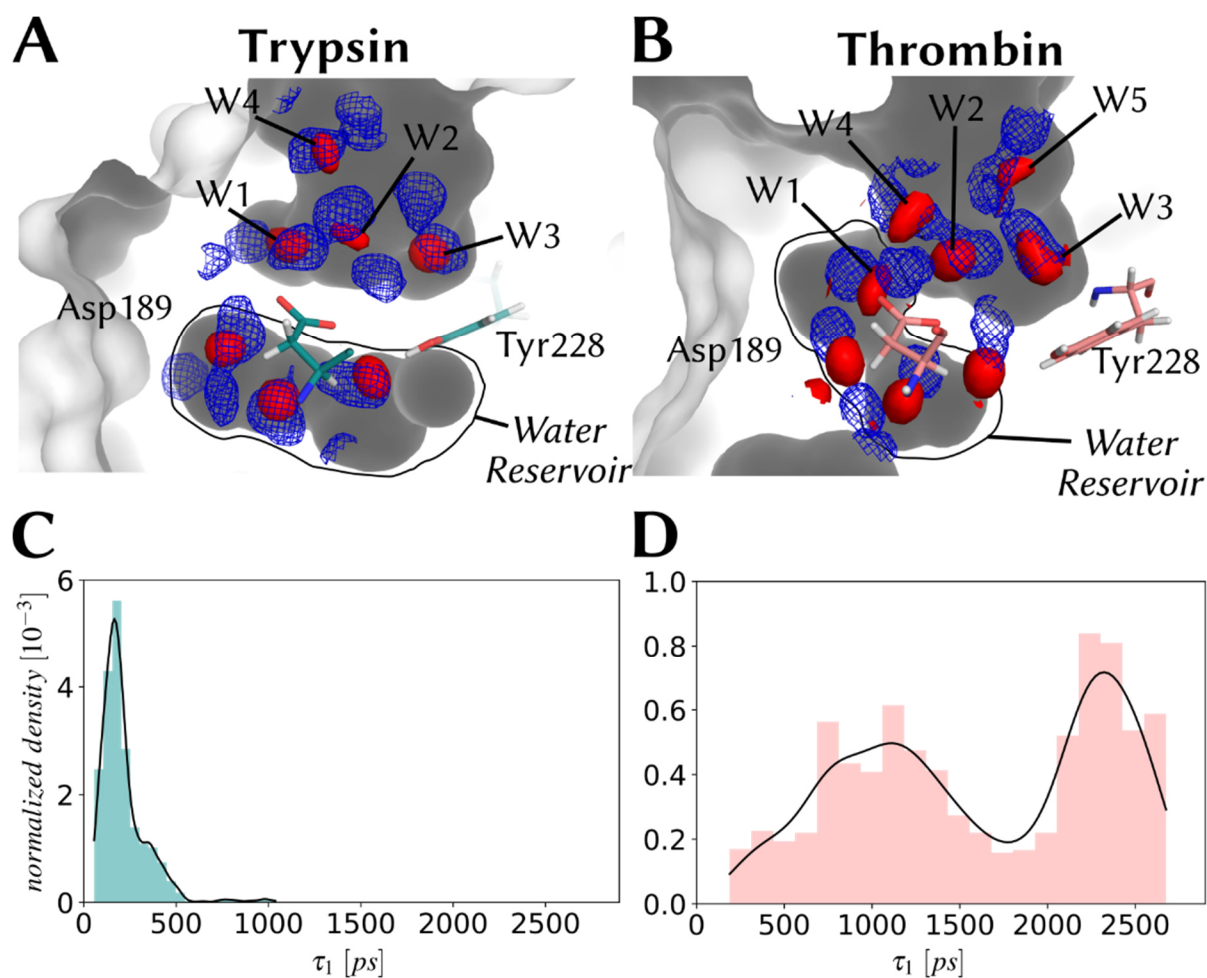


Figure 4-4: Solvent density map (blue mesh) and $2mF_o-DF_c$ electron density (red surface) in the S1 sub pocket of trypsin (**A**) and thrombin (**B**). The solvent density map is contoured at $2\rho^0$ and calculated from the distribution of water oxygen atoms in an MD simulation of the apo protein. The electron density map is contoured at 1.5σ (trypsin: 5MOP, thrombin: 2UUF). The plots in (**C**) and (**D**) show the normalized probability density distribution of the slow component of the MRT, τ_1 , for water molecules solvating Asp189 in trypsin and thrombin, respectively.

Table 4-3: Overview of water MRTs at key residues in the binding site.^{a)}

<i>Protein</i>	<i>Residue</i>	<i>Component</i>	<i>w</i>	<i>τ_1 [ps]</i>	<i>τ_2 [ps]</i>
<i>Thrombin</i>	Asp198	$\langle \bar{W}^{(s,0)} \rangle_{N_f}$	0.5 ± 0.1	1216.4 ± 537.8	12.0 ± 5.1
		$\langle \bar{W}^{(s,z)} \rangle_{N_f}$	0.5 ± 0.2	1153.4 ± 668.3	7.6 ± 2.8
		$\langle \bar{W}^{(s,y)} \rangle_{N_f}$	0.5 ± 0.2	1029.9 ± 444.9	8.4 ± 3.0
		$\langle \bar{W}^{(s,x)} \rangle_{N_f}$	0.5 ± 0.2	1012.2 ± 462.6	8.5 ± 3.3
<i>Trypsin</i>		$\langle \bar{W}^{(s,0)} \rangle_{N_f}$	0.5 ± 0.2	148.5 ± 41.0	15.5 ± 10.2
		$\langle \bar{W}^{(s,z)} \rangle_{N_f}$	0.5 ± 0.1	108.4 ± 34.3	3.2 ± 2.5
		$\langle \bar{W}^{(s,y)} \rangle_{N_f}$	0.5 ± 0.1	120.3 ± 35.6	4.3 ± 3.1
		$\langle \bar{W}^{(s,x)} \rangle_{N_f}$	0.5 ± 0.1	120.4 ± 35.2	4.5 ± 3.2
<i>Thrombin</i>	Tyr228	$\langle \bar{W}^{(s,0)} \rangle_{N_f}$	0.5 ± 0.1	8.6 ± 2.4	1.1 ± 0.4
		$\langle \bar{W}^{(s,z)} \rangle_{N_f}$	0.5 ± 0.1	6.9 ± 1.8	0.5 ± 0.2
		$\langle \bar{W}^{(s,y)} \rangle_{N_f}$	0.5 ± 0.1	7.4 ± 2.0	0.6 ± 0.2
		$\langle \bar{W}^{(s,x)} \rangle_{N_f}$	0.5 ± 0.1	7.3 ± 1.9	0.6 ± 0.2
<i>Trypsin</i>		$\langle \bar{W}^{(s,0)} \rangle_{N_f}$	0.5 ± 0.2	3.4 ± 1.4	0.9 ± 0.2
		$\langle \bar{W}^{(s,z)} \rangle_{N_f}$	0.5 ± 0.1	1.9 ± 0.7	0.4 ± 0.1
		$\langle \bar{W}^{(s,y)} \rangle_{N_f}$	0.5 ± 0.1	2.3 ± 0.9	0.5 ± 0.2
		$\langle \bar{W}^{(s,x)} \rangle_{N_f}$	0.5 ± 0.1	2.2 ± 0.8	0.5 ± 0.1

a) Error indicates ± 1 standard deviation from the mean value.

4.4.3 Investigating the Dissociation Mechanism

In the following, we investigate the dissociation mechanism of benzamidine and *N*-amidinopiperidine from the binding pocket of thrombin and trypsin by means of US simulations. The dissociation path was described by the distance between the geometric center of the amidino moiety of the ligand molecule and the terminal carboxylate group in the side chain of Asp189 found in the bottom of the S1 pocket (see Figure 4-5). Due to the high similarity between the two proteins as well as the two ligand molecules, the assigned reaction coordinate can be universally applied to all four protein-ligand complexes (benzamidine/trypsin, benzamidine/thrombin, *N*-amidinopiperidine/trypsin, *N*-amidinopiperidine/thrombin). The reaction coordinate was scanned from 3 to 10 Å in steps of 0.1 Å, resulting in 71 windows per protein-ligand complex. The PMF was estimated using the weighted histogram analysis method (WHAM) estimator.^{177,178}

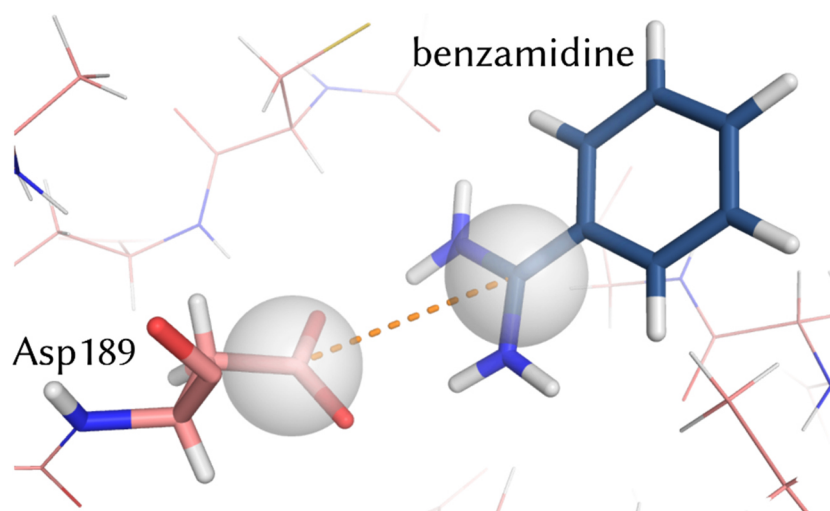


Figure 4-5: Binding pocket of thrombin in complex with benzamidine. The dashed orange line indicates the assigned reaction coordinate used for the umbrella sampling MD simulations. The reaction coordinate is defined as the distance between the centers of the carboxylate group (C_{β} , C_{γ} , $O_{\delta 1}$ and $O_{\delta 2}$) of Asp189 and the amidino moiety of the ligand. Note that the reaction coordinate is defined analogously for the other protein-ligand complexes.

4.4.4 Dissociation Mechanism of Trypsin Complexes

Initially, an overall similar global minimum on the PMF profile along the reaction coordinate was found at $d = 3.3 \text{ \AA}$ in the bound structures of both ligands in trypsin (see Figure 4-6A, state **a**). This global minimum matches perfectly well with the values found in the crystal structures (3.3 Å for *N*-amidinopiperidine and 3.2 Å for benzamidine). Both ligands adopt a bidentate salt bridge with Asp189, which is further stabilized by 2.8 water molecules on average in the case

of benzamidine (see state **a** in Figure 4-6B and the blue line in Figure 4-7A) and 3.3 water molecules in the case of *N*-amidinopiperidine (see state **a** in Figure 4-6C and the orange line in Figure 4-7A).

In the following, benzamidine passes a much steeper barrier on the FES (free energy surface) compared to *N*-amidinopiperidine (see step **b** in Figure 4-6A). This difference in barrier height of approximately 3 kcal·mol⁻¹ is due to the difference in the PMF of the protein-ligand complexes at the intermediate state **b**. The explanation for this difference observed at state **b** are differences in the solvation mechanisms of the amidino groups. At state **b**, water molecules intercalate between the partly dissociated amidino groups of the ligands and the carboxylate group of Asp189. These water molecules originate from a water reservoir located below Asp189 accommodating five water molecules in the case of benzamidine and six water molecules in the case *N*-amidinopiperidine. The water molecules in this reservoir remain fixed as long as the ligand molecule remains fully bound to the protein and thus blocks the only water exchange site to the water reservoir. For both protein-ligand complexes, the number of water molecules in the first hydration shell of Asp189 increases by 0.5 compared to the previous state **a** (see blue and orange lines in Figure 4-7A). In the case of benzamidine, one water molecule intercalates between the carboxylate group of Asp189 and the amidino group (see step **b** in Figure 4-6B), whereas in the case of *N*-amidinopiperidine two water molecules bridge between the ligand and Asp189 (see step **b** in Figure 4-6C). In the case of *N*-amidinopiperidine, these interactions are further stabilized by the water molecule found on top of Tyr228. Since in step **b** the major interaction between the ligand and the protein is broken, it can be attributed as one of the key steps in the ligand dissociation pathway.

In the final dissociation step **c**, the amidino groups of the ligand molecules orient toward the solvent-exposed part of the binding pocket (see step **c** in Figure 4-6B and C), whereas the apolar portion still penetrates into the S1 binding pocket. Both ligands flip their orientation upon dissociation instead of escaping the binding pocket with the apolar part leaving first. In the case of benzamidine, the amidine-carboxylate salt bridge is fully replaced by the coordination of one water molecule. Contrary, in the case of *N*-amidinopiperidine two water molecules interact with the abandoned carboxylate moiety of Asp189, while, at the same time, these water molecules are able to exchange with water molecules from the bulk water phase (see step **c** in Figure 4-6C). Thus, the additional water molecule found for *N*-amidinopiperidine is likely a consequence of the increased flow of water molecules into the pocket at a lower PMF than computed for benzamidine. The total (time-averaged) number of water molecules assembling

around Asp189 increases by 0.7 between the initial state **a** and the final state **c** in both trypsin complexes.

The free energy difference between bound and dissociated state of benzamidine in trypsin amounts to $-7.8 \text{ kcal}\cdot\text{mol}^{-1}$, which overestimates the experimentally determined value ($-6.3 \text{ kcal}\cdot\text{mol}^{-1}$, see also Table 4-1). For the corresponding *N*-amidinopiperidine complex, we computed a free energy of $-4.5 \text{ kcal}\cdot\text{mol}^{-1}$. This value matches rather well with the experimental value of $-4.7 \text{ kcal}\cdot\text{mol}^{-1}$. Note that we did not anticipate the full unbinding mechanism for any of the studied protein-ligand complexes, as in the present work we focused on perturbations of water structure in the binding pocket. The majority of these perturbations take place in the initial phase of the dissociation events once the ligand starts to escape from the binding pocket.

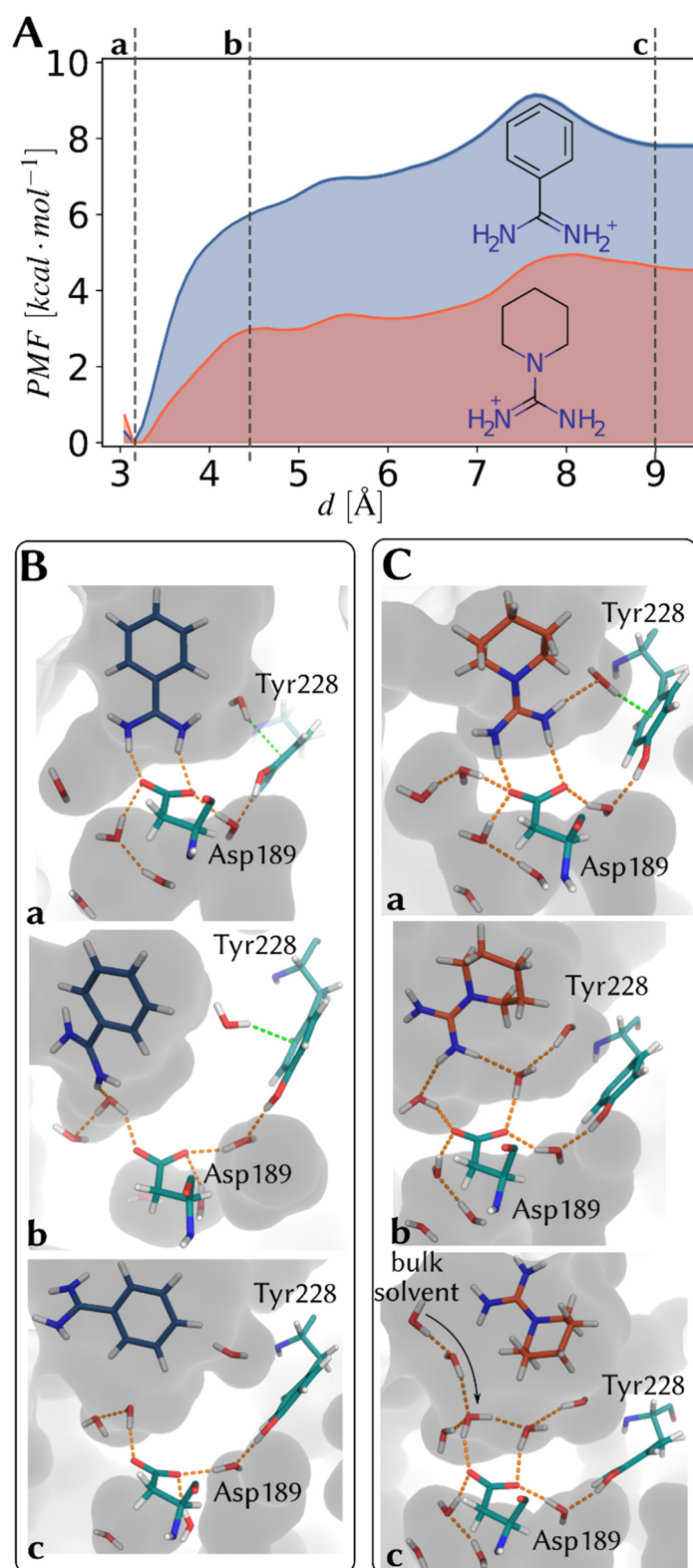


Figure 4-6: Overview of the dissociation mechanism of the protein-ligand complexes of benzamidine-trypsin (left) and N-amidinopiperidine-trypsin (right). **A:** PMF along the reaction coordinate d for the dissociation of trypsin-ligand complexes (see Figure 4-5 for the definition of the reaction coordinate); **B, C:** representative snapshots from the MD simulation at key steps **a, b** and **c** for benzamidine (**B**) and N-amidinopiperidine (**C**).

