



Molecular Simulation of Multivalent Ligand-Receptor Systems

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

eingereicht am
Fachbereich Mathematik und Informatik
- Fachrichtung Bioinformatik -
der Freien Universität Berlin

vorgelegt von
Alexander Bujotzek
aus Arnsberg

Berlin, 18. Februar 2013

Die vorliegende Arbeit wurde unter Anleitung von PD Dr. Marcus Weber von April 2009 bis Februar 2013 am Zuse-Institut Berlin, Abteilung Numerische Analysis und Modellierung, Arbeitsgruppe Mathematischer Molekülewurf durchgeführt.

1. Gutachter: PD Dr. Marcus Weber,
Numerische Analysis und Modellierung,
Mathematischer Molekülewurf,
Zuse-Institut Berlin

2. Gutachter: Prof. Dr. Oliver Seitz,
Institut für Chemie,
Bioorganische Synthese,
Humboldt-Universität zu Berlin

Tag der Disputation: 19.06.2013

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht.

Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum

Unterschrift

“The needs of the many outweigh the needs of the few – or the one.”

Spock - in STAR TREK: THE WRATH OF KHAN

Acknowledgements

A long, strange, and beautiful journey slowly comes to an end, and I want to seize this opportunity for thanking all the people who have accompanied me along the way.

First and foremost, I would like to thank my supervisor, PD Dr. Marcus Weber, for giving me the great opportunity to write a doctoral thesis in his group, and for patient and unconditional scientific and personal support at literally all the time during the many years I have spent at the Zuse institute.

I would like to thank my collaborators and fellow Ph.D. students from SFB-765, Dr. Min Shan, Frank Abendroth, Dr. Christian Scheibe, Dr. Hendrik Eberhard, Larissa von Krbek and Andreas Achazi, as well as their supervisors Prof. Dr. Rainer Haag, Prof. Dr. Oliver Seitz, Prof. Dr. Christoph Schalley and Prof. Dr. Beate Paulus for providing and sharing their exciting multivalent systems and research ideas. Without their hard work and dedication for our collaborative projects, finishing my thesis in this multidisciplinary form would not have been possible.

I would like to thank Ole Schütt for his tremendous programming efforts in the ZIB-MolPy project during his time in the working group. Without his dedication and skills, the software would not be in the shape it is in now, and it has been a great pleasure working with him.

I would like to thank Adam Nielsen for helping me with the P matrices, and Dr. Konstantin “Max” Fackeldey for moral and mathematical support on different occasions.

I would like to thank the HLRN support team at the Zuse institute, in particular Dr. Bernd Kallies, for providing and maintaining a powerful working environment (and the necessary NPL) that made the more demanding simulations possible.

I would like to thank all former and present members of the Weber group for a lot of memorable moments, in particular Vedat Durmaz for being a pleasant colleague and room mate (and I dare to say friend) for many years.

I would like to thank the organizers of SFB-765 and its graduate school, in particular Prof. Dr. Rainer Haag, Dr. Carlo Fastig, Katrin Wittig, Katharina Tebel and Prof. Dr. Hackenberger for creating a pleasant and multidisciplinary research environment for the doctoral students in SFB-765. The funding of SFB-765 by the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

Finally, I would like to thank my family and friends in Berlin for their love and support, especially my significant other, Wang Zheng, for her patience and encouragement.

Contents

Eidesstattliche Erklärung	iv
Acknowledgements	vi
List of Figures	xi
List of Tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Multivalency, an important concept in nature	1
1.2 Synthetic multivalency	3
1.3 Architecture and design of multivalent compounds	5
1.4 Molecular simulation of multivalent systems	8
2 Multivalent interactions – a theoretical background	13
2.1 Terms and definitions	13
2.2 Multivalent thermodynamics	16
2.2.1 The role of enthalpy	17
2.2.2 The role of entropy	19
2.3 Multivalent kinetics	29
2.4 Cooperativity in multivalent systems	35
3 Molecular simulation – a theoretical background	43
3.1 Basics of classical molecular simulation	43
3.1.1 Phase space and force field	45
3.1.2 The canonical ensemble	49
3.1.3 Markov chain Monte Carlo	53
3.1.4 Molecular dynamics	55
3.1.5 Thermostats	57
3.2 Conformational analysis with ZIBgridfree	60
3.2.1 Conformation dynamics	62
3.2.2 Internal coordinates	64
3.2.3 Implementing the potential modification	66
3.2.4 Adaptive refinement of the partitioning	68
3.2.5 Reweighting and cluster analysis	69
3.3 Conformational entropy estimation	73

4	Flexible spacer systems	79
4.1	Introduction	79
4.1.1	Estrogen receptor as a drug target	80
4.1.2	Bivalent inhibition of estrogen receptor	82
4.2	Study F1: Towards a rational spacer design for bivalent inhibition of estrogen receptor	84
4.2.1	Looking for spacer design paradigms	85
4.2.2	OEG spacers and novel semi-flexible spacer designs	86
4.2.3	Results	87
4.2.4	Discussion	95
4.3	Study F2: Conformational analysis of bivalent estrogen receptor-ligands – From intramolecular to intermolecular binding	98
4.3.1	Bivalent ligand design	98
4.3.2	Biological evaluation	100
4.3.3	Results and discussion (computational part)	106
4.4	Conclusion	111
4.5	Experimental setup	112
4.5.1	Modeling and simulation	112
4.5.2	Conformational entropy estimation	113
4.5.3	Visualization of conformational density	114
5	Rigid spacer systems	115
5.1	Introduction	115
5.2	Study R1: DNA-programmed spatial screening of carbohydrate-lectin interactions	117
5.2.1	Carbohydrate interactions benefit from multivalent presentation	118
5.2.2	Design and evaluation of LacNAc-PNA-DNA complexes	119
5.2.3	Results and discussion (computational part)	122
5.3	Study R2: DNA-controlled bivalent presentation of ligands for the estrogen receptor	126
5.3.1	Rigid and semi-rigid DNA scaffolds for the bivalent presentation of estrogen receptor ligands	127
5.3.2	Results and discussion (computational part)	130
5.4	Conclusion	137
5.5	Experimental setup	139
5.5.1	Study R1	139
5.5.2	Study R2	140
6	Multivalent binding processes	143
6.1	Introduction	143
6.2	How to presample a bivalent host-guest binding process	146
6.3	Results and discussion	149
6.3.1	Discretization of conformational space	149
6.3.2	Metastability analysis	150
6.4	Conclusion	159
6.5	Experimental setup	160
7	Conclusion and outlook	163

A	Validation of a ZIBgridfree-GROMACS hybrid sampling framework	167
A.1	Introduction	167
A.2	Results	169
A.2.1	Pentane <i>in vacuo</i>	169
A.2.2	Alanine dipeptide in water	172
A.3	Conclusion	176
A.4	Experimental setup	176
	Bibliography	179
	Zusammenfassung	199

List of Figures

1.1	A burr	1
1.2	The leukocyte adhesion cascade	2
1.3	Schematic representation of influenza virus entering a cellular host	3
1.4	Schematic representation of a virus binding to a cell surface	4
1.5	Different types of scaffolds for presenting multiple ligands	6
2.1	Schematized diagrams of dimeric IgA and pentameric IgM	14
2.2	Metal-EDTA chelate complex	15
2.3	Enthalpic change upon binding for a bivalent ligand	17
2.4	A simple lattice model of diffusion	19
2.5	Entropy loss upon binding for different scenarios	22
2.6	Entropy loss upon binding for different scenarios	24
2.7	Dually tyrosine-phosphorylated ITAM bound to the tSH2 domain of Syk .	25
2.8	Potential energy plots of monovalent and bivalent scenario	27
2.9	Conformational space representation of monovalent and bivalent scenario	28
2.10	Schematic of PEG-linked dimers binding to a tetrameric channel	30
2.11	Kinetic model of a trivalent ligand-receptor system	31
2.12	Stepwise equilibria defining the effective molarities for formation of the hexameric complex	34
2.13	Schematic representation of rebinding in a trivalent complex	34
2.14	Binding of a monovalent ligand to a bivalent receptor	36
2.15	Binding of a bivalent ligand to a bivalent receptor	38
2.16	Speciation profiles in the absence and in the presence of chelate cooperativity	38
3.1	Plot of a harmonic bond-stretching potential	47
3.2	Plot of a Lennard-Jones potential	48
3.3	Soft partitioning of a torsion angle by overlapping basis functions	64
3.4	A set of internal coordinates for describing the state of a bivalent host- guest system	65
3.5	Sampling of a torsion angle distribution with ZIBgridfree	67
3.6	A failed convergence test triggers an automatic refinement of the parti- tioning	68
3.7	Torsion angle distribution of the two torsion angles of n -pentane at 300 K	70
3.8	Schematic of a transition matrix in the presence of metastable subsets . .	72
4.1	Secondary structure representations of the crystal structures of the ER α ligand binding domain	81
4.2	Secondary structure representations of the crystal structures of the ER β ligand binding domain	83

4.3	Structures of the spacer designs involved in the computational study . . .	86
4.4	Two copies of DES coupled via amide to EG11, forming a bivalent ligand for ER	87
4.5	Histogram of bromine-bromine distances measured over 100 ns	88
4.6	Histogram of nitrogen-nitrogen distances measured over 100 ns	89
4.7	Impact of ligand attachment on spacer end-to-end distance mean and standard deviation	90
4.8	Conformational entropy comparison of bromine-capped spacers and DES-coupled spacers	90
4.9	Torsion angle types with different degrees of conformational entropy contribution	91
4.10	Conformational entropy contribution of different types types of torsion angles	92
4.11	Conformation density of bromine-capped spacer 3	93
4.12	Conformation density plot of spacer 3	93
4.13	Conformation density plot of spacer 5	94
4.14	In water, hydrophobic interactions promote stacked, sandwich-like conformations	95
4.15	Different scenarios of bivalent binding that can be modulated by the structure of spacer or scaffold	97
4.16	Structures of the ER agonist estradiol, the SERM raloxifene, its derivative LY156681, and the novel bivalent and monovalent RAL ligands	99
4.17	Relationship between the maximum spacer length and the relative binding affinity of of bivalent and monovalent ligands	103
4.18	Schematic of the ER α LBD dimer: Four different binding modes for bivalent ligands binding to ER α LBD	103
4.19	Aromatic and OEG regions of the ^1H NMR and ROESY spectra of OEG-tethered bivalent ligands in a solution of DMSO with D $_2$ O	105
4.20	Conformational density of compound 15 bound to ER α	107
4.21	Overlay of bivalent ligand 15 with LY156681, bound to the steroid binding pockets of ER α	108
4.22	Distance comparison between two LY156681 ligands and two RAL moieties as part of bivalent ligand 15 while bound to the ER α dimer	108
4.23	Conformational entropy loss of bivalent ligands 9–15 upon binding to ER α	109
4.24	Different sampling outcomes for compounds with short spacers	110
5.1	Crystal structure of ECL with <i>N</i> -linked oligosaccharide and lactose	118
5.2	Modular assembly of multipartite PNA-DNA complexes to control the valency and spatial arrangement of a multivalent display of glycoligands	119
5.3	Influence of the LacNAc-LacNAc distance on the binding affinities of the bivalent complexes	121
5.4	Mean force acting on the strand in the bent conformation with adjusted LacNAc-LacNAc distance	124
5.5	Conformations obtained by MD simulations in which a constant force acted on the strand to bend the complexes into a conformation suitable for bivalent binding	125