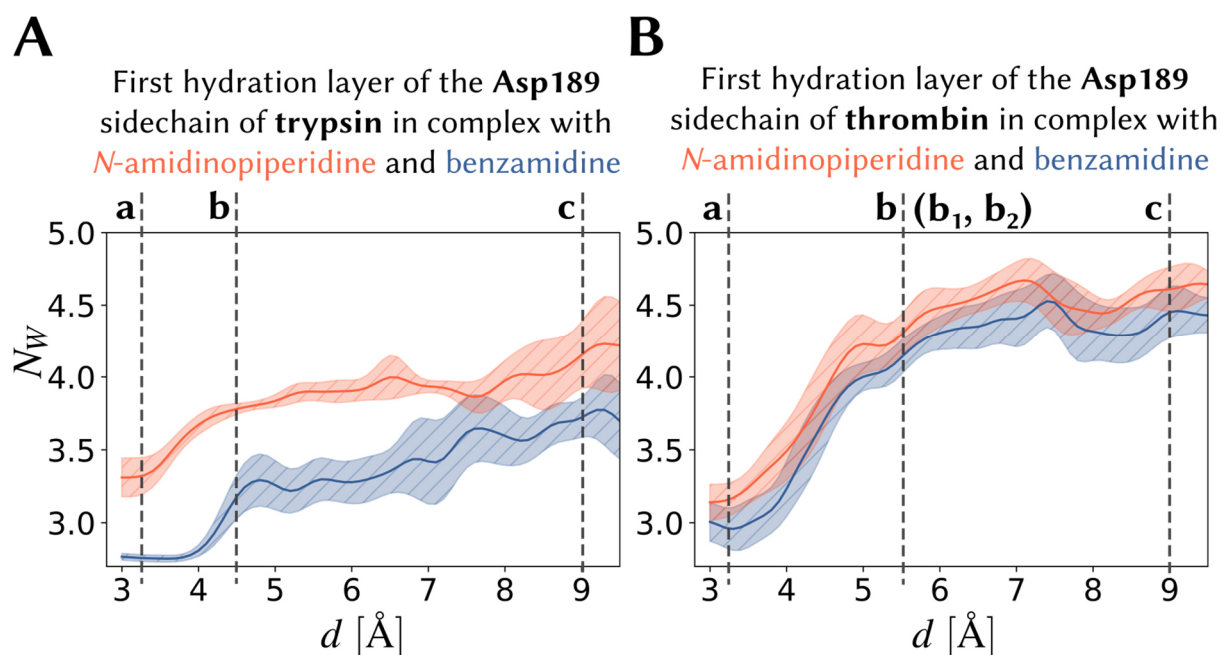


Figure 4-7: Number of water molecules, N_w , in the first hydration ($r < 4.1$ Å) layer of Asp189 along the reaction coordinate of the protein-ligand dissociation in (A) trypsin and (B) thrombin. The orange line corresponds to the protein-ligand complex formed with *N*-amidinopiperidine and the blue line corresponds to the protein-ligand complex formed with benzamide. The semi-transparent areas represent 1 standard deviation.

4.4.5 Dissociation Mechanism of Thrombin Complexes

In the case of thrombin both ligands are bound to the protein by forming a bidentate salt bridge between the amidino moiety and the carboxylate group of Asp189 (see step **a** in Figure 4-8B, C). In both protein-ligand complexes, Asp189 is solvated by approximately 3.0 water molecules on average (see orange and blue lines in Figure 4-7B). These water molecules are in exchange with the bulk solvent via two water channels (see step **a** in Figure 4-8B). Of these, the first one (water channel A) is located below the binding site of the ligand and a second one (water channel B) is located below the backbone atoms of Asp189. It must be noted, that in the crystal structure only water channel B was observed. Additionally, water channel B contains a sodium ion which is coordinated through multiple water molecules. Although water channel B is located at the same site as the water reservoir in trypsin (s. above), they must be treated differently. While the number of water molecules in the water reservoir of trypsin remains fixed up to state **c**, the number of water molecules can vary in the case of thrombin at each step during the ligand dissociation path.

In the following step **b**, both ligand molecules are able to adopt two different binding modes **b₁** and **b₂**. Of these two binding modes, **b₁** represents a stable intermediate with two interstitial water molecules mediating a contact between ligand and carboxylate group of Asp189 (see step

b₁ in Figure 4-8B, C). In the bound state **b**₂, both ligand molecules interact with the water molecule located on top of Tyr228 and another water molecule mediating the contact to Asp189 (see step **b**₂ in Figure 4-8B, C). On average, 4.3 water molecules (see orange line in Figure 4-7B) solvate the carboxylate group of Asp189 in binding mode **b** (i.e. the average over **b**₁ and **b**₂) of the *N*-amidinopiperidine complex. This is already close to the value of 4.5 water molecules that is achieved in the final state **c**. In both protein-ligand complexes, the two binding modes **b**₁ and **b**₂ are populated to about 50% each (see Figure S2 in the Supporting Information). The analogous state **b** in trypsin, occurs only with one single binding mode (see Figure S1 in the Supporting Information). The differences in free energy between states **a** and **b** are approximately 2 kcal·mol⁻¹ for benzamidine and 1 kcal·mol⁻¹ for *N*-amidinopiperidine. Contrary to trypsin, this difference on the FES is not accompanied by a steeper increase of the PMF next to state **a** in the case of benzamidine (see blue line in Figure 4-8A) compared to *N*-amidinopiperidine (see orange line in Figure 4-8A).

In the final state **c**, both ligands orient their amidino function towards the solvent, while still burying the apolar part in the binding pocket. In both cases, Asp189 is fully solvated by approximately 4.5 water molecules in the case of the *N*-amidinopiperidine complex and 4.3 water molecules in the case of benzamidine complex (see Figure 4-7B). In both complexes a rather similar rise in the number of water molecules assembling around Asp189 was observed. The time-averaged number of water molecules at this site increases by 1.5 in both complexes. This value is twice as high as in trypsin, indicating an enhanced flow of water molecules through the water channels in the case thrombin.

We computed a free energy difference of -2.3 kcal·mol⁻¹ between the bound and dissociated state of the thrombin-benzamidine complex. For the thrombin-*N*-amidinopiperidine complex, we computed a free energy difference of -1.5 kcal·mol⁻¹. Both values are calculated smaller compared to experimentally determined values and also suggest a different affinity ranking compared to experiment (-4.6 kcal·mol⁻¹ for benzamidine and -5.6 kcal·mol⁻¹ for *N*-amidinopiperidine, see Table 4-1). Possibly, the experimental values cover additional affinity contributions not considered in our simulations.

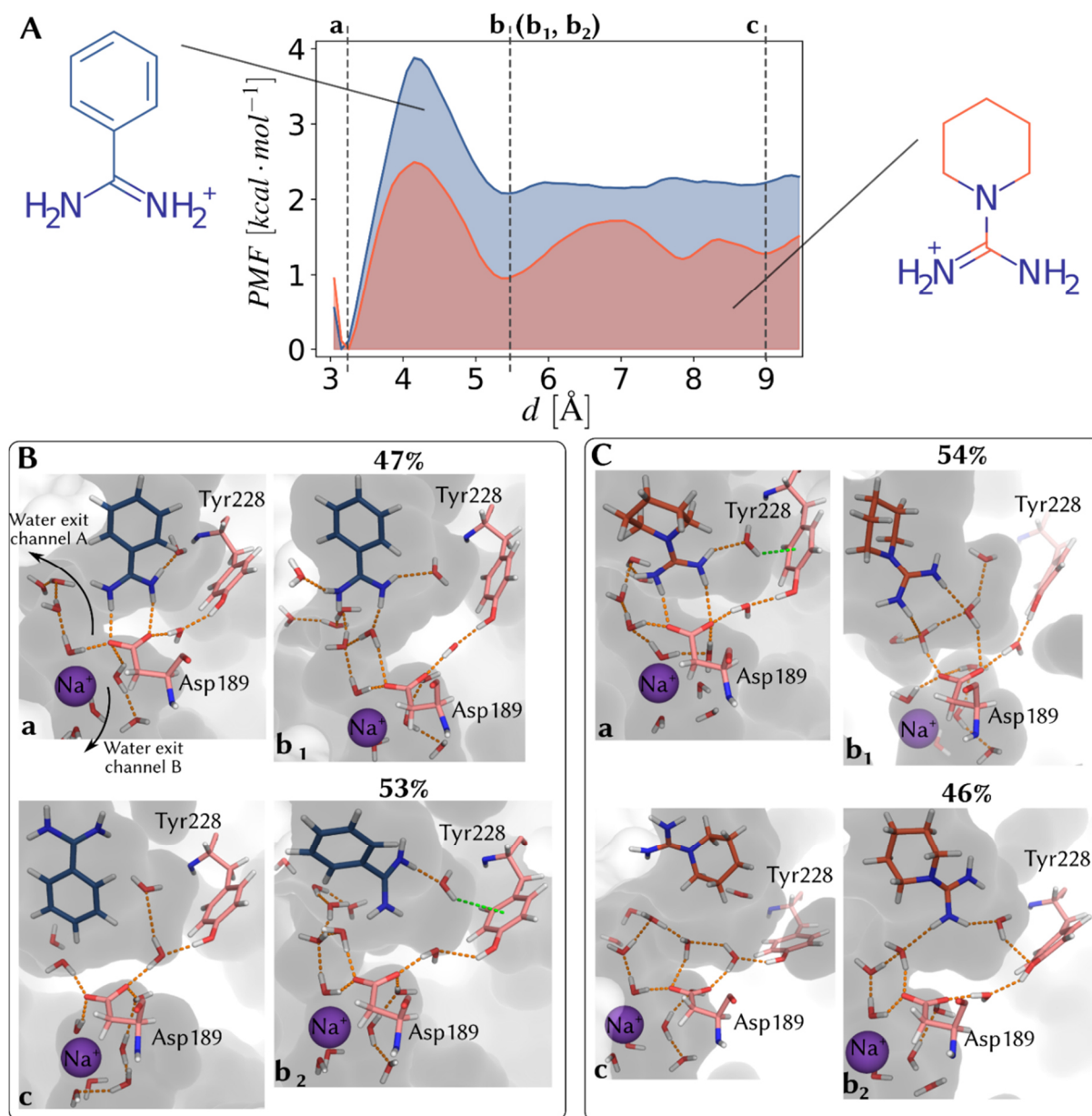


Figure 4-8: Overview of the dissociation mechanism of the protein-ligand complexes of benzamidine-thrombin and N-amidinopiperidine-thrombin. **A:** PMF along the reaction coordinate d for both thrombin complexes (see Figure 4-5 for the definition of the reaction coordinate); **B, C:** Representative snapshots from the MD simulation at key steps **a**, **b** and **c** for benzamidine (**B**) and N-amidinopiperidine (**C**).

4.4.6 Desolvation Time-scale during Ligand Dissociation

Generally, upon dissociation of the ligand molecule out of the protein binding pocket, the ligand molecule as well as the amino acids in the binding pocket have to change their interaction patterns. This change involves the perturbation of ligand-protein interactions along with modulations of the interactions with water molecules accommodating in the binding pocket. The time-scale at which water molecules undergo these modulations are investigated in the following paragraph. Mean residence time values are computed along the dissociation path of the ligand from the binding site using the trajectories from the US simulations. In the following, we will assume that the τ_l time constant dominates the overall kinetics of the water molecules in the binding pocket such that $\tau_{overall} \cong \tau_1$. This assumption is further supported by the observed huge difference (hundred to thousand fold) between τ_l and τ_2 for the water molecules in the binding pocket of the *apo* protein (see Table 4-3). Therefore, we will only investigate the τ_l time constant and not consider any contributions from τ_2 in our analysis of the time-scale of the water molecules upon ligand dissociation. In the following, the orientation time-scales of the water molecules are not being analyzed, as they are qualitatively identical to the translational MRT values (see Supporting Information).

4.4.7 Desolvation Time-Scale of Trypsin Complexes

In the initial state of the trypsin complexes (state **a**), the water molecules next to the carboxylate group of Asp189 hold completely differing MRT values between the two complexes (see Figure 4-9B). Most likely, the difference in water MRT is due to one more water molecule in the water reservoir in the case of the *N*-amidinopiperidine (c.f. Figure 4-6B and C). As already pointed out in the analysis of *NME-Asp-ACE* as well as in reference¹⁷⁶, water molecules adjacent to the charged side chain of aspartic acid show unfavorable water-water interactions, likely due to the fact that all O-H bond vectors point into the direction of the carboxylate group. This unfavorable state leads to low MRT values for the water molecules next to Asp189 in the case of the *N*-amidinopiperidine-trypsin complex. However, in the corresponding benzamidine complex one water molecule less is available to establish interactions to Asp189 leading to less unfavorable water-water interactions. Thus, the MRT of water molecules adjacent to Asp189 is higher in the case of the benzamidine complex compared with the *N*-amidinopiperidine complex. The water fluctuations in the first hydration shell of Asp189 are restricted only to water molecules in the water reservoir below Asp189. This is evidenced by the 500-fold

increased MRT for water molecules adjacent to the amidino moiety of *N*- amidinopiperidine compared to the corresponding moiety in benzamidine (see Figure 4-9A). These water molecules are shared between the first solvation layer of the amidino group of the ligand and the first solvation layer of the side chain of Asp189 and thus exclude water molecules below Asp189 (in the water reservoir).

In state **b** of the trypsin-complexes, the MRT of the water molecules next to the amidino moiety of *N*-amidinopiperidine is at its maximum value (1450 ps), whereas it is lower for the water molecules next to the benzamidine ligand (see Figure 4-9A, 500 ps). The latter MRTs reach their maximum at about 5.1 Å. The much higher MRT of water molecules next to the amidino moiety in the case of *N*-amidinopiperidine-trypsin complex is due to its additional water molecule, which is lacking in the corresponding benzamidine complex (c.f. Figure 4-6B and C).

In final state **c**, the residence time for the water molecules at amidino moiety and the ones at Asp189 have reached their reference values (see dotted lines in Figure 4-9A and B).

4.4.8 Desolvation Time-Scale of Thrombin Complexes

As already noted above, thrombin does not contain a water reservoir but water channels below Asp189 (see Figure 4-8B and C). These channels facilitate a constant flow of water molecules into and out of the binding pocket while the ligand is still bound. Due to this constant flow of water molecules, the MRT of water molecules next to the amidino moiety take the same value in state **a** of both thrombin complexes (Figure 4-9C). Moreover, the MRT of the water molecules next to Asp189 are also virtually identical in state **a** (Figure 4-9D).

In state **b**, the MRT of water molecules next to the amidino moiety in the benzamidine complex is 1050 ps and thus higher than the value of 600 ps computed for the corresponding *N*-amidinopiperidine complex. Similarly, water molecules next to Asp189 have higher MRTs in the case of the benzamidine complex (800 ps) compared to the *N*-amidinopiperidine (600 ps) complex. However, the observed standard deviations for the MRTs at state **b** are generally so high that no clear difference between the two complexes can be made. This is in line with the structural perspective, since both complexes seemed to be rather similar at this state (c.f. Figure 4-8B and C).

At state **c** of the thrombin complexes, the MRTs of the water molecules next to the ligand converge to the value found for the ligand in bulk solvent (4.1 ps, see dotted line in Figure 4-9C). The MRT values of water molecules next to Asp189 fluctuate between state **b** and state

c and do not fully reach the value of *apo* thrombin. This is most likely due to the large fluctuations and the multiple solvation processes observed in *apo* thrombin (see also Figure 4-4D). Also, it must be mentioned that even at $d = 10 \text{ \AA}$, the ligand is not fully unbound but all key interactions are already broken.