5.6	Conformation of complex 22 obtained by MD simulations in which a constant force acted on the strand to bend the complex into a conformation suitable for bivalent binding	126
5.7	Relative binding affinity (RBA) for the relevant complexes and compounds	128
5.8	The nucleotide-Ral conjugate incorporates a semi-flexible linker that ascertains an adequate distance between Ral moieties and DNA duplex . . .	130
5.9	Proposed binding modes of bivalent complexes Ral4₃ , Ral4₂ , and Ral4₄ to ER α	132
5.11	Overlay of 4Ral₀ -ER α and Ral2₁ /RAL2EG1-ER α complex adapting an intramolecular bivalent binding mode	135
5.12	Mean distance and mean torsion angle between to nucleobases divided by a distance of 15 bp as a function of the number of unpaired nucleotides .	136
6.1	Structures of the bivalent host-guest system DiC6-(G3+2H) and the monovalent control C6-(MonoG1+H)	147
6.2	An exemplary set of states from an ensemble of binding paths for a bivalent host-guest system generated with the vacuum shotgun method	147
6.3	Host guest systems C6-(MonoG1+H)-OTs and DiC6-(G3+2H)-2OTs after a 2 ns equilibration of the solvent mixture	148
6.4	Three distances between ammonium moiety and binding site form the internal coordinates for system C6-(MonoG1+H)-OTs	150
6.5	Mean potential energy and corrected discretization node weights for system C6-(MonoG1+H)-OTs	151
6.6	Mean potential energy and corrected discretization node weights for system DiC6-(G3+2H)-2OTs	152
6.7	Overlap integral matrix S for C6-(MonoG1+H)-OTs and DiC6-(G3+2H)-2OTs	153
6.8	The χ^T matrix of system C6-(MonoG1+H)-OTs groups the 16 discretization nodes into three metastable states	154
6.9	The χ^T matrix of system DiC6-(G3+2H)-2OTs groups the 31 discretization nodes into three metastable states	155
6.10	Snapshots from an MD simulation of system DiC6-(G3+2H)-2OTs that show the positioning of the OTs counter ion(s) during the binding process, and after formation of the doubly bound cyclic complex	156
A.1	Three-dimensional representation of n -pentane	169
A.2	Conformational weights of n -pentane	170
A.3	Conformational weights of n -pentane	172
A.4	Three-dimensional representation of alanine dipeptide	172
A.5	Conformational weights of alanine dipeptide	174
A.6	Conformational weights of alanine dipeptide	175

List of Tables

1.1	A selection of “wet” measurement methods for quantifying multivalent interactions	10
3.1	A selection of methods for reducing the computational cost of force field evaluations in classical molecular simulation	49
3.2	Different types of ensembles in statistical thermodynamics	51
4.1	Maximum spanned distance of spacers 1–6 in the fully extended conformation	87
4.2	Bromine-bromine distance data measured over 100 ns	88
4.3	Nitrogen-nitrogen distance data measured over 100 ns	89
4.4	Structure, maximum spacer length, relative binding affinity and effective concentration of bivalent and monovalent ligands	101
5.1	LacNAc–LacNAc presentation distances and angles as measured during 20 ns of MD simulation	123
6.1	Computational cost of ZIBgridfree in terms of nanoseconds of MD for systems C6-(MonoG1+H)-OTs and DiC6-(G3+2H)-2OTs	159
A.1	Conformational weights of <i>n</i> -pentane at 300 K, derived from a HMC simulation	170
A.2	Averaged conformational weights of <i>n</i> -pentane at 300 K, derived from ten runs of ZIBgridfree	171
A.3	Conformational weights of alanine dipeptide at 300 K in the <i>NVT</i> and in the <i>NpT</i> ensemble, derived from two 200 ns MD simulations	173
A.4	Averaged conformational weights of alanine dipeptide at 300 K in the <i>NVT</i> and in the <i>NpT</i> ensemble, derived from ten runs of ZIBgridfree	175

Abbreviations

AFM	A tom F orce M icro S copy
CNG	C yclic- N ucleotide- G ated channel
DES	D i E thyl S tilbestrol
DFT	D ensity F unctional T heory
DOSY	D iffusion O rdered S pectroscop Y
E₂	E stradiol
EDTA	E thylene D iamine T etraacetic A cid
EG	E thylene G lycol
EM	E ffective M olarity
ER	E strogen R eceptor
FRET	F örster (F luorescence) R esonance E nergy T ransfer
HMC	H ybrid M onte C arlo
hPG	hyperbranched P oly G lycerol
HPLC	H igh P erformance L iquid C hromatography
ITC	I sothermal T itration C alorimetry
LBD	L igand B inding D omain
MD	M olecular D ynamics
MSM	M arkov S tate M odel
NMR	N uclear M agnetic R esonance
OEG	O ligo(E thylene G lycol)
PCCA+	R obust P erron C luster C luster A nalysis
PEG	P oly(E thylene G lycol)
PLD	P EG L inked D imer
PNA	P eptide N ucleic A cid
RBA	R elative B inding A ffinity
RMSD	R oot M ean S quare D eviation
ROESY	R otating-frame nuclear O verhauser E ffect correlation S pectroscop Y
SD	S tochastic D ynamics
SERM	S elective E strogen R eceptor M odulator
SMD	S teered M olecular D ynamics

SPR	S urface P lasmon R esonance
TEM	T ransmission E lectron M icroscopy

*Gewidmet meinen Eltern, Elisabeth und Peter – für multivalente
Liebe und Unterstützung in allen Lebenslagen*

Chapter 1

Introduction

1.1 Multivalency, an important concept in nature



FIGURE 1.1: A burr – an example for large-scale multivalency in nature. Photo by Christian Fischer.

Multivalency is a key principle in nature. By means of multivalency, nature is able to create interactions that are both strong and reversible. A well known botanical example is the burdock, which, by displaying a great number of microscopic hooks on its prickly seed capsules (called burrs), manages to achieve firm attachment to the fur of animals or the clothing of men, thus providing an excellent mechanism for seed dispersal. As was recently pointed out by Haag *et al.* [1], man has copied this principle in order to create Velcro[®], where firm and reversible attachment of two opposing textile surfaces is achieved by presenting a very large number of tiny hooks on the one component, and an equally large number of tiny loops on the other component. When the two components meet, the hooks catch in the loops, and the two pieces fasten (or bind) temporarily. While the interaction between a single hook and a single loop is

negligibly weak, the strength of this connection is achieved only by the high number of simultaneously interacting sites – by multivalency.

In nature, this principle is observable not only on the macroscopic, but also on the molecular level. In a biological context, the tiny hooks and loops are represented by highly specific and complementary ligands and receptors.

The human immune system, for example, depends on multivalent interactions in order to route leukocytes to sites of tissue damage or injury, a fact that has been pointed out by Whitesides and coworkers [2]: During an inflammatory response, stimuli such as histamine and thrombin cause endothelial cells to mobilize selectin molecules (a family of receptors that mediate cell adhesion) from stores inside the cell to the cell surface. As the leukocyte rolls along the blood vessel wall, the lectin-like domain of the endothelial selectin binds to sialyl-Lewis^X (sLe^X) carbohydrate groups presented on proteins on the leukocyte, which slows the cell and allows it to leave the blood vessel and enter the site of infection. The low-affinity nature of selectins is what allows the characteristic “rolling” action attributed to leukocytes during the leukocyte adhesion cascade. In order to slow down the leukocyte at the right position and initiate firm adhesion and transmigration of the leukocyte into the injured tissue, the multivalent presentation of both selectin receptors and carbohydrate ligands is a key factor (Figure 1.2).

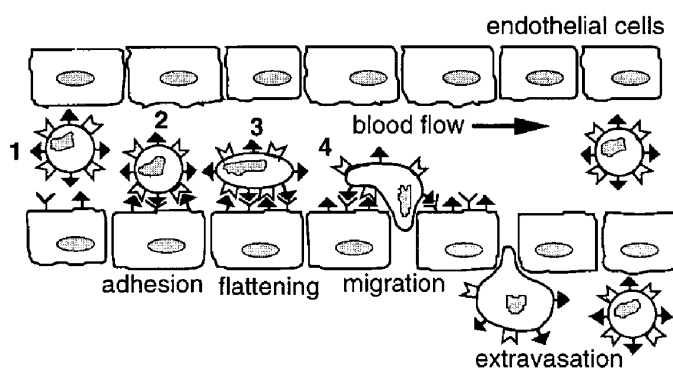


FIGURE 1.2: The leukocyte adhesion cascade, a fundamental biological process that is dependent on multivalent interactions [2]. (1) By expression and display of E- and P-selectins on the surface of nearby endothelial cells, a leukocyte is directed to a site of injury. (2) Through sLe^X, the leukocyte interacts polyvalently with the E- and P-selectins on the endothelial surface. In addition, sLe^X moieties on the endothelial cell interact with L-selectins expressed on the surface of the leukocyte. (3) The shape of the leukocyte flattens, which increases the number of interaction sites. (4) The migration of the leukocyte through the endothelial surface and extravasation into the injured tissue is mediated by integrins. Figure taken from [2], reprinted with kind permission of John Wiley and Sons.

Not only the immune system, but also its assailants make use of multivalency. The influenza virus, for instance, attaches to its host cells on the bronchial epithelial surface

via multiple copies of the hemagglutinin molecule at the same time (Figure 1.3). Hemagglutinin binds to sialic acid, a terminal sugar that is presented on many glycoproteins of the cell surface. As hemagglutinin itself is a trimer with three sialic acid binding sites, the valency of the interaction is increased further. Once a firm attachment between cell surface and virus is achieved, the virus is engulfed by a portion of cell membrane and thus enters the cell via endocytosis. Once inside the host cell, the virus induces a sequence of steps in order to release its contents, including the viral RNA genome, into the host cell's cytoplasm [2].

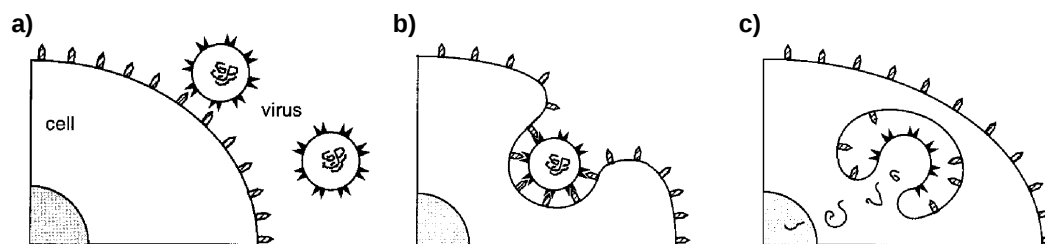


FIGURE 1.3: Schematic representation of influenza virus entering a cellular host [2]. a) The virus attaches to the cell surface through polyvalent interactions of viral hemagglutinin and cellular sialic acid sugars. b) The virus is engulfed and endocytosed by the cell. c) After endocytosis, acidification triggers the fusion of the viral envelope with the endosomal membrane, upon which the viral RNA genome and replication machinery is released into the cytoplasm. Figure taken from [2], reprinted with kind permission of John Wiley and Sons.

The latter example shows that multivalency in nature is not only a means to enhance the binding potency of individually weak interactions, but is also used in order to preorganize and control geometries on length scales that surpass the size and range of individual ligands and receptors by several orders of magnitude. In this case, the multivalent array of hemagglutinin molecules on the virus surface promotes the formation of a very large molecular architecture, namely the endosome that will engulf the virus and transport it into the cell [1].

1.2 Synthetic multivalency

The purposeful use of multivalent chemical compounds for different applications suggests itself. Directly related to the example of the importance of multivalency for viral infection strategies, Whitesides and coworkers conducted a comparison between mono- and multivalent inhibition of influenza virus [3, 4]. The authors presented a polymeric inhibitor containing 20 % sialic acid on a polyacrylamide backbone with a very low inhibition constant of only 600 pM, compared to an inhibition constant of 2 mM for its monomeric counterpart.

In 2000, Kitov *et al.* developed a pentavalent, water-soluble carbohydrate ligand for the inhibition of Shiga-like toxin [5]. Shiga toxins circumvent their low binding affinity by binding simultaneously to five or more cell-surface carbohydrates. *In vitro*, the pentameric inhibitor named STARFISH reached a subnanomolar inhibitory activity with an IC_{50} value of 0.24 nM, a ten million-fold increase compared to the monovalent ligand ($IC_{50} = 0.21$ mM).

More recently, Schwefel *et al.* investigated the binding of a tetravalent neoglycopeptide ligand to the plant protein wheat germ agglutinin (WGA) [6]. For this compound, consisting of four *N*-acetylglucosamine (GlcNAc) moieties tethered to a cyclic peptide backbone at distances tailored in order to fit the binding sites of the protein, they reported an IC_{50} value of 0.9 μ M, which corresponds to a 6400-fold increase in inhibitory activity per GlcNAc unit.

The above examples show that bridging adjacent protein binding sites by multivalent scaffolds can transform low-affinity ligands into potent inhibitors – a finding that opens up new perspectives (and challenges¹) for drug design. In a way, the rise of therapeutic monoclonal antibodies (“biologics”) in the recent years may also be a vehicle for transferring the concept of multivalency into everyday pharmaceutical applications [7].