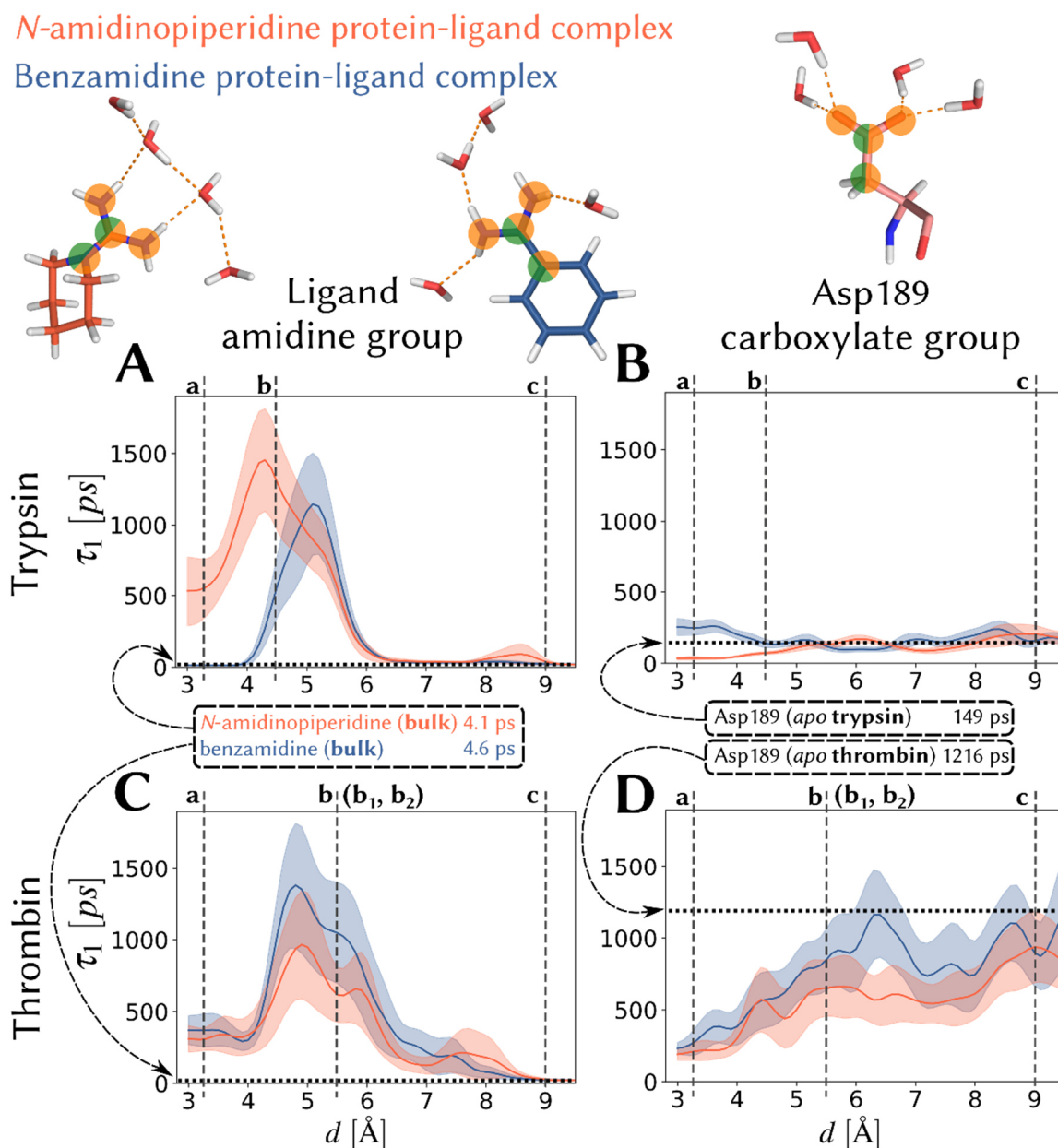


Figure 4-9: Overview of the slow MRT component τ_1 calculated for all US windows along the reaction coordinate using the LoCorA approach. The displayed water MRTs were computed using eq. (4-4) in the local solute coordinate system of the ligand amidino group (left column **A**, **C**) or the carboxylate group of Asp189 (right column **B**, **D**) calculated for protein-ligand complexes of trypsin and thrombin. The dotted lines indicate the MRT values for the pure ligands in bulk solvent (**A**, **C**) or Asp189 in apo trypsin (**B**) or apo thrombin (**D**).

4.5 Discussion

In the present contribution, we first investigated the spatial and temporal fluctuations of water molecules in the *apo* state of thrombin and trypsin. We then elaborated on the mechanism of the dissociation process of benzamidine and *N*-amidinopiperidine from the two proteins using US simulations. During this elaboration, we focused on the spatial and temporal fluctuation of water molecules upon ligand dissociation from the binding site along the predefined reaction coordinate.

4.5.1 Water Escape Mechanisms in the *apo* Binding Sites

We found that water molecules show very long MRTs (in the range of ns) next to the carboxylate group of Asp189 in the S1 pocket of thrombin, whereas around Asp189 in trypsin, the computed MRTs corresponds only to one tenth (only in the range of 100 ps) of the values in thrombin. This indicates that the two very similar proteins show very distinct solvation and desolvation mechanisms for Asp189, which is a key residue in substrate recognition. This difference likely contributes to the different selectivity profiles of the two proteins, as a ligand (or a substrate) has to compete with water molecules about the binding to the side chain of Asp189. In trypsin, the water molecules associated with Asp189 escape more frequently from the first hydration shell. Therefore, a ligand has a high probability to find a dewetted binding position at Asp189, accordingly it can bind more easily. Due to the principle of microscopic reversibility, these same considerations increase the barrier for the ligand dissociation from the binding site. The critical role of the water molecules in the binding mechanism is also reflected by the generally higher barriers on the FES for trypsin compared to thrombin.

In addition, we found only a single most probable value for the MRT in trypsin, whereas a bimodal Gaussian distribution was found for thrombin (Figure 4-5). This suggests that only a single water escape mechanism exists in trypsin. For thrombin, two major escape mechanisms are possible which reflect the fact that thrombin has multiple entry and exit channels to the S1 binding pocket. Through these channels, water molecules can exchange with bulk solvent, which was also noted elsewhere.¹⁷⁹

We argue that the occurrence of two different water escape mechanisms in thrombin has a functional role for the protein. We can only speculate about this functionality, but we believe that external factors, such as ligands that bind to a remote site or modulations of the ionic strength due to changes in the local salt concentration, can alter the preference for these two

water escape mechanisms. Thus, a shift in the probability of these mechanisms also shifts the protein's ability to recognize ligand molecules, as the water molecules associated with the first hydration shell of Asp189 leave more frequently into the bulk. Consequently, trypsin should experience less influence of external factors as only one single solvent escape mechanism can be observed.

4.5.2 Water Molecules and the Ligand Dissociation Mechanism

We found different mechanisms taking place during the dissociation of *N*-amidinopiperidine and benzamidine from the protein binding pockets of thrombin and trypsin. The difference between the mechanisms in the two proteins is mainly due a different water inventory below Asp189: In the case of trypsin, this inventory is called water reservoir and has a fixed number of water molecules as soon as a ligand molecule is accommodated in the protein binding pocket. In the case of thrombin, this inventory is called water channel and is proposed to have a varying number of water molecules, independent of the binding state of the protein.

Upon the dissociation of benzamidine from trypsin, one water molecule first intercalates between ligand and Asp189 in the S1 binding pocket. This is in contrast to the dissociation of *N*-amidinopiperidine from trypsin, where two water molecules intercalate between the ligand and Asp189. These observations confirm our previous investigations on the role of water molecules in the binding mechanism of trypsin complexes.⁵⁴ It must be noted that the PMF profile likely will increase at higher reaction coordinate values. However, for our considerations mainly concerning the water molecules in the binding pocket, the scanned range of reaction coordinates is sufficient.

In the present contribution, we found that *N*-amidinopiperidine assembles water molecules with long MRTs in the fully bound end-state with trypsin. This is in contrast to the corresponding end-state of the benzamidine-trypsin complex, as the water molecules in this complex exhibit shorter MRTs. Interestingly, the opposite distribution of MRTs was found for water molecules next to Asp189 in the two complexes of trypsin. These observations are explained with one additional water molecule in the *N*-amidinopiperidine-trypsin complex. The additional water molecule leads to unfavorable water-water interactions next to Asp189 and thus facilitates shorter MRTs of water molecules in the first hydration shell of Asp189. This seems only to affect water molecules in the water reservoir below Asp189 and not the water molecules next to the amidine group of the *N*-amidinopiperidine ligand. The additional water molecule in the complex of *N*-amidinopiperidine is seemingly recruited by the ligand in state **c** of the binding

reaction (see Figure 4-5C), as at this stage water molecules from the bulk water phase are in exchange with the protein binding pocket. The overall higher range of MRTs computed for the water molecules next to the amidine group in the *N*-amidinopiperidine-trypsin complex indicates a higher desolvation barrier as compared to the corresponding benzamidine-trypsin complex. The difference in barrier height may explain the lower binding affinity of *N*-amidinopiperidine towards trypsin (c.f. Table 4-1). It must be noted that one cannot directly deduce the barrier height from the time-constants, since the pre-exponential factor (for an Arrhenius-type analysis) is not known. For this purpose, one would have to carry out the same analysis at different temperatures and obtain the pre-exponential factor as well as the activation barrier from an Arrhenius plot.

Concerning the thrombin complexes of benzamidine and *N*-amidinopiperidine, we found a different solvation mechanism upon ligand dissociation than for trypsin. For both complexes, the MRTs of the ligand-associated water molecules are similar in the fully bound state, which is in contrast to the deviating MRTs observed for the two trypsin complexes in the fully bound state. This can be explained by the two water channels (see Figure 4-8B) present in thrombin but absent in trypsin. These exit channels enable the escape of water molecules from the binding site by a path that is not blocked by the ligand molecule. Thus, the MRTs of water molecules in thrombin do not depend on the ligand molecule that is bound to the binding pocket. However, in trypsin it clearly depends on the type of ligand molecule that is accommodated in the binding pocket, as there no water channels exist and the water molecules must enter through the same path as the ligand.

We computed lower MRTs for ligand-associated water molecules in the intermediate states of the *N*-amidinopiperidine-thrombin complex compared to the corresponding benzamidine-complex, although it is not completely clear why this difference occurs. These intermediate states involve bridging water molecules between the amidino group of the ligand and the carboxylate group of Asp189. We conclude that the low MRTs of water molecules adjacent to *N*-amidinopiperidine indicate a lower desolvation barrier of the intermediate states in the *N*-amidinopiperidine compared to the corresponding barrier in the benzamidine complex. The constant increase of the MRTs of water molecules at Asp189 in both complexes (see Figure 4-9D) indicates a constantly increasing (desolvation) barrier for water molecules and consequently lower the frequency of their escape from the binding pocket upon ligand dissociation.

A further, quite remarkable difference between the thrombin and trypsin complexes, is the fact

that thrombin exhibits a bimodal distribution with two geometries in state **b** (see Figure 4-8B, C and Figure S2 in the Supporting Information), whereas trypsin only shows a pathway with one single intermediate state (Figure 4-6B, C and Figure S1 in the Supporting Information). Thus, in thrombin the intermediate state **b** is stabilized entropically, whereas in trypsin no entropic contribution was observed.

4.6 Conclusion

In the first part of this contribution, we investigated the solvation mechanism of the binding pocket of *apo* thrombin and trypsin. We found that two fundamentally different solvation mechanisms are at play in these two proteins. These differences are due to the occurrence of water channels, which are present in thrombin, but are absent in trypsin. Trypsin on the other hand has a so-called water reservoir, which is a water inventory holding a fixed number of water molecules. Our mechanistic considerations and analysis of MRTs of ligand-associated water molecules led to the conclusion that the desolvation time-scale is dependent on the ligand and the (fixed) number of water molecules in the water reservoir in the case of trypsin, whereas it is quite independent from the ligand in the case of thrombin due to water channels.

Our investigation sheds light on the presently unpopular but physically reasonable idea that ligand binding mechanisms are not only driven by protein-ligand interactions, but also by solvation barriers of the protein as well as the ligand molecule. And even for such similar proteins as thrombin and trypsin, drastic differences are observed that are hard to record by experiment alone. Selectivity of drug molecules towards a specific target protein is important for the development of successful drug molecules. With our contribution, we highlight the concept of solvation barriers as an additional dimension in the development for selective drug molecules.

Our approach, *LoCorA*, is integrated into a software package that can be obtained from the GitHub page of the lead author of this contribution (<https://github.com/wutobias>).

4.7 Materials and Methods

In this section, we will outline the procedure that was used for conducting the various biased and unbiased MD simulations of this work. Also, we will give additional details for the calculation of the TCF from the *LoCorA* approach. In the first part, we will explain the structure preparation procedure, followed by the procedure applied for the generation of structure and parameter files. This is followed by a description of our protocol used for carrying out unbiased as well as biased (i.e. umbrella sampling) MD simulations. For the whole procedure, we used the Amber16 program package.¹¹⁷ From the Amber package, we used *pmemd* for all minimization runs and the GPU implementation *pmemd.cuda*¹²⁰⁻¹²² for all MD runs.

During our study, we derived all TCFs and MRTs from an NVT ensemble of the system. This is quite uncommon in the context of residence time calculations or derivation of kinetic properties in general. Usually one would use an NVE ensemble and not have a thermostat actively changing the velocity distribution of the system. However, in our case it was necessary, since we had to derive the PMF and MRT from the same set of trajectories. In order to calculate the PMF, it was necessary to have a molecular ensemble with a defined thermodynamic temperature, which made the use of a thermostat inevitable to our approach.

4.7.1 Structure Preparation

The structures were obtained from the PDB website. For trypsin, we used 5MOP, 5MOQ and 5MNP as input for the MD simulations of the *apo* protein, the benzamidine complex and the *N*-amidinopiperidine complex, respectively. For thrombin, we used 2UUF, 4UEH and 4UE7 for the MD simulations of the *apo* protein, the benzamidine complex and the *N*-amidinopiperidine complex, respectively. All structures were prepared (building missing atoms, assigning protonation states) using MOE.¹¹³ Also, we used MOE to assign am1-bcc charges^{180,181} for the ligand molecules. The protein atoms were treated with the *FF14SB* amber force field⁵⁷ and the ligand is treated using the *GAFF* force field.¹¹⁶ Missing parameters for the ligand molecule were assigned using *parmchk2* from the AmberTools17 package.¹¹⁷ All parameters were combined using *tLEaP* and all atoms were embedded into a truncated octahedron simulation box filled with water molecules. Throughout all simulations, we used the TIP4P-Ew water model.^{118,182} In order to ensure net charge neutrality, we used the *addions2* utility of *tLEaP* and added one sodium ion to the *apo* simulation box of thrombin, one sodium ion to the benzamidine-thrombin complex and one chlorine ion to the *N*-amidinopiperidine-

thrombin complex to achieve overall charge neutrality. In the case of trypsin, we added eight chlorine ions to the *apo* protein, nine chlorine ions to the benzamidine-trypsin complex and nine chlorine ions to the *N*-amidinopiperidine-trypsin complex. For the simulation boxes of the ligand molecules, we placed one ligand molecule in the simulation box and added one chlorine counter ion. The ligand molecules were treated as protonated in all simulations, according to our calculations using the protonate3D utility of MOE. All simulation boxes contained 13348 water molecules for the protein-ligand complexes as well as the *apo* proteins, and 2300 water molecules for the ligand molecules in solution.

4.7.2 Unbiased MD simulations

During all simulation runs, we applied periodic boundary conditions using the periodic-mesh Ewald technique as implemented in Amber16 *pmemd.cuda* together with a 9 Å real-space distance cutoff. Furthermore, all bonds involving hydrogen atoms were constrained using the SHAKE¹¹⁹ algorithm. All runs were carried out in triplicates and each run was started from a different (assigned randomly) random seed for the velocities.

For the *apo* structures as well as the ligand molecules in solution, we carried out classical (unbiased) MD simulations. We initially performed an energy minimization of the system while keeping the solute heavy atoms fixed to their crystallographic positions using a harmonic spring potential with a force constant of 25 kcal·mol⁻¹·Å⁻². This energy minimization was carried out using 250 steps of steepest descent and 250 steps of conjugate gradient minimization. In an additional second energy minimization, the force constant was reduced to 2 kcal·mol⁻¹·Å⁻², all other parameters were kept similar to the first minimization. Then, the system was heated to 300 K within 25 ps using an integration time step of 1 fs, while still keeping the atoms fixed with a force constant of 25 kcal·mol⁻¹·Å⁻². The system was kept at this temperature for all following runs using a Langevin dynamics thermostat with a collision frequency of $\gamma = 2 \text{ ps}^{-1}$. The integration time step was increased to 2 fs and the restraints were switched off gradually, while equilibrating the system within 100 ps to a target pressure of 1 bar using the Berendsen barostat¹¹⁹. In a final step, the system was equilibrated under NVT conditions for a duration of 1 ns without any restraints.

Final production MD trajectories were carried out for 200 ns for the *apo* proteins as well as the ligand molecules in solution. The coordinates of all atoms were saved to disk every 0.5 ps.

4.7.3 Umbrella Sampling MD simulations

For the biased sampling of configurations along the reaction coordinate (see Figure 4-5 for the definition of the reaction coordinate), the latter was divided into 71 equally spaced windows with a width of 0.1 Å ranging from 3.0 to 10.0 Å. Each window was sampled in triplicates and each replicate was started from a randomly chosen snapshot extracted from a 1 ns MD run of the fully bound protein-ligand complex. These short 1 ns MD runs for every protein-ligand complex were generated by the protocol for unbiased MD simulations as introduced above. Furthermore, each run was started from a different (assigned randomly) random seed for the velocities.

The starting structure for each window in the system was optimized with 250 steps of steepest descent energy minimization followed by 250 steps of conjugate gradient energy minimization. After that, the system was heated to 300 K within 25 ps using an integration time step of 1 fs. At this temperature, the system was equilibrated to a target pressure of 1 bar within 50 ps under NPT conditions using the Berendsen thermostat, followed by a 50 ps NVT equilibration run. The equilibrated protein-ligand complexes were pulled gently to the target reaction coordinate value within 1 ns using the steered MD^{183,184} functionality in *pmemd.cuda*.

At the final reaction coordinate value, the system was again minimized, heated and equilibrated as carried out right before the steered MD step. However, this time the system was restrained using a harmonic potential centered at the target reaction coordinate value with a force constant of 5 kcal·mol⁻¹·Å⁻².

Final production MD runs were carried out for 10 ns for each of the three replicates in each window. Similar to the simulations of the *apo* proteins and the unbound ligand molecules, the coordinates of all atoms were saved to disk every 0.5 ps.