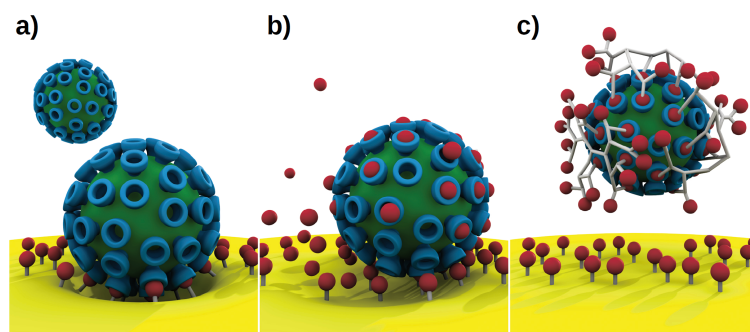


FIGURE 1.4: Schematic representation of a) a virus binding to a cell surface prior to infection, in the presence of b) a conventional monovalent inhibitor and c) a multivalent inhibitor using a dendritic scaffold. By interacting with multiple viral receptors simultaneously, the multivalent compounds form a screening around the viral surface that forestalls binding to the cell surface. Figure taken from [1], reprinted with kind permission of John Wiley and Sons.

Apart from the obvious applications in medicine and biochemistry, the concept of multivalency shows great promise for the structurally defined construction of new functional molecules in supramolecular chemistry [8] and in material sciences [9]. Multivalent molecular systems might serve as a starting point for the directed self-assembly of increasingly complex structures (e.g. functional “molecular machines”), or a chemical structuring of surfaces on the nanometer scale [10].

¹The challenges are, for example, to create biocompatible and, if necessary, cell-penetrating spacers and scaffolds

Current synthetic multivalent systems are not on par with the diversity and functionality of biological multivalent systems. Nonetheless, small synthetic multivalent host-guest systems that are being studied are providing helpful insights into the mechanisms of multivalent binding [11]. In contrast to biological systems, the systematic variation of synthetic structures is typically easier to realize, allowing for the study of very specific, structure-related effects based on larger series of compounds. Furthermore, the quantitative analysis of synthetic systems is facilitated by the fact that the number of interacting sites is controllable. Last but not least, the spacer and scaffold structures that are used to connect, present and preorganize the binding moieties can be varied and modified [1], which leads to one of the main topics of this thesis, namely the design of successful architectures for the multivalent presentation of ligands.

1.3 Architecture and design of multivalent compounds

Typically, a monovalent ligand is reducible to structural elements that are directly involved in binding to its receptor binding site, be it a highly specific enzymatic binding site, or a rather promiscuous, non-enzymatic binding site on a cell surface receptor. The design of multivalent ligands, by contrast, involves that multiple binding moieties (or “ligands” in the original sense) are tethered to a molecular spacer or scaffold. The task of the spacer is to connect the binding moieties and to present them at more or less defined distances. A scaffold serves the same purpose as a spacer, but offers additional levels of organization or architecture, based on its structural properties (Figure 1.5).

The linkage between binding moiety and spacer or scaffold, in turn, is denoted as linker.² While in the most basic case, the linker consists of a single chemical group that remains from the linkage reaction, it can also have a more elaborate structure that is meant to serve a certain purpose, such as enforcing a certain distance or angle of presentation between spacer and binding moiety. Obviously, unlike the chemically or enzymatically labile linkers used for applications in drug delivery, the linkage between spacer and tethered binding moiety is required to be stable under the given conditions [12, 13].

The rational design of multivalent compounds has to encompass the choice of binding moiety (known monovalent binders are an obvious choice), the choice of spacer tethering position and linkage group (often dictated by what is feasible from the perspective of synthetic chemistry), and the choice of spacer or scaffold structure. The latter two are largely responsible for how the binding moieties are presented to the receptor, and thus represent a key factor for the realization of multivalent binding. Depending on their

²The nomenclature in the literature is somewhat misleading, as linear spacers are frequently denoted as linkers, too. In the course of this thesis, the term linker will refer only to the linkage between binding moiety and a higher unit of organization, such as the spacer.

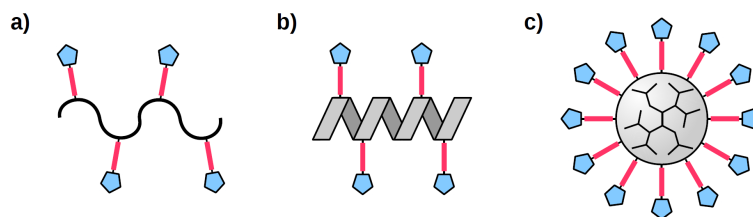


FIGURE 1.5: Schematic of different types of spacers and scaffolds for presenting multiple ligands: a) Linear polymeric spacer (flexible), b) Linear polymeric spacer/scaffold (rigid), and c) spherical dendritic polymer scaffold (nanoparticle-like). Ligand moieties are shown in blue, linkers in red.

structural properties, different parameters of ligand presentation can be adjusted. A flexible chain-like spacer will restrict the spacing of binding moieties to a certain distance range. A rigid nucleic acid spacer will also restrict the angle of presentation. If, for instance, the binding moieties are presented on a spherical dendritic polymer such as hyperbranched polyglycerol (hPG) [14, 15], the size of the polymer will determine the lateral density of presented binding moieties and the curvature of the binding interface. The adjustment of ligand presentation parameters opens up a wide range of possibilities of optimizing and tailoring the spacer or scaffold system with regard to its target receptor(s), given that the corresponding structural information is available. In fact, one can find many examples in the literature where even subtle changes in the spacer or scaffold structure were connected to very distinct changes in binding affinity [6, 16, 17].

Of course, the degree of preorganization that can be achieved also has to be related to the scale and valency of the interacting entities. If the goal is to create a high-affinity bivalent ligand for a dimeric protein such as estrogen receptor, preorganization and spacer design on the Ångström scale is meaningful (and probably necessary, as will be discussed in Chapters 4 and 5). If, however, the target is a undefined number of mobile receptors on the cell surface, or a multitude of interaction sites on the glycocalix, preorganization has to resort to a coarser level of architecture, e.g. the choice of shape and curvature of the functionalized nanoparticle to be used, and/or the density of ligands to be presented.

A lack of structural knowledge can be compensated either by systematic screening (e.g. different spacer lengths, or different polymeric scaffolds) or by allowing for increased variability in the presentation of the binding moieties. The latter can be implemented by using very flexible spacer or scaffold structures, which however increases the entropic penalty of binding (to be discussed in Chapter 2) and possibly evokes certain side effects that can be difficult to control (to be discussed in Chapter 4). Interestingly, multivalent architectures can also create new possibilities of gaining structural information about a given receptor. Chapter 5 will elaborate on an approach devised by Seitz and coworkers [18], where linear nucleic acid scaffolds equipped with multiple copies of a ligand are

used for screening the distance between the binding sites of a dimeric protein. Results obtained by this comparatively inexpensive method of screening can be used as a basis for further optimization, even in the absence of a crystal structure.

Using methods of molecular simulation, this thesis will make an effort to (i) contribute to the process of designing spacers and scaffolds for multivalent chemical systems and (ii) investigate the mechanisms of multivalent enhancement, possibly by finding traces of a cooperative effect within the binding process.

The first scientific goal, (i), is addressed in Chapters 4 and 5, respectively, where the results of a number of studies on flexible and rigid spacer and scaffold structures are presented. These studies, conducted in close cooperation with experimental scientists, attempt to enhance the understanding of the dynamics of spacer and scaffold structures and their interactions with ligand moieties as well as the receptor by means of conformational analysis. The computational results are related to “wet” experimental measurements wherever possible. It will become clear that spacer and scaffold structures represent “more than innocent spectators” [1] in the process of multivalent binding. Based on the findings, certain guidelines for the design of spacers and scaffolds for the multivalent presentation of ligands are suggested.

Chapter 6, in turn, will focus on (ii), the more general mechanisms of multivalent binding by analyzing the molecular simulation of a bivalent host-guest binding process in explicit solvent – to the best knowledge of the author, the first one that is being reported. A large part of Chapter 6 will deal with the technical difficulties that are connected with the sampling of binding processes by means of molecular simulation. The ZIBgridfree algorithm, reimplemented, enhanced and validated in the course of this thesis, is introduced as a tool for tackling some of these difficulties. In summary, Chapters 4–6 represent the “applied” part of this thesis.

The applied part of this thesis is complemented by a theoretical part, constituted by Chapters 2 and 3. First, Chapter 2 will provide the theoretical background on multivalent interactions – i.e. the “chemical theory” of this thesis – addressing the theoretical concepts that are available for describing the subtle effects that are at work in multivalent ligand-receptor systems, followed by Chapter 3, providing a theoretical background on molecular simulation, with a focus on statistical ensembles and a selection of algorithms necessary for mimicking experimental conditions in simulations.

The next section will feature a very brief introduction on the field of molecular simulation, also addressing the questions if molecular simulation is able contribute to the study of multivalent systems in the first place, as well as how and where it might complement the use of “wet” experimental methods in a meaningful manner.

1.4 Molecular simulation of multivalent systems

A new discipline between theory and experiment

According to Schmidt, Weber, Noé *et al.* [19], the aim of molecular simulation is the calculation of macroscopic quantities, or observables, in order to interpret or predict the outcome of experiments. The observables can be divided into static (equilibrium) and dynamic (non-equilibrium) properties. The static observables encompass structural quantities (e.g. structural parameters of proteins) and thermodynamic quantities (e.g. interactions energies, surface tension, or phase equilibria). The dynamic, non-equilibrium properties range from classic observables of fluids such as heat transmission, diffusion or viscosity to highly elaborate problems such as the prediction of protein folding, a notorious challenge for science in the 21st century.

The massive use of computers in quantum chemistry began as early as the 1960s, and shaped the field of theoretical – or computational – chemistry. The application of computational methods to biochemical systems, became possible only with the advance of microprocessor technology that was brought about by the 1970s and 1980s, and led to the emergence of computational biology and molecular simulation, the latter in particular directed at the challenge of computational drug design by means of molecular modeling [19]. While the most daring promise of molecular modeling, the purely *in silico* design of novel drug molecules, could not be fulfilled³, the field has since then brought forth many valuable tools for the study and design of molecular systems, in particular when combined with the power of structure determination, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The emergence of bioinformatics in the last decade has further facilitated the use of computational resources in the life sciences, in particular by creating and crosslinking comprehensive genomic, (bio)chemical and structure-based databases. The RCSB Protein Data Bank is a prime example for this development [21]. In summary, it is safe to state that molecular simulation has become an established tool, and has gained some renown even by non-theoretical scientists.

Today, molecular simulation, favored by further advancements in single-processor performance and the wide availability of parallel computing architectures, is applied to increasingly large and complex systems. One notable example in this respect is the 50 ns all-atom molecular dynamics (MD) simulation of satellite tobacco mosaic virus (STMV) in explicitly modeled solvent, conducted by Schulten and coworkers [22]. At the same time, recent developments such as cloud computing and the adaption of MD simulation code for use on highly parallel (and yet inexpensive and energy-efficient) graphics

³It can be argued that this is not in particular due to imperfect molecular modeling methods, but more so caused by a lack of understanding of the principles that govern ligand-receptor binding. The static “lock and key”-principle, for example, has been proven as a too naive view on this problem [20].

processing units (GPUs) are raising the appeal of molecular simulation for smaller research facilities, where large computing clusters are usually not maintainable: While the cost of computer hardware is ever decreasing, the cost of running and maintaining a large computing cluster is becoming increasingly incalculable due to rising energy costs. Despite the fact that technical innovations and theoretical progress (e.g. novel efficient algorithms for calculating non-bonded atomic interactions, cp. Chapter 3.1.1) have increased the applicability of molecular simulation to relevant systems (“relevant” often is equivalent to “complex”), the field is still facing certain limitations. The most severe limitation of molecular simulation is the so-called sampling problem.

All observables in molecular simulation are calculated as the expected values of a representative statistical ensemble⁴. In order to obtain meaningful values for the desired observables, a thorough sampling of the underlying statistical ensemble is necessary, a sampling that has to be obtained by stepwise numerical propagation of the dynamics of the system. The dynamics of molecular systems, in turn, is governed by small but very fast processes, such as the oscillations of chemical bonds and torsion angles. The timescale of these processes is of femtoseconds ($1 \text{ fs} = 10^{-15} \text{ s}$), and they have to be propagated correctly in order to describe the molecular processes on any other timescale (the computational cost of calculating even a single one femtosecond time step of MD can be significantly high). Many relevant molecular processes, however, only begin to be observable in the upper nanosecond regime, and more often take place in the timescale of microseconds, and above. Consequently, the sampling problem is rooted in the multiscale nature of molecular systems [19].