4.7.4 PMF Analysis

The PMF along the reaction coordinate was obtained by means of WHAM^{177,178} as implemented in the program *wham*.¹⁸⁵

4.7.5 Trajectory processing with *LoCorA*

The trajectories from the biased and unbiased simulations were post-processed using *LoCorA*, which is an in-house developed program available to the scientific community at <https://github.com/wutobias>. The local coordinate system of the amino acid side chains and the ligand portions were defined as outlined in Figure 4-2 of the Theoretical Background section.

In order to obtain an error estimate for the MRT parameters in eqs. (4-4) and (4-5), we performed block bootstrapping on the time-series of the survival function $B_j^{(S)}(t)$ (see eq. (4-1)). The same bootstrapping blocks that were used for $B_j^{(S)}(t')$, were also used in the bootstrapping of the axis-vectors for the water coordinate system $\vec{W}_{j,t'}^{(S,x)}$, $\vec{W}_{j,t'}^{(S,y)}$, $\vec{W}_{j,t'}^{(S,z)}$. In order to obtain a globally optimal solution of the parameters in eqs. (4-4) and (4-5), we applied a short basin-hopping¹³⁴ optimization run in conjunction with the L-BFGS-B^{186,187} local minimizer. The basin-hopping optimization run evolved for a maximum of 30 steps and was allowed to stop if no improved solution was found after 10 steps. The parameter optimization was carried out using the SciPy package¹⁸⁸ for scientific computing in Python.

All TCF were calculated from 1000-fold block bootstrapping and each block had a length of 6 ns. For the final analysis, we discarded all bootstrapped solutions to eqs. (4-4) and (4-5) that had an R^2 of less than 0.95 with the computed TCF from the MD simulation.

Input files for *LoCorA* will be provided as part of the Supporting Information upon publication of this manuscript.

4.8 Supporting Material

4.8.1 Temporal Properties of Bulk Water Molecules

Initially, we calculated the MRT of water molecules around an individual water molecule in the pure bulk water phase. The calculation of the MRT in bulk water, i.e. the MRT of a water molecule in the local coordinate system of another water molecule, is especially insightful. In the bulk state, water molecules usually exhibit faster relaxation behavior than in the environment of a solute molecule. Therefore, it serves as a reference state and also allows for the comparison with experimentally determined translational and orientational time constants of water molecules from NMR or fsIR (femto-second infrared).

Throughout this study, we used the four-site water model TIP4P-Ew.^{118,182} An overview about the calculated MRTs calculated from a 1 ns MD trajectory (0.01 ps per frame in the MD trajectory) is summarized in

Table S4-4. We found that the translational MRTs for the TIP4P-Ew water model (5.6 ± 0.6 ps) are in good agreement with reference values from the SPC-E water model (6.0 ps).¹⁷³ Note that for this comparison we used a TCF with a tolerance time ($t^*=2$ ps) according to the IMM approach (see the Theoretical Background section). The orientational MRTs are differing from the ones calculated by the SPC-E water model and also deviate from the ones found by experiment. This is most likely due to the different definitions of this quantity: In our work, we do not make any assumptions about the state of the water molecule at the beginning of a time series (i.e. at $t' = 0$). In other studies,^{189,190} the orientation times are calculated for water molecules that just have lost a hydrogen bond to another water molecule and interimly tumble in space before they establish a new hydrogen bond to an adjacent water molecule. Due to this concept, it is also termed reorientation time and has led to the development of the jump-model.¹⁸⁹ In our work, the orientational MRTs are shorter (1.2 ± 0.2 ps) than the above defined reorientation gap calculated for the SPC-E water model (2.5 ps)¹⁸⁹ or the experiment suggesting 2-7.5 ps.^{189,190} Most likely, this is due to the fact that water molecules that lost a hydrogen bond to another water molecule are still weakly bound to previously contacted water molecule. Effectively they are still under the mutual influence of the dipoles of each other and thereby experience enhanced orientation time constants compared to our model of water orientation, which includes water molecules in all possible states. Nonetheless, we want to emphasize that our approach suggests values falling close to those references. Clearly, the interpretation of reorientation times based on swapped hydrogen bonding seems amenable, but probably will be

misleading in the case of hydrophobic environments, where no generalized short-ranged geometric preferences between hydrogen-bond acceptor and donor (between solvent and solute) is given as in the bulk water phase.

Equations (4) and (5) reflect the bimodal behavior of the water translation and orientation kinetics in the first hydration shell. This characteristic behavior is further reflected in a plot of the raw TCF data (Figure S4-10): At very short lag times, the transition from fast to slow relaxation can be seen as an offset in the TCF (at about 0.1 ps). The initial fast decay corresponds to water molecules that do not stabilize sufficiently well in the first hydration shell and leave immediately. These water molecules can be involved in fast recrossing processes at the boundary between first and second hydration shell. Since the orientational TCF is conditioned on the translational TCF, the orientational TCF cannot decay any faster than the translational TCF. In our considerations, the decay of the orientational TCF is about twice as fast as the corresponding decay of the translational TCF. Most likely, this is due to the suggested jump-mechanism¹⁸⁹ of water molecules undergoing the breaking and making of hydrogen bonds accompanied by large angular jumps.

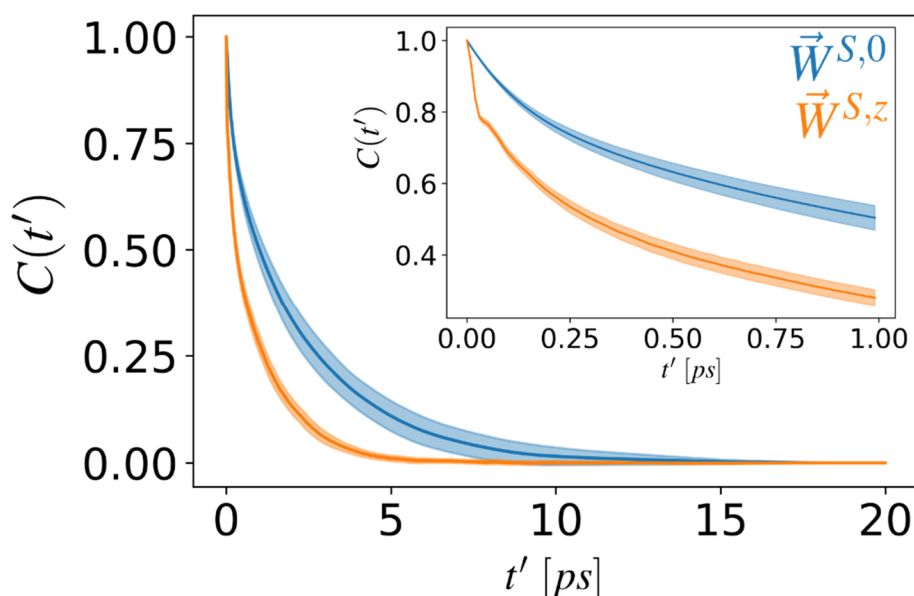


Figure S4-10: Time-correlation function for the translation (blue) and orientation with respect to the z-axis (orange) of water molecules in bulk water. The transparent area indicates ± 1 standard deviation.

Table S4-4: Residence times and weighting factors calculated for bulk water.^{a)}

Water Model	Component ^{b)}	τ_{int} [ps] ^{c)}	τ [ps] ^{c)}	τ_1 [ps] ^{e)}	τ_2 [ps] ^{e)}	w [ps] ^{e)}
<i>TIP4P-Ew</i> ($t^*=0$ ps)	$\langle \vec{W}^{(s,0)} \rangle_{N_f}$	2.0 ± 0.4	2.4 ± 0.5	2.7 ± 0.6	0.2 ± 0.1	0.4 ± 0.2
	$\langle \vec{W}^{(s,z)} \rangle_{N_f}$	0.8 ± 0.1	1.1 ± 0.1	1.3 ± 0.1	0.1 ± 0.1	0.5 ± 0.1
	$\langle \vec{W}^{(s,y)} \rangle_{N_f}$	0.9 ± 0.2	1.2 ± 0.2	1.4 ± 0.3	0.1 ± 0.1	0.5 ± 0.1
	$\langle \vec{W}^{(s,x)} \rangle_{N_f}$	0.9 ± 0.1	1.2 ± 0.1	1.4 ± 0.1	0.1 ± 0.1	0.5 ± 0.1
<i>TIP4P-Ew</i> ($t^*=2$ ps) ^{f)}	$\langle \vec{W}^{(s,0)} \rangle_{N_f}$	4.6 ± 0.4	5.6 ± 0.6	5.8 ± 0.7	0.1 ± 0.1	0.4 ± 0.3
	$\langle \vec{W}^{(s,z)} \rangle_{N_f}$	1.0 ± 0.1	1.5 ± 0.2	1.6 ± 0.2	0.1 ± 0.1	0.5 ± 0.1
	$\langle \vec{W}^{(s,y)} \rangle_{N_f}$	1.2 ± 0.2	1.6 ± 0.3	1.8 ± 0.3	0.1 ± 0.1	0.5 ± 0.1
	$\langle \vec{W}^{(s,x)} \rangle_{N_f}$	1.2 ± 0.1	1.6 ± 0.2	1.8 ± 0.2	0.1 ± 0.1	0.5 ± 0.2
<i>SPC-E</i> ($t^*=2$ ps)	$\langle \vec{W}^{(s,0)} \rangle_{N_f}$	n.a. ^{k)}	6.0 ^{g)}	n.a.	n.a.	n.a.
<i>SPC-E</i> ($t^*=0$ ps)	$\langle \vec{W}^{(s,x)} \rangle_{N_f}$	1.7 ^{h),j)}	2.5 ^{h),j)}	n.a.	n.a.	n.a.
<i>Experiment</i>	$\langle \vec{W}^{(s,x)} \rangle_{N_f}$	n.a.	$2-7.5$ ^{i),j)}	n.a.	n.a.	n.a.

a) Mean values and standard deviations are obtained from 1000 block bootstrapping attempts based on a 1 ns trajectory. Each block was 50 ps in length and the time step between two MD frames in the block was 0.01 ps.

b) Reference coordinate system for components is based on the local coordinate system of another (central) water molecule with similar coordinate system definition as for all other water molecules (see Figure 2 in the main text).

c) Calculated from a full integral over the TCF.

d) Calculated from fitting a single exponential function to the TCF.

e) Calculated from eqs. (4) and (5).

f) Calculated with a transient recrossing time of $t^*=2$ ps.

g) See Ref. [173]

h) See Ref. [189]

i) See Refs. [189,190]

j) Calculated from the reorientation behavior of water molecules in water-water hydrogen bonding.

n.a.: a value for this quantity is not available

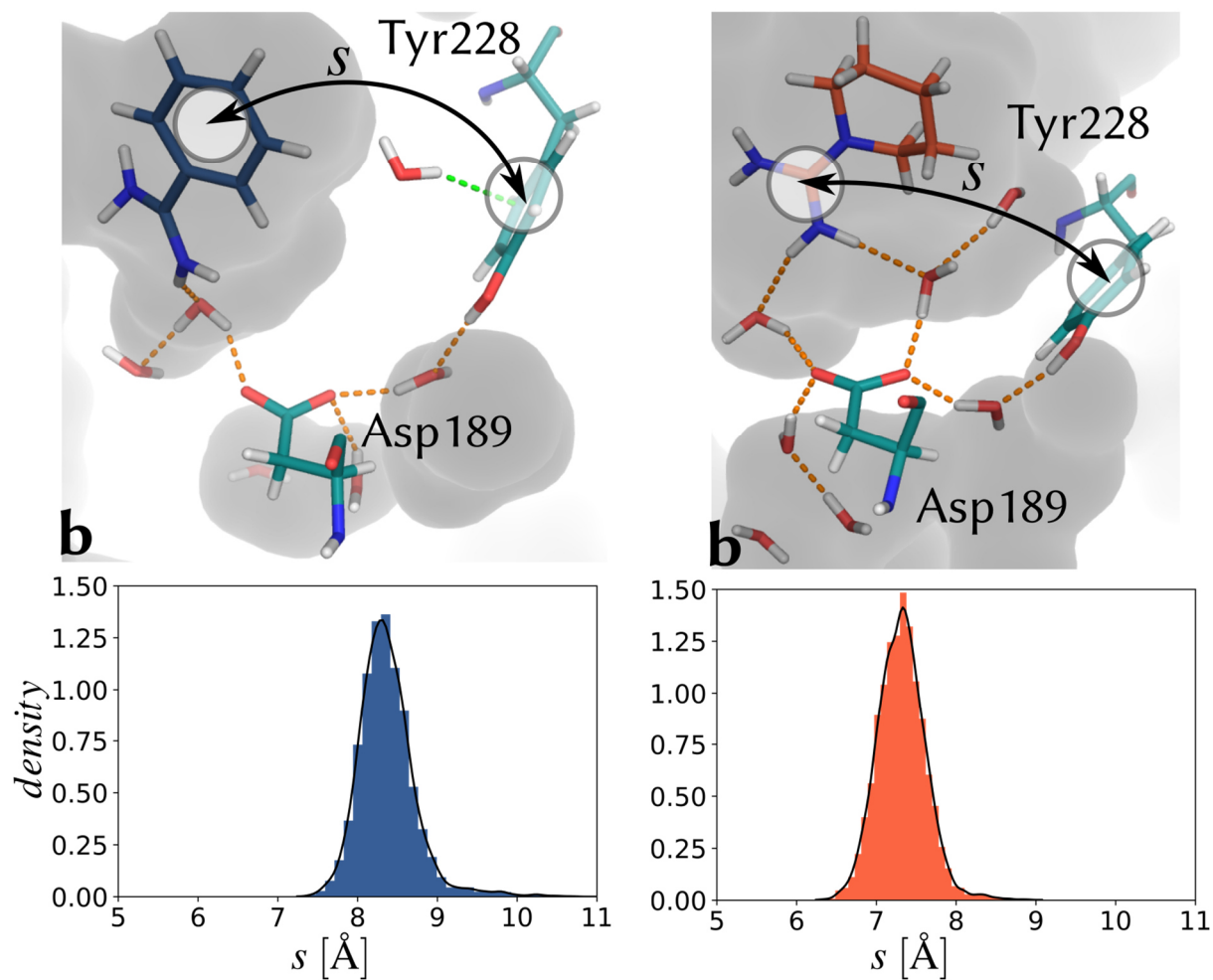
Distance amidine and Tyr228 sidechain
at **b** ($d = 4.5 \text{ \AA}$) in trypsin-ligand complexes

Figure S4-11: Distribution of the distance between the amidine group of the ligand and the aromatic side chain portion of tyr228 in trypsin at reaction coordinate value of $d = 4.5 \text{ \AA}$ (corresponds to state **b** in the main text).

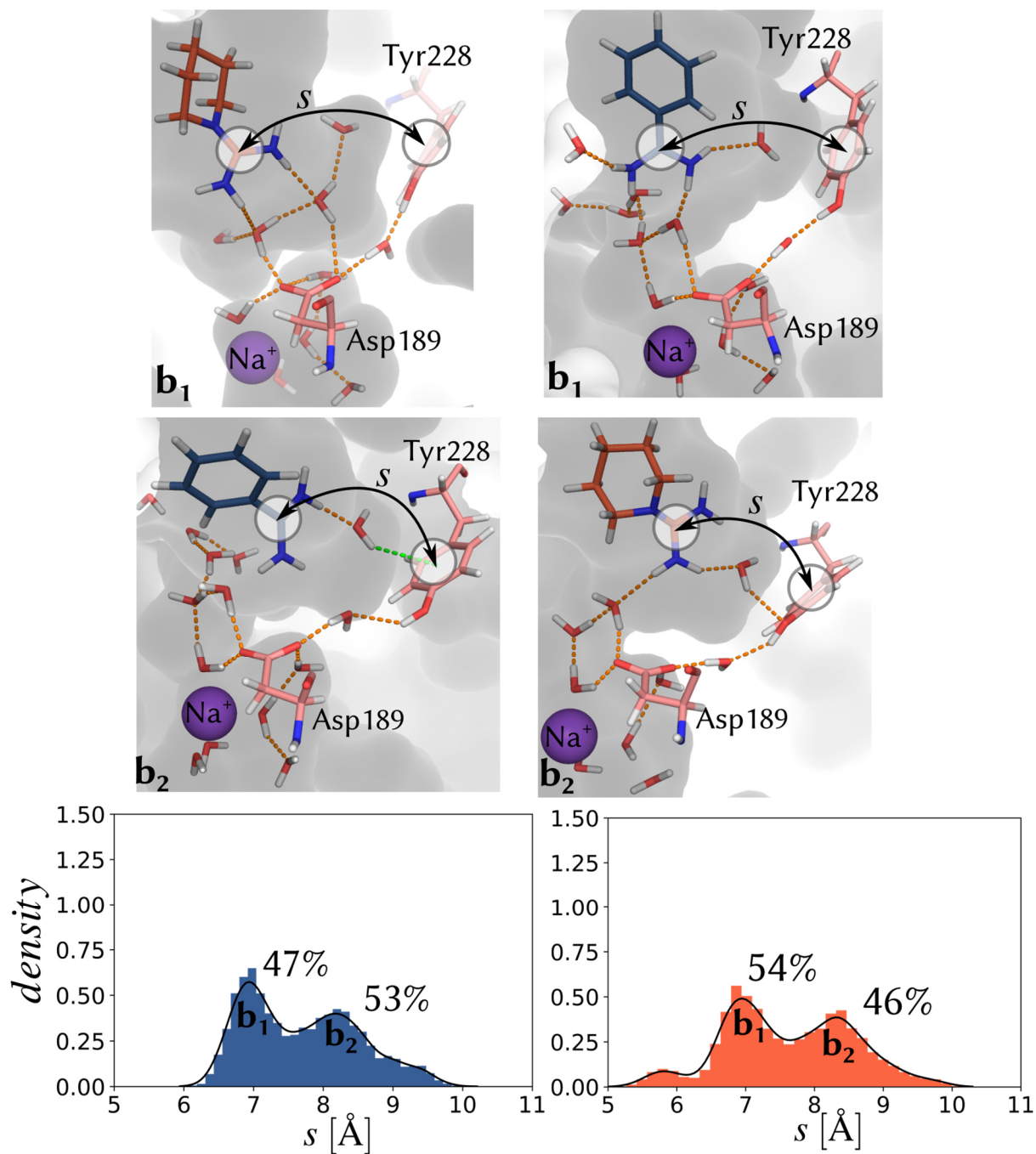
Distance between amidine and Tyr228 sidechain at **b** ($d = 5.5 \text{ \AA}$) thrombin-ligand complexes

Figure S4-12: Distribution of the distance between the amidine group of the ligand and the aromatic side chain portion of tyr228 in thrombin at reaction coordinate value of $d = 4.5 \text{ \AA}$ (corresponds to state **b** in the main text).