Currently, the temporal limits for MD simulations of “large” (biological) systems are in the microsecond range. More simulation performance can be “bought” by using implicit solvent models or coarse-grained approaches⁵, at the expense of precision and validity of the simulation outcome. Simulations including explicitly modeled solvent, in turn, require significant computing power, and generate large amounts of (highly redundant) data that is tedious to handle and to analyze. In summary, molecular simulation is still a developing field that requires further innovations in order to live up to the expectations of experimental scientists with regard to its predictive capability.

Application to multivalent systems

In order to study the properties of multivalent systems, a variety of “wet” experimental methods is available (Table 1.1). The measurements cover thermodynamic quantities

⁴The notion of statistical ensembles is elaborated in chapter 3.1.2.

⁵In contrast to all-atom force fields, e.g. Amber [23], coarse-grained (CG) force fields such as MARTINI [24] merge groups of several atoms in terms of larger “beads” in order to reduce the overall number of particles. The approach is successful in particular for large and hydrophobic molecules such as lipids.

such as equilibrium constants and binding free energies (including separate terms for the change of enthalpy and the change of entropy), structural parameters such as aggregate size and shape, and kinetic parameters such as rate constants. In addition to purely observational studies, it is possible to directly interact with multivalent systems by methods such as atomic force microscopy (AFM), and thus to derive certain quantities, e.g. the binding force of a multivalent compound to a functionalized surface [1].

The availability of various methods for deriving quantitative information about multivalent interactions directly from experiments gives rise to the question in how far the use of computational methods, and in particular molecular simulation (given its limitations), are meaningful for studying multivalency.

Method	Measured quantity	Measure for multivalency
surface plasmon resonance (SPR) spectroscopy	refraction index changes upon ligand-receptor interaction (mass-dependent)	association constant IC ₅₀ kinetics
NMR spectroscopy	diffusion coefficient from DOSY measurements NMR integrals or shifts as a function of ligand-receptor ratio or NMR titration	size information for detecting multivalently induced complex formation free energy of binding temperature-dependent measurements: binding enthalpy and entropy activation parameters ΔG^\ddagger , ΔH^\ddagger , ΔS^\ddagger (e.g. by line-shape analysis)
isothermal titration calorimetry (ITC)	heat as a function of the ligand-receptor or host-guest ratio	association constant binding enthalpy binding entropy free energy of binding
atomic force microscopy (AFM)	pulling force as a function of the intermolecular distance	binding force
transmission electron microscopy (TEM)	number of (multivalent) nanoparticles per aggregate visualization of individual bindings in a multivalent complex	degree of aggregation induced by multivalent functionalization stability against aggregation as a function of multivalent functionalization
high performance liquid chromatography (HPLC)	difference in retention as a function of polarity, molecular size or binding affinity	determination of equilibrium constants by competitive measurements

TABLE 1.1: A selection of “wet” measurement methods for quantifying multivalent interactions as compiled by Haag *et al.* [1]. For a complete list, including examples and references, refer to the original Ref. [1].

In fact, there are various reasons that justify the use of modeling and simulation approaches in this context: (i) Experimental methods typically yield macroscopic, averaged

values that give no immediate information about the microscopic processes in the system.⁶ Modeling of the microscopic processes, in turn, may help to test a hypothesis that was put forward and can facilitate the interpretation of experimental results. Moreover, the mere visualization of a given system on the molecular level may suffice to provide deeper insights and spark new ideas. (ii) Performing molecular simulation prior to experiments can help to narrow down the number of compounds that have to be tested, e.g. by filtering out candidates that are structurally inapt, saving both time and costs. (iii) By molecular simulation, it is possible to estimate quantities that are not directly derivable from current experimental methods, e.g. the conformational entropy change upon binding in different parts of a small multivalent ligand.

The latter point can be a double-edged sword, as the outcome of a simulation should in some form be verifiable by experiment. In summary, however, the benefits of studying multivalency by means of molecular simulation are evident, in particular as it is a phenomenon where subtle and evasive microscopic effects play an important role. This is elaborated in detail in the next chapter.

While molecular simulation of multivalent compounds is worthwhile, it is also challenging. This is mainly due to two reasons: Obviously, (i) multivalent systems are larger than comparable monovalent systems (if available), and (ii) current modeling and simulation methods are tailored for small, drug-like⁷ molecules. Drug-like molecules tend to have rather rigid structures with a high content of aromatic groups, a property that facilitates the use of computational high-throughput approaches such as molecular docking.

The first point, (i), refers foremost to the system size in terms of the overall number of particles. The atomistic simulation of, for example, polyvalently functionalized nanoparticles binding to a cell surface is beyond current technical means. It also refers to the system size in terms of conformational space volume: Highly flexible structures such as polyethylene glycol (PEG) chains, used as linkers or spacers, are not large in terms of the overall number of particles, but have a voluminous conformational space that needs to be sampled properly in order to make meaningful predictions.

The latter point, (ii), has an impact on the choice of computational tools that are applicable for studying multivalent interactions. Traditionally, ligand-receptor interactions are evaluated in a rather “static” way in order to limit the computational cost, e.g. by calculating and comparing single-point (e.g. molecular docking [26, 27]) or short trajectory (e.g. the linear interaction model [28]) interactions energies. The special properties of multivalent ligand-receptor interactions, however, are rooted in “dynamic” phenomena (to be discussed in Chapter 2) that can only be captured by studying larger ensembles

⁶An example is the measurement of “off” rate constants in kinetics. Experiments will only yield the averaged “off” rate constant for the whole complex, which gives little to none information about how the microscopic kinetics of a multivalently bound complex works.

⁷According to Lipinski’s Rule of Five [25], a rule of thumb for “druglikeness”, the majority of drug-like molecules is small, with a molecular weight of less than 500 u, an equivalent of about 20 to 70 atoms.

of states. In other words, it is necessary to move from a static, state-based view of ligand-receptor binding to a more dynamic perspective that focuses on processes – the topic of Chapter 6.