N-amidinopiperidine protein-ligand complex

Benzamidine protein-ligand complex

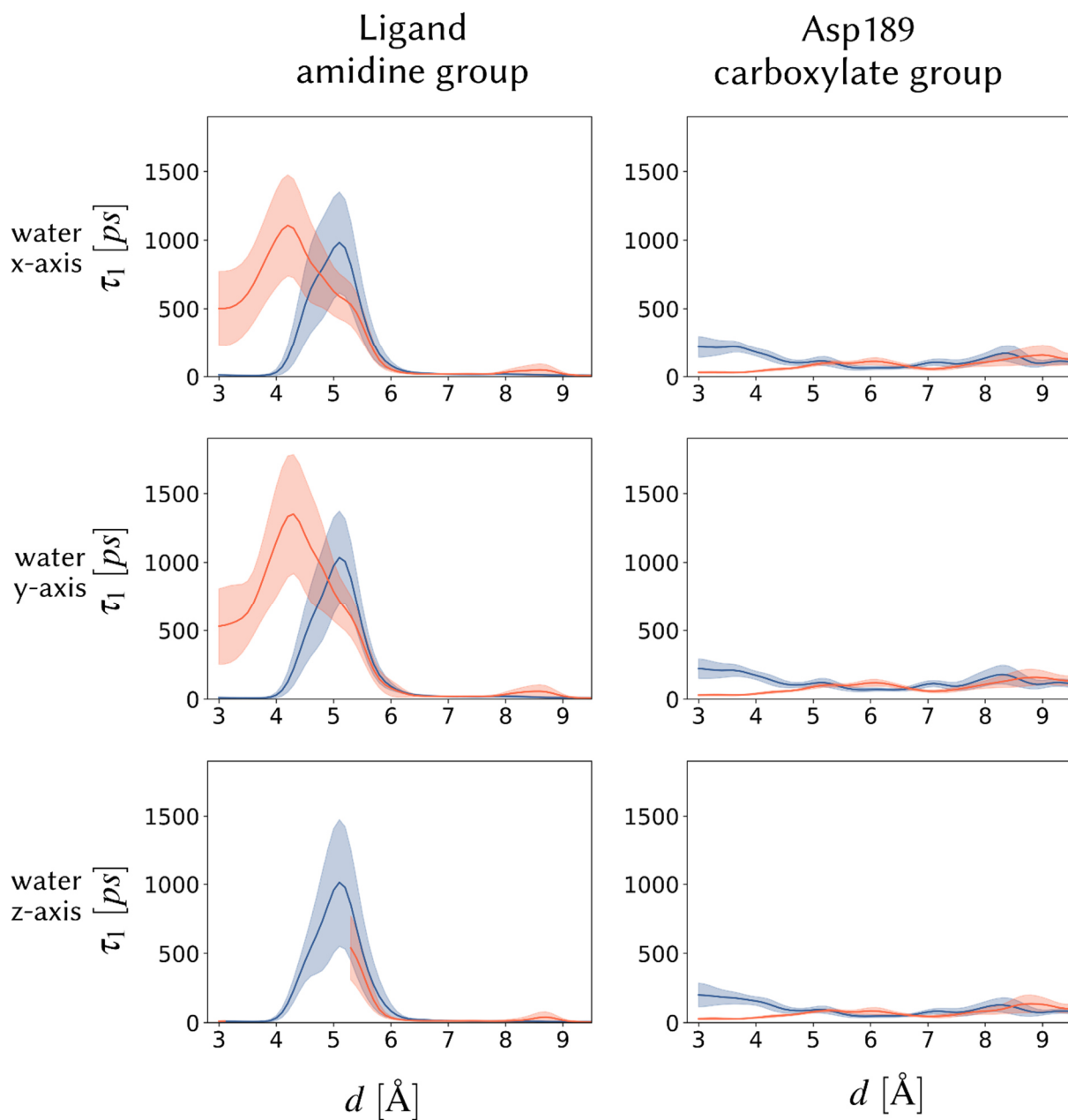
Trypsin

Figure S4-13: Orientation time-constants of water molecules adjacent to the amidino group of the ligand (left column) and the Asp189 side chain (left column) for stages along the dissociation path in trypsin.

N-amidinopiperidine protein-ligand complex

Benzamidine protein-ligand complex

Thrombin

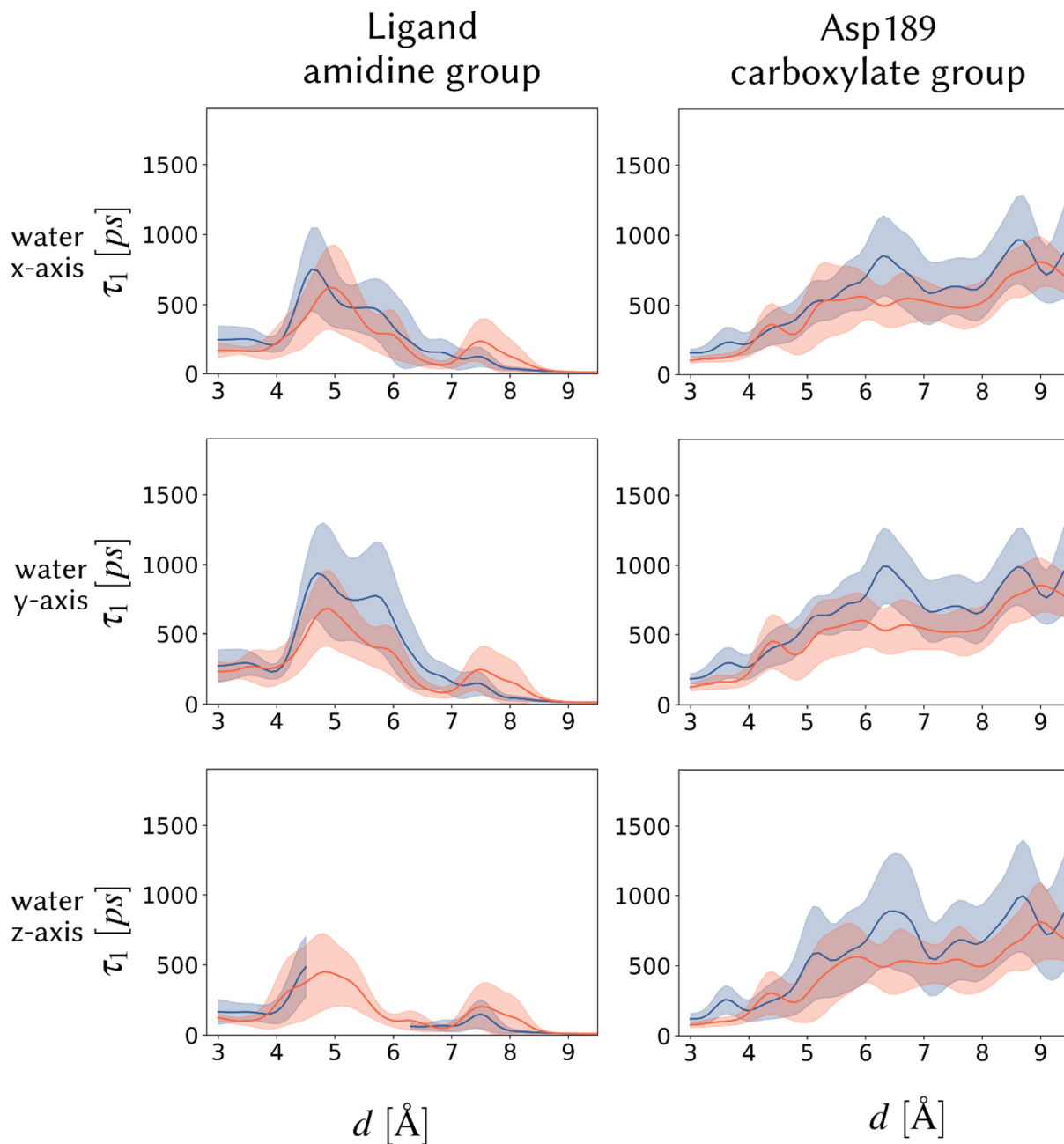


Figure S4-14: Orientation time-constants of water molecules adjacent to the amidino group of the ligand (left column) and the Asp189 side chain (left column) for stages along the dissociation path in trypsin.

5 Additional Studies

Remark

In this section, a set of additional studies is presented that were conducted and published during the time of my doctoral studies at Marburg University. In almost all studies, molecular interactions involving water molecules are in the focus of the investigation. All of these studies involved collaborative work with experimental research such as protein crystallography or ITC. For almost all of these studies (see introductory remarks for author contribution statement), I designed and conducted the computational modelling work. Due to the limited space in this doctoral dissertation, only the abstract and an author contribution statement are listed for each individual study.

Strategies for Late-stage Optimization: Profiling Thermodynamics by Preorganization and Salt Bridge Shielding

Sandner, A; Hüfner-Wulsdorf, T.; Heine, A.; Steinmetzer, T.; Klebe, G, *in revision*.

Introductory Remark

In this study, I carried out molecular dynamics simulations and according to these, suggested preorganization mechanisms that explained the relationship between the thermodynamic profile and the observed structure. Furthermore, I supported the interpretation of the experimental findings.

Abstract

Structural fixation of a ligand in its bioactive conformation may, due to entropic reasons, improve affinity. We present a congeneric series of thrombin ligands with a variety of functional groups triggering preorganization prior to binding. Fixation in solution and complex formation have been characterized by crystallography, ITC and MD simulations. First, we show why these preorganizing modifications do not affect the overall binding mode and how key interactions are preserved. Next, we demonstrate how preorganization thermodynamics are largely dominated by enthalpy, rather than entropy due to the significant population of low-energy conformations. Furthermore, a salt bridge is shielded by actively reducing its surface exposure and thus, leading to an enhanced enthalpic binding profile. Our results suggest that the consideration of the ligand solution ensemble by molecular dynamics simulation is necessary to predict preorganizing modifications that enhance the binding behavior of already promising binders.

Paradoxically, Most Flexible Ligand Binds Most Entropy-Favored: Intriguing Impact of Ligand Flexibility and Solvation on Drug-Kinase Binding.

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. *J. Med. Chem.* **2018**, *61*, 5922-5933.

Introductory Remark

In this study, I carried out molecular dynamics simulations in combination with solvation and conformation entropy calculations. Furthermore, I validated my calculations with data obtained from NMR experiments (experiments carried out by H.R.A. Jonker), particularly to study the ligand properties in aqueous solution prior to protein binding.

Abstract

Biophysical parameters can accelerate drug development, e.g. rigid ligands may reduce entropic penalty and improve binding affinity. We studied systematically the impact of ligand rigidification on thermodynamics using a series of fasudil derivatives inhibiting protein kinase A by crystallography, isothermal titration calorimetry, nuclear magnetic resonance and molecular dynamics simulations. The ligands varied in their internal degrees of freedom but conserve the number of heteroatoms. Counterintuitively, the most flexible ligand displays the entropically most favored binding. As experiment shows, this cannot be explained by higher residual flexibility of ligand, protein or formed complex nor by a deviating or increased release of water molecules upon complex formation. NMR and crystal structures show no differences in flexibility and water release although strong ligand-induced adaptations are observed. Instead, the flexible ligand entraps more efficiently water molecules in solution prior to protein binding and by releasing these waters, the favored entropic binding is observed.

On the Implication of Water on Fragment-to-Ligand Growth in Kinase Binding Thermodynamics

Wienen-Schmidt, B.*; Wulsdorf, T.*; Jonker, H. R.A.; Saxena, K.; Kudlinzki, D.; Linhard, V.; Sreeramulu, S.; Heine, A.; Schwalbe, H.; Klebe, G. *ChemMedChem* **2018**, *13*, 1988-1996.

*these authors contributed equally.

Introductory Remark

In this study, I carried out molecular dynamics simulations and solvation thermodynamics calculations. I supported the interpretation of experimental findings and suggested possible SARs according to my calculations.

Abstract

A ligand-binding study is presented focusing on thermodynamics of fragment expansion. The binding of four compounds with increasing molecular weight to protein kinase (PKA) was analyzed. The ligands display affinities between low-micromolar to nanomolar potency despite their low-molecular weight. Binding free energies were measured by isothermal titration calorimetry (ITC), revealing a trend towards more entropic and less enthalpic binding with increase in molecular weight. All protein-ligand complexes were analyzed by crystallography and solution NMR spectroscopy. Crystal structures and solution NMR data are highly consistent and no major differences in complex dynamics across the series are observed that would explain the differences in the thermodynamic profiles. Instead, molecular dynamics simulations reveal that the thermodynamic trends result either from differences in the solvation patterns of the conformationally more flexible ligands in aqueous solution prior to protein binding or local shifts of the water structure in the ligand-bound state. Our data thus provide evidence that changes in the solvation pattern constitute an important parameter for the understanding of thermodynamic data in protein-ligand complex formation.

Price for Opening the Transient Specificity Pocket in Human Aldose Reductase upon Ligand Binding: Structural, Thermodynamic, Kinetic, and Computational Analysis

Rechlin, C.; Scheer, F.; Terwesten, F.; Wulsdorf, T.; Pol, E.; Fridh, V.; Toth, P.; Diederich, W.E.; Heine, A.; Klebe, G. *ACS Chemical Biology* **2017**, *12*, 1397-1415.

Introductory Remark

For this study, I carried out molecular dynamics simulations and hydration site analysis. Based on the results of my calculations, I supported the formulation of a hypothetical induced fit binding mechanism.

Abstract

Insights into the thermodynamic and kinetic signature of the transient opening of a protein-binding pocket are presented resulting from accommodation of suitable substituents attached to a given parent ligand scaffold. As target, we selected human aldose reductase, an enzyme involved in development of late-stage diabetic complications. To recognize a large scope of substrate molecules, this reductase opens a transient specificity pocket. The pocket-opening step was studied by X-ray crystallography, microcalorimetry and surface plasmon resonance using a narrow series of 2-carbamoyl-phenoxy-acetic acid derivatives. Molecular dynamics simulations suggest that pocket opening only occurs once an appropriate substituent is attached to the parent scaffold. Transient pocket opening of the uncomplexed protein is hardly recorded. Hydration-site analysis suggests that up to five water molecules penetrating into the opened pocket cannot stabilize this state. Sole substitution with a benzyl group stabilizes the opened state and the energetic barrier for opening is estimated to about 5 kJ/mol. Additional decoration of the pocket-opening benzyl substituent with a nitro group results in a huge enthalpy-driven potency increase, whereas an isosteric carboxylic-acid group reduces potency 1000-fold and binding occurs without pocket opening. We suggest a ligand induced-fit mechanism for the pocket-opening step, which however, does not represent the rate-determining step in binding kinetics.

A False-Positive Screening Hit in Fragment-Based Lead Discovery: Watch out for the Red Herring

Cramer, J.; Schiebel, J.; Wulsdorf, T.; Grohe, K.; Najbauer, E.E.; Ehrmann, F.R.; Radeva, N.; Zitzer, N.; Linne, U.; Linser, R.; Heine, A.; Klebe, G. *Angew. Chem. Int. Ed.* **2017**, *56*, 1908–1913.

Introductory Remark

For this study, I calculated NICS values, electrophilicity indices and partial charges on DFT level using quantum chemistry methods. With my findings, I supported experimental findings about the reactivity of the investigated fragment molecule.

Abstract

With the rising popularity of fragment-based approaches in drug development, more and more attention has to be devoted to the detection of false-positive screening results. In particular, the small size and low affinity of fragments drives screening techniques to their limit. The pursuit of a false-positive hit can cause significant loss of time and resources. Here, we present an instructive and intriguing example about the origin of misleading assay results for a fragment that emerged as most potent binder for the aspartic protease endothiapepsin (EP) across multiple screening assays. This molecule shows its biological effect mainly after conversion to another entity through a reaction cascade that involves major rearrangements of its heterocyclic scaffold. The formed ligand binds EP through an induced-fit mechanism involving remarkable electrostatic interactions. Structural information in the initial screening proved to be crucial for the identification of this false-positive hit.

Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes.

Schiebel, J.; Gaspari, R.; Wulsdorf, T.; Ngo, K.; Sohn, C.; Schrader, Tobias E.; Cavalli, A.; Ostermann, A.; Heine, A.; Klebe, G. *Nature Communications* **2018**, *9*, 1-30.

Introductory Remark

In this study, I analyzed residence times and orientation times of water molecules in the *apo* binding pocket of trypsin. Furthermore, I compared the computed distribution of water orientations with the ones observed during neutron scattering experiments.

Abstract

Hydrogen bonds are key interactions determining protein-ligand binding affinity and therefore fundamental to any biological process. Unfortunately, explicit structural information about hydrogen positions and thus H-bonds in protein-ligand complexes is extremely rare and similarly the important role of water during binding remains poorly understood. Here, we report on neutron structures of trypsin determined at very high resolutions ≤ 1.5 Å in uncomplexed and inhibited state complemented by X-ray and thermodynamic data and computer simulations. Our structures show the precise geometry of H-bonds between protein and the inhibitors N-amidinopiperidine and benzamidine along with the dynamics of the residual solvation pattern. Prior to binding, the ligand-free binding pocket is occupied by water molecules characterized by a paucity of H-bonds and high mobility resulting in an imperfect hydration of the critical residue Asp189. This phenomenon likely constitutes a key factor fueling ligand binding via water displacement and helps improving our current view on water influencing protein–ligand recognition.

Diamandoid Amino Acid-Based Peptide Kinase A Inhibitor Analogues

Müller, J.; Kirschner, R. A.; Berndt, J. P.; Wulsdorf, T.; Metz, A.; Hrdina, R.; Schreiner, P. R.; Geyer, A.; Klebe, G. *ChemMedChem* **2019**, *14*, 663-672.

Introductory Remark

In this study, I carried out molecular dynamics simulations together with Dr. Alexander Metz. I analyzed the molecular dynamics trajectories using MMGBSA calculations in conjunction with various structural descriptors and suggested (together with Dr. Alexander Metz) a set of promising peptides for synthesis.

Abstract

The incorporation of diamandoid amino acids (DAAs) into peptide-like drugs is a general strategy to improve lipophilicity, membrane permeability, and metabolic stability of peptidomimetic pharmaceuticals. We designed and synthesized five novel peptidic DAA-containing kinase inhibitors of protein kinase A using a sophisticated molecular dynamics protocol and solid-phase peptide synthesis. By means of a thermophoresis binding assay, NMR, and crystal structure analysis, we determined the influence of the DAAs on the secondary structure and binding affinity in comparison to the native protein kinase inhibitor, which is purely composed of proteinogenic amino acids. Affinity and binding pose are largely conserved. One variant showed 6.5-fold potency improvement, most likely related to its increased side chain lipophilicity. A second variant exhibited slightly decreased affinity presumably due to loss of hydrogen-bond contacts to surrounding water molecules of the first solvation shell.

Impact of Surface Water Layers on Protein-Ligand Binding: How Well Are Experimental Data Reproduced by Molecular Dynamics Simulations in a Thermolysin Test Case?

Betz, M.; Wulsdorf, T.; Krimmer, S. G.; Klebe, G. J. *Chem. Inf. Model* **2016**, *56*, 223-233.

Introductory Remark

In this study, I supported the interpretation of the results and also provided python scripts that enabled the spatial integration over hydration sites.

Abstract

Drug binding involves changes of the local water structure around proteins including water rearrangements across surface-solvation layers around protein and ligand portions exposed to the newly formed complex surface. For a series of thermolysin-binding phosphoramidates, we discovered that variations of the partly exposed P2'-substituents modulate binding affinity up to 10 kJ·mol⁻¹ with even larger enthalpy/entropy partitioning of the binding signature. The observed profiles cannot be completely explained by desolvation effects. Instead, the quality and completeness of the surface water network wrapping around the formed complexes provide an explanation for the observed structure–activity relationship. We used molecular dynamics to compute surface water networks and predict solvation sites around the complexes. A fairly good correspondence with experimental difference electron densities in high-resolution crystal structures is achieved; in detail some problems with the potentials were discovered. Charge-assisted contacts to waters appeared as exaggerated by AMBER, and stabilizing contributions of water-to-methyl contacts were underestimated.

6 References

- (1) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists Bread and Butter. *J Med Chem* **2016**, *59* (9), 4385–4402. <https://doi.org/10.1021/acs.jmedchem.6b00153>.
- (2) Zon, L. I.; Peterson, R. T. In Vivo Drug Discovery in the Zebrafish. *Nat Rev Drug Discov* **2005**, *4* (1), 35–44. <https://doi.org/10.1038/nrd1606>.
- (3) Gibbs, J. B. Mechanism-Based Target Identification and Drug Discovery in Cancer Research. *Sci* **2000**, *287* (17), 1969–1973.
- (4) Sams-dodd, F. Target-Based Drug Discovery: Is Something Wrong? *Drug Discov Today* **2005**, *10* (2), 139–147.
- (5) Blundell, T. L.; Jhoti, H.; Abell, C. High-Throughput Crystallography for Lead Discovery in Drug Design. *Nat Rev Drug Discov* **2002**, *1* (1), 45–54. <https://doi.org/10.1038/nrd706>.
- (6) Kuhn, P.; Wilson, K.; Patch, M. G.; Stevens, R. C. The Genesis of High-Throughput Structure-Based Drug Discovery Using Protein Crystallography. *Curr Opin Chem Biol* **2002**, *6* (5), 704–710. [https://doi.org/10.1016/S1367-5931\(02\)00361-7](https://doi.org/10.1016/S1367-5931(02)00361-7).
- (7) Salon, J. a; Lodowski, D. T.; Palczewski, K. The Significance of G Protein-Coupled Receptor. *Pharmacol Rev* **2011**, *63* (4), 901–937. <https://doi.org/10.1124/pr.110.003350.901>.
- (8) Sugiki, T.; Furuita, K.; Fujiwara, T.; Kojima, C. Current NMR Techniques for Structure-Based Drug Discovery. *Molecules* **2018**, *23* (1), 1–27. <https://doi.org/10.3390/molecules23010148>.
- (9) Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T. L.; Homans, S. W.; Kessler, H.; Luchinat, C.; et al. Perspectives on NMR in Drug Discovery: A Technique Comes of Age. *Nat Rev Drug Discov* **2008**, *7*, 738–745. <https://doi.org/10.1111/j.1365-2621.1965.tb01855.x>.
- (10) Pellecchia, M.; Sem, D. S.; Wüthrich, K. NMR in Drug Discovery. *Nat Rev Drug Discov* **2002**, *1* (3), 211–219. <https://doi.org/10.1038/nrd748>.
- (11) Muhammed, M. T.; Aki-Yalcin, E. Homology Modeling in Drug Discovery: Overview, Current Applications, and Future Perspectives. *Chem Biol Drug Des* **2019**, *93* (1), 12–20. <https://doi.org/10.1111/cbdd.13388>.
- (12) González, P. M.; Acharya, C.; MacKerell, A. D.; Polli, J. E. Inhibition Requirements of the Human Apical Sodium-Dependent Bile Acid Transporter (HASBT) Using Aminopiperidine Conjugates of Glutamyl-Bile Acids. *Pharm Res* **2009**, *26* (7), 1665–1678. <https://doi.org/10.1007/s11095-009-9877-3>.
- (13) Ekins, S.; Mirny, L.; Schuetz, E. G. A Ligand-Based Approach to Understanding Selectivity of Nuclear Hormone Receptors PXR, CAR, FXR, LXR α , and LXR β . *Pharm Res* **2002**, *19* (12), 1788–1800. <https://doi.org/10.1023/A:1021429105173>.
- (14) Aparoy, P.; Kumar Reddy, K.; Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors. *Curr Med Chem* **2012**, *19* (22), 3763–3778. <https://doi.org/10.2174/092986712801661112>.
- (15) Cappel, D.; Hall, M. L.; Lenselink, E. B.; Beuming, T.; Qi, J.; Bradner, J.; Sherman, W. Relative Binding Free Energy Calculations Applied to Protein Homology Models. *J Chem Inf Model* **2016**, *56* (12), 2388–2400. <https://doi.org/10.1021/acs.jcim.6b00362>.

-
- (16) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J Chem Inf Model* **2017**, *57* (12), 2911–2937. <https://doi.org/10.1021/acs.jcim.7b00564>.
- (17) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J Am Chem Soc* **2015**, *137* (7), 2695–2703. <https://doi.org/10.1021/ja512751q>.
- (18) Zhang, H.; Jiang, W.; Chatterjee, P.; Luo, Y. Ranking Reversible Covalent Drugs: From Free Energy Perturbation to Fragment Docking. *J Chem Inf Model* **2019**, *59* (5), 2093–2102. <https://doi.org/10.1021/acs.jcim.8b00959>.
- (19) Hung, C. L.; Chen, C. C. Computational Approaches for Drug Discovery. *Drug Dev Res* **2014**, *75* (6), 412–418. <https://doi.org/10.1002/ddr.21222>.
- (20) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew Chem. - Int Ed* **2002**, *41* (15), 2644–2676.
- (21) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting Binding Modes, Binding Affinities and “hot Spots” for Protein-Ligand Complexes Using a Knowledge-Based Scoring Function. *Perspect Drug Discov Des* **2000**, *20*, 115–144. <https://doi.org/10.1023/A:1008781006867>.
- (22) Sprous, D. G.; Palmer, R. K.; Swanson, J. T.; Lawless, M. QSAR in the Pharmaceutical Research Setting: QSAR Models for Broad, Large Problems. *Curr Top Med Chem* **2010**, *10* (6), 619–637. <https://doi.org/10.2174/156802610791111506>.
- (23) Neves, B. J.; Braga, R. C.; Melo-Filho, C. C.; Moreira-Filho, J. T.; Muratov, E. N.; Andrade, C. H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front Pharmacol* **2018**, *9* (NOV), 1–7. <https://doi.org/10.3389/fphar.2018.01275>.
- (24) Weber, A.; Böhm, M.; Supuran, C. T.; Scozzafava, A.; Sotriffer, C. A.; Klebe, G. 3D QSAR Selectivity Analyses of Carbonic Anhydrase Inhibitors: Insights for the Design of Isozyme Selective Inhibitors. *J Chem Inf Model* **2006**, *46* (6), 2737–2760. <https://doi.org/10.1021/ci600298r>.
- (25) Yoshida, F.; Topliss, J. G. QSAR Model for Drug Human Oral Bioavailability. *J Med Chem* **2000**, *43* (13), 2575–2585. <https://doi.org/10.1021/jm0000564>.
- (26) Gesty-Palmer, D.; Luttrell, L. M. *Refining Efficacy. Exploiting Functional Selectivity for Drug Discovery*, 1st ed.; Elsevier Inc., 2011; Vol. 62. <https://doi.org/10.1016/B978-0-12-385952-5.00009-9>.
- (27) Tan, L.; Yan, W.; McCorvy, J. D.; Cheng, J. Biased Ligands of G Protein-Coupled Receptors (GPCRs): Structure-Functional Selectivity Relationships (SFSRs) and Therapeutic Potential. *J Med Chem* **2018**, *61* (22), 9841–9878. <https://doi.org/10.1021/acs.jmedchem.8b00435>.
- (28) Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic Shotguns versus Magic Bullets: Selectively Non-Selective Drugs for Mood Disorders and Schizophrenia. *Nat Rev Drug Discov* **2004**, *3* (April), 353–359.
- (29) Ortiz, A.; Gomez-Puertas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Vinas, E.; Morreale, A.; Murcia, M.; Wang, K. Computational Approaches to Model Ligand Selectivity in Drug Design. *Curr Top Med Chem* **2005**, *6* (1), 41–55. <https://doi.org/10.2174/156802606775193338>.

- (30) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **1996**, *23*, 3–25.
- (31) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov Today Technol* **2004**, *1* (4), 337–341. <https://doi.org/10.1016/j.ddtec.2004.11.007>.
- (32) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J Med Chem* **2002**, *45* (12), 2615–2623. <https://doi.org/10.1021/jm020017n>.
- (33) Hersh, D. S.; Wadajkar, A. S.; Roberts, N. B.; Perez, J. G.; Connolly, N. P.; Frenkel, V.; Winkles, J. A.; Woodworth, G. F.; Kim, A. J. Evolving Drug Delivery Strategies to Overcome the Blood Brain Barrier. *Curr Pharm Des* **2016**, *22*, 1177–1193.
- (34) Li, X.; Chen, L.; Cheng, F.; Wu, Z.; Bian, H.; Xu, C.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods. *J Chem Inf Model* **2014**, *54* (4), 1061–1069. <https://doi.org/10.1021/ci5000467>.
- (35) Cronin, M. T. D.; Dearden, J. C. QSAR in Toxicology. 1. Prediction of Aquatic Toxicity. *Quant Struct Relatsh.* **1995**, *14* (1), 1–7. <https://doi.org/10.1002/qsar.19950140102>.
- (36) Baron, R.; McCammon, J. A. Molecular Recognition and Ligand Association. *Annu Rev Phys Chem* **2013**, *64* (1), 151–175. <https://doi.org/10.1146/annurev-physchem-040412-110047>.
- (37) Lehn, J. M. Towards Complex Matter: Supramolecular Chemistry and Self-Organization. *PNAS* **2002**, *99* (8), 4763–4768. <https://doi.org/10.1017/S1062798709000805>.
- (38) Lehn, J.-M. Supramolecular Chemistry Receptors, Catalysts, and Carriers. *Sci.* **1985**, *227* (4689), 849–856.
- (39) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem* **2010**, *53* (14), 5061–5084. <https://doi.org/10.1021/jm100112j>.
- (40) Kuhn, B.; Mohr, P.; Stahl, M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. *J Med Chem* **2010**, *53* (6), 2601–2611. <https://doi.org/10.1021/jm100087s>.
- (41) Borhani, D. W.; Shaw, D. E. The Future of Molecular Dynamics Simulations in Drug Discovery. *J Comput Aided Mol Des* **2012**, *26* (1), 15–26. <https://doi.org/10.1007/s10822-011-9517-y>.
- (42) Martin, S. F.; Clements, J. H. Correlating Structure and Energetics in Protein-Ligand Interactions: Paradigms and Paradoxes. *Annu Rev Biochem* **2013**, *82*, 267–93. <https://doi.org/10.1146/annurev-biochem-060410-105819>.
- (43) Wienen-Schmidt, B.; Jonker, H. R. A.; Wulsdorf, T.; Gerber, H. D. H.-D.; Saxena, K.; Kudlinzki, D.; Sreeramulu, S.; Parigi, G.; Luchinat, C.; Heine, A.; et al. Paradoxically, Most Flexible Ligand Binds Most Entropy-Favored: Intriguing Impact of Ligand Flexibility and Solvation on Drug-Kinase Binding. *J Med Chem* **2018**, *61* (14), 5922–5933. <https://doi.org/10.1021/acs.jmedchem.8b00105>.
- (44) DeLorbe, J. E.; Clements, J. H.; Teresk, M. G.; Benfield, A. P.; Plake, H. R.; Millspaugh, L. E.; Martin, S. F. Thermodynamic and Structural Effects of Conformational Constraints in Protein-Ligand Interactions. Entropic Paradox Associated with Ligand Preorganization. *J Am Chem Soc* **2009**, *131* (46), 16758–16770. <https://doi.org/10.1021/ja904698q>.
- (45) Benfield, A. P.; Teresk, M. G.; Plake, H. R.; DeLorbe, J. E.; Millspaugh, L. E.; Martin, S. F. Ligand Preorganization May Be Accompanied by Entropic Penalties in Protein-

- Ligand Interactions. *Angew Chem. - Int Ed* **2006**, *45* (41), 6830–6835. <https://doi.org/10.1002/anie.200600844>.
- (46) Biela, A.; Nasief, N. N.; Betz, M.; Heine, A.; Hangauer, D.; Klebe, G. Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin. *Angew Chem Int Ed Engl* **2013**, *52*, 1822–8. <https://doi.org/10.1002/anie.201208561>.
- (47) Betz, M.; Wulsdorf, T.; Krimmer, S. G.; Klebe, G. Impact of Surface Water Layers on Protein-Ligand Binding: How Well Are Experimental Data Reproduced by Molecular Dynamics Simulations in a Thermolysin Test Case? *J Chem Inf Model* **2016**, *56* (1), 223–233. <https://doi.org/10.1021/acs.jcim.5b00621>.
- (48) Biela, A.; Betz, M.; Heine, A.; Klebe, G. Water Makes the Difference: Rearrangement of Water Solvation Layer Triggers Non-Additivity of Functional Group Contributions in Protein-Ligand Binding. *Chemmedchem* **2012**, *7*, 1423–34. <https://doi.org/10.1002/cmcd.201200206>.
- (49) Biela, A.; Sielaff, F.; Terwesten, F.; Heine, A.; Steinmetzer, T.; Klebe, G. Ligand Binding Stepwise Disrupts Water Network in Thrombin: Enthalpic and Entropic Changes Reveal Classical Hydrophobic Effect. *J Med Chem* **2012**, *55* (13), 6094–110. <https://doi.org/10.1021/jm300337q>.
- (50) Snyder, P. W.; Mecinovic, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc Natl Acad Sci U S A* **2011**, *108* (44), 17889–17894. <https://doi.org/10.1073/pnas.1114107108>.
- (51) Ball, P. Water as an Active Constituent in Cell Biology. *Chem Rev* **2008**, *108* (1), 74–108. <https://doi.org/10.1021/cr068037a>.
- (52) Biedermann, F.; Nau, W. M.; Schneider, H. J. The Hydrophobic Effect Revisited - Studies with Supramolecular Complexes Imply High-Energy Water as a Noncovalent Driving Force. *Angew Chem. - Int Ed* **2014**, *53* (42), 11158–11171. <https://doi.org/10.1002/anie.201310958>.
- (53) Krimmer, S. G.; Cramer, J.; Schiebel, J.; Heine, A.; Klebe, G. How Nothing Boosts Affinity: Hydrophobic Ligand Binding to the Virtually Vacated S1' Pocket of Thermolysin. *J Am Chem Soc* **2017**, *139* (30), 10419–10431. <https://doi.org/10.1021/jacs.7b05028>.
- (54) Schiebel, J.; Gaspari, R.; Wulsdorf, T.; Ngo, K.; Sohn, C.; Schrader, T. E.; Cavalli, A.; Ostermann, A.; Heine, A.; Klebe, G. Intriguing Role of Water in Protein-Ligand Binding Studied by Neutron Crystallography on Trypsin Complexes. *Nat Commun* **2018**, *9* (1), 1–30. <https://doi.org/10.1038/s41467-018-05769-2>.
- (55) Bayly, 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* (40), 10269–10280.
- (56) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712–725. <https://doi.org/10.1002/prot>.
- (57) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>.

-
- (58) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J Am Chem Soc* **1988**, *110* (4), 1657–1666. <https://doi.org/10.1021/ja00214a001>.
- (59) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J Am Chem Soc* **1996**, *118* (15), 11225–11236.
- (60) Van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J Phys Chem A* **2001**, *105* (41), 9396–9409. <https://doi.org/10.1021/jp004368u>.
- (61) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; et al. The ReaxFF Reactive Force-Field: Development, Applications and Future Directions. *Npj Comput Mater* **2016**, *2* (September 2015). <https://doi.org/10.1038/npjcompumats.2015.11>.
- (62) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew Chem. - Int Ed* **2009**, *48* (7), 1198–1229. <https://doi.org/10.1002/anie.200802019>.
- (63) Stanton, C. L.; Kuo, I. F. W.; Mundy, C. J.; Laino, T.; Houk, K. N. QM/MM Metadynamics Study of the Direct Decarboxylation Mechanism for Orotidine-5'-Monophosphate Decarboxylase Using Two Different QM Regions: Acceleration Too Small to Explain Rate of Enzyme Catalysis. *J Phys Chem B* **2007**, *111* (43), 12573–12581. <https://doi.org/10.1021/jp074858n>.
- (64) Faraji, S.; Groenhof, G.; Dreuw, A. Combined QM/MM Investigation on the Light-Driven Electron-Induced Repair of the (6-4) Thymine Dimer Catalyzed by DNA Photolyase. *J Phys Chem B* **2013**, *117* (35), 10071–10079. <https://doi.org/10.1021/jp401662z>.
- (65) Rajamani, R.; Naidoo, K. J.; Gao, J. Implementation of an Adaptive Umbrella Sampling Method for the Calculation of Multidimensional Potential of Mean Force of Chemical Reactions in Solution. *J Comput Chem* **2003**, *24* (14), 1775–1781. <https://doi.org/10.1002/jcc.10315>.
- (66) Dewar, M. J. S.; Zuebis, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J Am Chem Soc* **1985**, *107* (13), 3902–3909. <https://doi.org/10.1021/ja00299a024>.
- (67) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J Mol Model* **2007**, *13* (12), 1173–1213. <https://doi.org/10.1007/s00894-007-0233-4>.
- (68) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScoreCSD-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of near-Native Ligand Poses and Better Affinity Prediction. *J Med Chem* **2005**, *48* (20), 6296–6303. <https://doi.org/10.1021/jm050436v>.
- (69) Verdonk, M. L.; Ludlow, R. F.; Giangreco, I.; Rathi, P. C. Protein-Ligand Informatics Force Field (PLIff): Toward a Fully Knowledge Driven “Force Field” for Biomolecular Interactions. *J Med Chem* **2016**, *59* (14), 6891–6902. <https://doi.org/10.1021/acs.jmedchem.6b00716>.
- (70) Gohlke, H.; Klebe, G. Drugscore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J Med Chem* **2002**, *45* (19), 4153–4170. <https://doi.org/10.1021/jm020808p>.
- (71) Wade, R. C.; Mazar, M. H.; McCammon, J. A.; Quiocho, F. A. A Molecular Dynamics Study of Thermodynamic and Structural Aspects of the Hydration of Cavities in Proteins. *Biopolymers* **1991**, *31* (8), 919–931. <https://doi.org/10.1002/bip.360310802>.

-
- (72) Helms, V.; Wade, R. C. Thermodynamics of Water Mediating Protein-Ligand Interactions in Cytochrome P450cam: A Molecular Dynamics Study. *Biophys J* **1995**, *69* (3), 810–824. [https://doi.org/10.1016/S0006-3495\(95\)79955-6](https://doi.org/10.1016/S0006-3495(95)79955-6).
- (73) Woo, H.-J.; Dinner, A. R.; Roux, B. B. Grand Canonical Monte Carlo Simulations of Water in Protein Environments. *J Chem Phys* **2004**, *121* (13), 6392–400. <https://doi.org/10.1063/1.1784436>.
- (74) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. Strategies to Calculate Water Binding Free Energies in Protein-Ligand Complexes. *J Chem Inf Model* **2014**, *54* (6), 1623–1633. <https://doi.org/10.1021/ci400674k>.
- (75) Bortolato, A.; Tehan, B. G.; Bodnarchuk, M. S.; Essex, J. W.; Mason, J. S. Water Network Perturbation in Ligand Binding: Adenosine A 2A Antagonists as a Case Study. *J Chem Inf Model* **2013**, *53* (7), 1700–1713. <https://doi.org/10.1021/ci4001458>.
- (76) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J Am Chem Soc* **2015**, *137* (47), 14930–14943. <https://doi.org/10.1021/jacs.5b07940>.
- (77) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J Phys Chem B* *113*, 13337–13346.
- (78) 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* (9), 2817–2831. <https://doi.org/10.1021/ja0771033>.
- (79) Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. Contribution of Explicit Solvent Effects to the Binding Affinity of Small-Molecule Inhibitors in Blood Coagulation Factor Serine Proteases. *Chemmedchem* **2011**, *6*, 1049–66. <https://doi.org/10.1002/cmde.201000533>.
- (80) Ramsey, S.; Nguyen, C.; Salomon-Ferrer, R.; Walker, R. C.; Gilson, M. K.; Kurtzman, T. Solvation Thermodynamic Mapping of Molecular Surfaces in AmberTools: GIST. *J Comput Chem* **2016**, 2029–2037. <https://doi.org/10.1002/jcc.24417>.
- (81) Uehara, S.; Tanaka, S. AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking. *Molecules* **2016**, *21* (11). <https://doi.org/10.3390/molecules21111604>.
- (82) Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J Chem Theory Comput* **2014**, *10* (7), 2769–2780. <https://doi.org/10.1021/ct401110x>.
- (83) Nguyen, C. N.; Young, T. K.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril. *J Chem Phys* **2012**, *137* (14), 1–17. <https://doi.org/10.1063/1.4751113>.
- (84) *SZMAP 1.0.0*; Openye Scientific Software Inc.: Santa Fe, NM, USA, 2011.
- (85) Grant, J. A. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J Comput Chem* **2001**, *22* (6), 608–640. <https://doi.org/10.1002/jcc.1032>.
- (86) Gerogiokas, G.; Calabro, G.; Henchman, R. H.; Southey, M. W. Y.; Law, R. J.; Michel, J. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *J Chem Theory Comput* **2014**, *10* (1), 35–48. <https://doi.org/10.1021/ct400783h>.
- (87) Gerogiokas, G.; Southey, M.; Mazanetz, M.; Heifetz, A.; Bodkin, M.; Law, R.; Michel, J. Evaluation of Water Displacement Energetics in Protein Binding Sites with Grid Cell Theory. *Phys Chem Chem Phys* **2015**, *17* (13). <https://doi.org/10.1039/C4CP05572A>.
- (88) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. Evaluation of Host-Guest Binding Thermodynamics of Model Cavities with Grid Cell Theory. *J Chem Theory Comput* **2014**, *10* (9), 4055–4068. <https://doi.org/10.1021/ct500368p>.

-
- (89) Lazaridis, T.; Paulaitis, M. E. Entropy of Hydrophobic Hydration: A New Statistical Mechanical Formulation. *Fluid Phase Equilib* **1993**, *83* (C), 43–49. [https://doi.org/10.1016/0378-3812\(93\)87005-L](https://doi.org/10.1016/0378-3812(93)87005-L).
- (90) Lazaridis, T.; Karplus, M. Orientational Correlations and Entropy in Liquid Water. *J Chem Phys* **1996**, *105* (2), 4294–4316. <https://doi.org/10.1063/1.472247>.
- (91) Lazaridis, T. Solvent Reorganization Energy and Entropy in Hydrophobic Hydration. *J Phys Chem B* **2000**, *104*, 4964–4979.
- (92) Ichihara, O.; Shimada, Y.; Yoshidome, D. The Importance of Hydration Thermodynamics in Fragment-to-Lead Optimization. *Chemmedchem* **2014**, *9* (12), 2708–2717. <https://doi.org/10.1002/cmdc.201402207>.
- (93) Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein-Ligand Binding. *J Am Chem Soc* **2013**, *135* (41), 15579–84. <https://doi.org/10.1021/ja4075776>.
- (94) Balius, T. E.; Fischer, M.; Stein, R. M.; Adler, T. B.; Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T.; Shoichet, B. K. Testing Inhomogeneous Solvation Theory in Structure-Based Ligand Discovery. *Pnas* **2017**, 1–8. <https://doi.org/10.1073/pnas.1703287114>.
- (95) 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* (1), 418–425. <https://doi.org/10.1021/acs.jctc.7b00592>.
- (96) Krimmer, S. G.; Betz, M.; Heine, A.; Klebe, G. Methyl, Ethyl, Propyl, Butyl: Futile but Not for Water, as the Correlation of Structure and Thermodynamic Signature Shows in a Congeneric Series of Thermolysin Inhibitors. *Chemmedchem* **2014**, *9*, 833–46. <https://doi.org/10.1002/cmdc.201400013>.
- (97) Snyder, P. W.; Lockett, M. R.; Moustakas, D. T.; Whitesides, G. M. Is It the Shape of the Cavity, or the Shape of the Water in the Cavity? *Eur Phys J Spec Top* **2014**, *223* (5), 853–891. <https://doi.org/10.1140/epjst/e2013-01818-y>.
- (98) Bodnarchuk, M. S. Water, Water, Everywhere... It's Time to Stop and Think. *Drug Discov Today* **2016**, *21* (7), 1139–1146. <https://doi.org/10.1016/j.drudis.2016.05.009>.
- (99) Rechlin, C.; Scheer, F.; Terwesten, F.; Wulsdorf, T.; Pol, E.; Fridh, V.; Toth, P.; Diederich, W. E.; Heine, A.; Klebe, G. Price for Opening the Transient Specificity Pocket in Human Aldose Reductase upon Ligand Binding: Structural, Thermodynamic, Kinetic, and Computational Analysis. *ACS Chem Biol* **2017**, *12* (5), 1397–1415. <https://doi.org/10.1021/acscchembio.7b00062>.
- (100) Wienen-Schmidt, B.; Wulsdorf, T.; Jonker, H. R. A.; Saxena, K.; Kudlinzki, D.; Linhard, V.; Sreeramulu, S.; Heine, A.; Schwalbe, H.; Klebe, G. On the Implication of Water on Fragment-to-Ligand Growth in Kinase Binding Thermodynamics. *ChemMedChem* **2018**, *13* (18). <https://doi.org/10.1002/cmdc.201800438>.
- (101) Nguyen, C.; Gilson, M. K.; Young, T. Structure and Thermodynamics of Molecular Hydration via Grid Inhomogeneous Solvation Theory. *arXiv* **2011**, 16.
- (102) Cui, G.; Swails, J. M.; Manas, E. S. SPAM: A Simple Approach for Profiling Bound Water Molecules. *J Chem Theory Comput* **2013**, *9* (12), 5539–5549. <https://doi.org/10.1021/ct400711g>.
- (103) Biela, A.; Khayat, M.; Tan, H.; Kong, J.; Heine, A.; Hangauer, D.; Klebe, G. Impact of Ligand and Protein Desolvation on Ligand Binding to the S1 Pocket of Thrombin. *J Mol Biol* **2012**, *418* (5), 350–366. <https://doi.org/10.1016/j.jmb.2012.01.054>.

- (104) Baum, B.; Mohamed, M.; Zayed, M.; Gerlach, C.; Heine, A.; Hangauer, D.; Klebe, G. More than a Simple Lipophilic Contact: A Detailed Thermodynamic Analysis of Nonbasic Residues in the S1 Pocket of Thrombin. *J Mol Biol* **2009**, *390*, 56–69. <https://doi.org/10.1016/j.jmb.2009.04.051>.
- (105) Baum, B.; Muley, L.; Heine, A.; Smolinski, M.; Hangauer, D.; Klebe, G. Think Twice: Understanding the High Potency of Bis(Phenyl)Methane Inhibitors of Thrombin. *J Mol Biol* **2009**, *391* (3), 552–564. <https://doi.org/10.1016/j.jmb.2009.06.016>.
- (106) Chelsea Collins. Back to Basics: Distinguishing Key Interactions in Protein-Ligand Relationships with Thrombin as a Model Protein. Diploma Thesis, Marburg University: Marburg, 2014.
- (107) Anna Sandner. Thermodynamic and Crystallographic Characterization of Preorganized Thrombin Ligands. Diploma Thesis, Marburg University: Marburg, 2015.
- (108) Winquist, J.; Geschwindner, S.; Xue, Y.; Gustavsson, L.; Musil, D.; Deinum, J.; Danielson, U. H. Identification of Structural-Kinetic and Structural-Thermodynamic Relationships for Thrombin Inhibitors. **2013**.
- (109) Rühmann, E.; Betz, M.; Heine, A.; Klebe, G. Fragment Binding Can Be Either More Enthalpy-Driven or Entropy-Driven: Crystal Structures and Residual Hydration Patterns Suggest Why. *J Med Chem* **2015**, *58* (17), 6960–6971. <https://doi.org/10.1021/acs.jmedchem.5b00812>.
- (110) Gerlach, C.; Smolinski, M.; Steuber, H.; Sottriffer, C. A.; Heine, A.; Hangauer, D. G.; Klebe, G. Thermodynamic Inhibition Profile of a Cyclopentyl and a Cyclohexyl Derivative towards Thrombin: The Same but for Different Reasons. *Angew Chem. - Int Ed* **2007**, *46* (44), 8511–8514. <https://doi.org/10.1002/anie.200701169>.
- (111) Steinmetzer, T.; Baum, B.; Biela, A.; Klebe, G.; Nowak, G.; Bucha, E. Beyond Heparinization: Design of Highly Potent Thrombin Inhibitors Suitable for Surface Coupling. *Chemmedchem* **2012**, *7* (11), 1965–1973. <https://doi.org/10.1002/cmde.201200292>.
- (112) Ahmed, H. U.; Blakeley, M. P.; Cianci, M.; Cruickshank, D. W. J.; Hubbard, J. a.; Helliwell, J. R. The Determination of Protonation States in Proteins. *Acta Crystallogr Sect Biol Crystallogr* **2007**, *63* (8), 906–922. <https://doi.org/10.1107/S09074444907029976>.
- (113) *Molecular Operating Environment (MOE), Version 2015.10 (2016) Chemical Computing Group Inc., Montreal.*
- (114) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J Comput Chem* **1995**, *16* (11), 1357–1377. <https://doi.org/10.1002/jcc.540161106>.
- (115) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09 Revision C.01*; Gaussian Inc. Wallingford CT 2009.
- (116) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J Comput Chem* **2004**, *25* (9), 1157–1174. <https://doi.org/10.1002/jcc.20035>.
- (117) Case, D. A.; Cerutti, D. S.; Cheatham, T. E. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; et al. *AMBER 2017*; University of California, San Francisco, 2017.
- (118) Horn, H. W.; Swope, W. C.; Pitara, J. W. Characterization of the TIP4P-Ew Water Model: Vapor Pressure and Boiling Point. *J Chem Phys* **2005**, *123* (19). <https://doi.org/10.1063/1.2085031>.

- (119) 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. [https://doi.org/10.1016/0021-9991\(77\)90098-5](https://doi.org/10.1016/0021-9991(77)90098-5).
- (120) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without Compromise - A Mixed Precision Model for GPU Accelerated Molecular Dynamics Simulations. *Comput Phys Commun* **2013**, *184* (2), 374–380. <https://doi.org/10.1016/j.cpc.2012.09.022>.
- (121) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Grand, S. L.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012**, *8* (5), 1542–1555. <https://doi.org/10.1021/ct200909j>.
- (122) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **2013**, *9* (9), 3878–3888. <https://doi.org/10.1021/ct400314y>.
- (123) Roe, D. R.; Cheatham III, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Com* **2013**, *9* (7), 3084–3095. <https://doi.org/10.1021/ct400341p>.
- (124) Matteo Aldeghi. *SplitVolume.Py*; <https://github.com/gosldorf/BiggerGist>.
- (125) Steven Ramsey. *Gistpp*; github.com/gosldorf/gist-post-processing.
- (126) *The PyMOL Molecular Graphics System, Version 1.8*; Schrödinger, LLC.
- (127) DeLano, W. L. *The PyMOL Molecular Graphics System*; Delano Scientific, San Carlos, 2002.
- (128) Miller, B. R.; Mcgee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py : An Efficient Program for End-State Free Energy Calculations. *J Chem Theory Comput* **2012**, *8*, 3314–3321.
- (129) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. An MM / 3D-RISM Approach for Ligand Binding Affinities. *J Phys Chem* **2010**, *114*, 8505–8516.
- (130) Onufriev, A.; Bashford, D.; Case, D.A. Modification of the Generalized Born Model Suitable for Macro- Molecules. *J Phys Chem B* **2000**, *104*, 3712–3720.
- (131) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct Funct Bioinforma* **2004**, *55* (2), 383–394. <https://doi.org/10.1002/prot.20033>.
- (132) Chandler, D.; Singh, Y.; Richardson, D. M. Excess Electrons in Simple Fluids. I. General Equilibrium Theory for Classical Hard Sphere Solvents. *J Chem Phys* **1984**, *81* (4), 1975–1982. <https://doi.org/10.1063/1.447820>.
- (133) Ichiye, T.; Chandler, D. Hypernetted Chain Closure Reference Interaction Site Method Theory of Structure and Thermodynamics for Alkanes in Water. *J Phys Chem* **1988**, *92* (18), 5257–5261. <https://doi.org/10.1021/j100329a037>.
- (134) Wales, D.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J Phys Chem A* **1997**, *101* (97), 5111–5116. <https://doi.org/10.1021/jp970984n>.
- (135) Francesco Biscani; Dario Izzo. *Pagmo2 & PyGMO (Version v2.10)*; <http://doi.org/10.5281/zenodo.2529931>.
- (136) Kraft, D. *A Software Package for Sequential Quadratic Programming*; Institut für Dynamik der Flugsysteme: Oberpfaffenhofen, 1988.
- (137) Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multi-Objective Genetic Algorithm: NSGAI. **2002**, *6* (2), 182–197.
- (138) Schechter, I.; Berger, A. On the Size of the Active Site in Proteases. I: Papain. *Biochem Biophys Res Commun* **1967**, *27* (2), 157–162.

- (139) *RDKit, Open-Source Cheminformatics*; <http://www.rdkit.org>.
- (140) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci* **1999**, *39* (5), 868–873. <https://doi.org/10.1021/ci9903071>.
- (141) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation Changes upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on PKa Calculations and ITC Experiments. *J Mol Biol* **2007**, *367* (5), 1347–1356. <https://doi.org/10.1016/j.jmb.2007.01.022>.
- (142) Dullweber, F.; Stubbs, M. T.; Musil, D.; Stürzebecher, J.; Klebe, G. Factorising Ligand Affinity: A Combined Thermodynamic and Crystallographic Study of Trypsin and Thrombin Inhibition. *J Mol Biol* **2001**, *313* (3), 593–614. <https://doi.org/10.1006/jmbi.2001.5062>.
- (143) Böhm, M.; Stürzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and FXa. *J Med Chem* **1999**, *42* (3), 458–477.
- (144) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J Am Chem Soc* **1988**, *110* (18), 5959–5967. <https://doi.org/10.1021/ja00226a005>.
- (145) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J Med Chem* **1994**, *37* (24), 4130–4146. <https://doi.org/10.1021/jm00050a010>.
- (146) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J Med Chem* **1985**, *28* (7), 849–57. <https://doi.org/10.1021/jm00145a002>.
- (147) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J Comput Aided Mol Des* **1991**, *5* (6), 545–552. <https://doi.org/10.1007/BF00135313>.
- (148) Sandner, A.; Hüfner-Wulsdorf, T.; Steinmetzer, T.; Klebe, G. Strategies in Late-Stage Optimization: Preorganization and Ligand-Induced Salt-Bridge Shielding in Thrombin. *submitted*.
- (149) Stubbs, M. T.; Oschkinat, H.; Mayr, I.; Huber, R.; Angliker, H.; Stone, S. R.; Bode, W. The Interaction of Thrombin with Fibrinogen. *Eur.J.Biochem.* **1992**, *206*, 187–195.
- (150) Swallow, S. Fluorine in Medicinal Chemistry. *Prog Med Chem* **2015**, *54*, 65–133. <https://doi.org/10.1016/bs.pmch.2014.11.001>.
- (151) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J Chem Inf Comput Sci* **1998**, *38* (3), 511–522. <https://doi.org/10.1021/ci970429i>.
- (152) Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and the Cambridge Structural Database. *J Chem Inf Model* **2010**, *50* (50), 572–584.
- (153) *OMEGA 3.1.1.2: OpenEye Scientific Software, Santa Fe, NM*; <http://www.eyesopen.com>.
- (154) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J Chem Phys* **1984**, *81* (8), 3684–3690. <https://doi.org/10.1063/1.448118>.

- (155) Klebe, G. The Use of Thermodynamic and Kinetic Data in Drug Discovery: Decisive Insight or Increasing the Puzzlement? *ChemMedChem* **2015**, *10* (2), 229–231. <https://doi.org/10.1002/cmde.201402521>.
- (156) Schmidtke, P.; Javier Luque, F.; Murray, J. B.; Barril, X. Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. *J Am Chem Soc* **2011**, *133* (46), 18903–18910. <https://doi.org/10.1021/ja207494u>.
- (157) Gaspari, R.; Rechlin, C.; Heine, A.; Bottegoni, G.; Rocchia, W.; Schwarz, D.; Bomke, J.; Gerber, H.-D.; Klebe, G.; Cavalli, A. Kinetic and Structural Insights into the Mechanism of Binding of Sulfonamides to Human Carbonic Anhydrase by Computational and Experimental Studies. *J Med Chem* **2015**, *acs.jmedchem.5b01643*. <https://doi.org/10.1021/acs.jmedchem.5b01643>.
- (158) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat Commun* **2015**, *6* (May). <https://doi.org/10.1038/ncomms8653>.
- (159) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc Natl Acad Sci* **2011**, *108* (32), 13118–13123. <https://doi.org/10.1073/pnas.1104614108>.
- (160) Heyden, M.; Havenith, M. Combining THz Spectroscopy and MD Simulations to Study Protein-Hydration Coupling. *Methods* **2010**, *52* (1), 74–83. <https://doi.org/10.1016/j.ymeth.2010.05.007>.
- (161) Liepinsh, E.; Otting, G.; Wüthrich, K. NMR Observation of Individual Molecules of Hydration Water Bound to DNA Duplexes: Direct Evidence for a Spine of Hydration Water Present in Aqueous Solution. *Nucleic Acids Res* **1992**, *20* (24), 6549–6553. <https://doi.org/10.1093/nar/20.24.6549>.
- (162) Zhou, D.; Bryant, R. G. Water Molecule Binding and Lifetimes on the DNA Duplex d(CGCGAATTCGCG)₂. *J Biomol NMR* **1996**, *8* (1), 77–86. <https://doi.org/10.1007/BF00198141>.
- (163) Jorge, C.; Marques, B. S.; Valentine, K. G.; Wand, A. J. *Characterizing Protein Hydration Dynamics Using Solution NMR Spectroscopy*, 1st ed.; Elsevier Inc., 2019; Vol. 615. <https://doi.org/10.1016/bs.mie.2018.09.040>.
- (164) Mattea, C.; Qvist, J.; Halle, B. Dynamics at the Protein-Water Interface from ¹⁷O Spin Relaxation in Deeply Supercooled Solutions. *Biophys J* **2008**, *95* (6), 2951–2963. <https://doi.org/10.1529/biophysj.108.135194>.
- (165) Szyk, \Lukasz; Yang, M.; Nibbering, E. T. J.; Elsaesser, T. Ultrafast Vibrational Dynamics and Local Interactions of Hydrated DNA. *Angew Chem. - Int Ed* **2010**, *49* (21), 3598–3610. <https://doi.org/10.1002/anie.200905693>.
- (166) Fogarty, A. C.; Laage, D. Water Dynamics in Protein Hydration Shells: The Molecular Origins of the Dynamical Perturbation. *J Phys Chem B* **2014**, *118* (28), 7715–29. <https://doi.org/10.1021/jp409805p>.
- (167) Sterpone, F.; Stirnemann, G.; Laage, D. Magnitude and Molecular Origin of Water Slowdown next to a Protein. *J Am Chem Soc* **2012**, *134* (9), 4116–4119. <https://doi.org/10.1021/ja3007897>.
- (168) Duboué-Dijon, E.; Fogarty, A. C.; Hynes, J. T.; Laage, D. Dynamical Disorder in the DNA Hydration Shell. *J Am Chem Soc* **2016**, *138* (24), 7610–7620. <https://doi.org/10.1021/jacs.6b02715>.
- (169) Ho, S. J.; Brighton, T. A. Ximelagatran: Direct Thrombin Inhibitor. *Vasc Health Risk Manag* **2006**, *2* (1), 49–58. <https://doi.org/10.2147/vhrm.2006.2.1.49>.

- (170) Straub, A.; Roehrig, S.; Hillisch, A. Oral, Direct Thrombin and Factor Xa Inhibitors: The Replacement for Warfarin, Leeches, and Pig Intestines? *Angew Chem. - Int Ed* **2011**, *50* (20), 4574–4590. <https://doi.org/10.1002/anie.201004575>.
- (171) Rühmann, E.; Betz, M.; Fricke, M.; Heine, A.; Schäfer, M.; Klebe, G. Thermodynamic Signatures of Fragment Binding: Validation of Direct versus Displacement ITC Titrations. *Biochim Biophys Acta - Gen Subj* **2015**, *1850* (4), 647–656. <https://doi.org/10.1016/j.bbagen.2014.12.007>.
- (172) Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. Residence Times of Water Molecules in the Hydration Sites of Myoglobin. *Biophys J* **2000**, *79* (6), 2966–2974. [https://doi.org/10.1016/S0006-3495\(00\)76533-7](https://doi.org/10.1016/S0006-3495(00)76533-7).
- (173) Laage, D.; Hynes, J. T. On the Residence Time for Water in a Solute Hydration Shell: Application to Aqueous Halide Solutions. *J Phys Chem B* **2008**, *112* (26), 7697–7701. <https://doi.org/10.1021/jp802033r>.
- (174) Saha, D.; Supekar, S.; Mukherjee, A. Distribution of Residence Time of Water around DNA Base Pairs: Governing Factors and the Origin of Heterogeneity. *J Phys Chem B* **2015**, *119* (34), 11371–11381. <https://doi.org/10.1021/acs.jpcc.5b03553>.
- (175) Impey, R. W.; Madden, P. A.; McDonald, I. R. Hydration and Mobility of Ions in Solution. *J Phys Chem* **1983**, *87* (25), 5071–5083. <https://doi.org/10.1021/j150643a008>.
- (176) 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* (34), 8743–8756. <https://doi.org/10.1021/acs.jpcc.6b01094>.
- (177) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. *J Comput Chem* **1995**, *16* (11), 1339–1350. <https://doi.org/10.1002/jcc.540161104>.
- (178) Roux, B. The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput Phys Commun* **1995**, *91* (1–3), 275–282. [https://doi.org/10.1016/0010-4655\(95\)00053-I](https://doi.org/10.1016/0010-4655(95)00053-I).
- (179) Kurisaki, I.; Barberot, C.; Takayanagi, M.; Nagaoka, M. Dewetting of S1-Pocket via Water Channel upon Thrombin-Substrate Association Reaction. *J Phys Chem B* **2015**, *119* (52), 15807–15812. <https://doi.org/10.1021/acs.jpcc.5b09581>.
- (180) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-mbox-BCC Model: II. Parameterization and Validation. *J Comput Chem* **2002**, *23* (16), 1623–41. <https://doi.org/10.1002/jcc.10128>.
- (181) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J Comput Chem* **2000**, *21* (2), 132–146.
- (182) Horn, H. W.; Swope, W. C.; Pitner, 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* (20), 9665–9678. <https://doi.org/10.1063/1.1683075>.
- (183) Park, S.; Tajkhorshid, E.; Schulten, K.; Jensen, M. Ø. Energetics of Glycerol Conduction through Aquaglyceroporin GlpF. *Proc Natl Acad Sci* **2002**, *99* (10), 6731–6736. <https://doi.org/10.1073/pnas.102649299>.
- (184) Crespo, A.; Martí, M. A.; Estrin, D. A.; Roitberg, A. E. Multiple-Steering QM-MM Calculation of the Free Energy Profile in Chorismate Mutase. *J Am Chem Soc* **2005**, *127* (19), 6940–6941. <https://doi.org/10.1021/ja0452830>.
- (185) Grossfield, A. *An Implementation of WHAM: The Weighted Histogram Analysis Method*; 2004.

-
- (186) Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans Math Softw* **1997**, *23* (4), 550–560. <https://doi.org/10.1145/279232.279236>.
- (187) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J Sci Comput* **1995**, *16* (5), 1190–1208. <https://doi.org/10.1137/0916069>.
- (188) Jones, E.; Oliphant, T.; Peterson, P.; others. *SciPy: Open Source Scientific Tools for Python*; 2001.
- (189) Laage, D.; Hynes, J. T. On the Molecular Mechanism of Water Reorientation. *J Phys Chem B* **2008**, *112* (45), 14230–14242. <https://doi.org/10.1021/jp805217u>.
- (190) Boisson, J.; Stirnemann, G.; Laage, D.; Hynes, J. T. Water Reorientation Dynamics in the First Hydration Shells of F- and I-. *Phys Chem Chem Phys* **2011**, *13* (44), 19895. <https://doi.org/10.1039/c1cp21834d>.

Acknowledgements

Mein ausgesprochener Dank gilt meinem Doktorvater und Mentor Prof. Dr. Gerhard Klebe. Sein Enthusiasmus und Hingabe zu naturwissenschaftlicher Forschung gepaart mit viel Geduld und Empathie sind einmalig und unersetzbar. Ich möchte mich sowohl für das entgegengebrachte Vertrauen, als auch für die viele gemeinsame Arbeit bedanken bei der ich so viel Neues lernen durfte. Darüber hinaus gilt mein ausgesprochener Dank Prof. Dr. Peter Kolb. Zum einen bedanke ich mich für die Übernahme des Zweitgutachtens meiner Doktorarbeit, zum anderen aber auch für seine stetige Bereitschaft zu Diskussionen und konstruktivem Feedback. Im Weiteren bedanke ich mich bei Frau Prof. Dr. Cornelia Keck und Herrn Prof. Dr. Moritz Bünemann dafür, dass sie sich dazu bereiterklärt haben meiner Prüfungskommission anzugehören.

Ein ausgesprochener Dank gilt meinen Kollegen Dr. Michael Betz und Dr. Alexander Metz für die umfassende Einführung in die Thematik und das Programmieren. Weiterer Dank gilt den ehemaligen und aktuellen Computer-Administratoren der AG Klebe und AG Kolb für die stetige und hingabevolle Pflege unseres Computer-Netzwerkes: Dr. Felix Terwesten, Dr. Timo Krotzky, Thomas Rickmeyer, Dr. Michael Betz, Dr. Denis Schmidt, Dr. Corey Taylor, Frank Balzer, Victor Lim und Matthäus Drabek. Ein besonderer Dank gilt Lydia Hartleben, die immer zuverlässig und schnell jegliche organisatorische Arbeit übernommen hat und auch für große bürokratische Knoten einfache Lösungen gefunden hat.

Ganz herzlich möchte ich mich bei meinem Bürokollegen Steffen Glöckner bedanken. Nicht nur, dass wir uns stets fernab unserer Forschungsarbeit austauschen konnten, sondern auch, dass wir zusammen konstruktiv über unsere, thematisch maximal unterschiedliche, Forschungsarbeit miteinander sprechen konnten, machte die gemeinsame Zeit unvergesslich.

Pharmazeutisch-chemische Forschung ist Team-Arbeit. Daher möchte ich meinen Kollegen und Kooperationspartnern danken, allen voran Dr. Barbara Wienen-Schmidt, Dr. Henry Jonker, Prof. Dr. Andreas Heine, Dr. Felix Terwesten, Dr. Chris Rechlin, Dr. Johannes Schiebel, Dr. Jonathan Cramer, Dr. Steffan Krimmer, Dr. Janis Müller, Dr. Alexander Metz, Anna Sandner, Prof. Dr. Torsten Steinmetzer, Khang Ngo, Dr. Engi Hassaan und Francesca Magari. Meinen beiden Vertiefer-Studenten Janik Hedderich und Torben Gutermuth möchte ich danken für ihr großes Interesse an der Thematik und ihre großartige Mitarbeit an unseren Projekten.

I'd like to acknowledge Dr. Mike Gilson (UCSD), Dr. Tom Kurtzman (CUNY) for hosting me a research visit in spring and fall of 2016. Further thanks go out to the members of their research groups at UCSD and CUNY: Dr. Steven Ramsey, Anthony Cruz, Dr. Kamran Haider, Dr. Ido Ben-Shalom and Dr. Niel Henriksen.

Mein abschließender Dank gilt meiner Familie. Bei meinen Eltern möchte ich mich bedanken, da ich mich bei allen Plänen und Vorhaben stets auf ihre Unterstützung verlassen konnte. Bei meiner Frau Rebekka bedanke ich mich für ihre Loyalität und unermüdliche moralische Unterstützung während meines Studiums und meiner Promotion.

Erklärung

Ich versichere, dass ich meine Dissertation

Expanding the Toolbox for Computational Analysis in Rational Drug Discovery: Using Biomolecular Solvation to Predict Thermodynamic, Kinetic and Structural Properties of Protein-Ligand Complexes

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe. Alle vollständig oder sinngemäß übernommenen Zitate sind als solche gekennzeichnet.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den.....

.....
Tobias Hüfner