What remains is to find strategies that allow for the sampling of these large ensembles. In the course of this thesis, different approaches are taken. The first one is long-time MD simulation, which basically is the attempt to overpower the sampling problem by using more computing resources. This is expensive in terms of resources, but a proven way to capture molecular processes. The second approach is coarse-graining, where realism and detail with regard to the molecular system under investigation is sacrificed in order to access larger timescales. Recently, Numata *et al.* [29] reported how coarse-grained simulations of PEG spacers in a bivalent ligand-receptor system can be used to estimate the effective concentration of ligand, depending on the topology of the receptor. Another recent example is the study of rebinding effects in multivalent systems by Weber *et al.* [30], relying in parts on Monte Carlo simulations of a highly simplified bivalent ligand-receptor system. Finally, the third approach is importance sampling, and other computational methods that aim at lessening the problems associated with large systems by using enhanced sampling algorithms [31]. As an example for this class of methods, this thesis will present the ZIBgridfree sampling algorithm (Chapter 3.2) and its application to a small multivalent system (Chapter 6.3). An evaluation of the method in terms of sampling accuracy can be found in Appendix A.

Chapter 2

Multivalent interactions – a theoretical background

2.1 Terms and definitions

Following the nomenclature summarized by Haag *et al.* [1], interactions between an m -valent receptor and an n -valent ligand are considered to be multivalent ($m, n > 1$, also $m \neq n$). Interactions between multiple monovalent ligands ($n = 1$) with a multivalent receptor are not multivalent. In this case, however, the binding of the monovalent ligands is favored due to the symmetry effect [32–34], as an m -valent receptor will bind m times more monovalent ligands than a monovalent receptor with only one binding site, a fact that, depending on the ligand concentration, will also promote the occurrence of immediate rebinding after dissociation from the receptor. Both the symmetry effect as well as rebinding have to be taken into account to multivalent ligands as well. Furthermore, one can discriminate between homo- and hetero-multivalent ligands and receptors, which either have multiple identical binding sites (homo), or multiple non-identical binding sites (hetero) [1].

Multivalent interactions at interfaces (e.g. cell membranes or functionalized nanolayers) or on larger structures (e.g. viruses or nanoparticles) are often denoted as polyvalent interactions. These systems are characterized by the high valency of binding sites ($m, n \gg 10$) and large-scale (and predominantly lateral) interactions. Although the terms polyvalency and multivalency eventually describe the same functional concept, this differentiation in nomenclature can be helpful for providing a certain context. The focus of this thesis is laid on isolated, homo-multivalent interactions of low valency¹.

¹Atomistic simulation of upper nanoscale (and above) polyvalent systems is still out of reach for classical molecular simulation, even on current supercomputers (cp. Chapter 1.4).

Two more noteworthy terms related to the concept of multivalency are avidity (from the context of biochemistry and immunology) and chelation, or chelate effect (from the context of inorganic coordination chemistry).

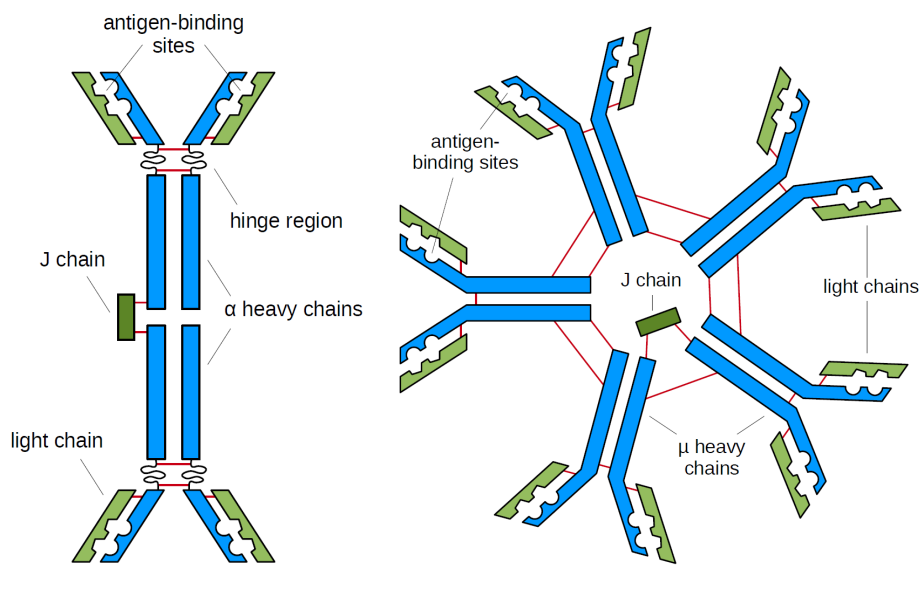


FIGURE 2.1: Schematized diagrams of dimeric IgA (left) and pentameric IgM (right). The subunits are held together by disulfide bonds (red). A single J (joining) chain is disulfide-bonded between the tails of two heavy chains. IgA is the principal antibody class in secretions such as saliva and tears. During the formation of the IgM pentamer, the addition of each successive four-chain IgM subunit requires a J chain, which, except for the final one, is discarded afterwards. Unlike other antibody classes, the IgM subunit does not have a flexible hinge region within its heavy chain, possibly an explanation for the comparatively low binding affinity of its antigen-binding sites [35].

The first term, avidity, was coined in order to describe the binding behavior of immunoglobulins (or antibodies) of different valency. The five antibody isotypes of mammals can be divided into bivalent (monomeric IgD, IgE and IgG), tetravalent (homodimeric IgA, see Figure 2.1, left) and decavalent (homopentameric IgM, see Figure 2.1, right) structural classes, which in turn differ in their biological properties, functional locations and ability to deal with different antigens [36].

In contrast to the term affinity, which in this context corresponds to the strength of a single antibody-antigen interaction, avidity denotes the combined synergistic strength of all antibody-antigen interactions involved. Given that the antigen has multiple binding sites, avidity is supposed to be more than the sum of the individual antibody-antigen binding affinities. The decavalent antibody IgM, for instance, the major antibody class secreted into the blood in the early stages of a primary antibody response, is known to have a lower affinity per antigen-binding site compared to its bivalent counterparts, antibodies IgD, IgE and IgG, but was shown to compensate for this drawback by its higher avidity – arguably related to its higher valency [37].

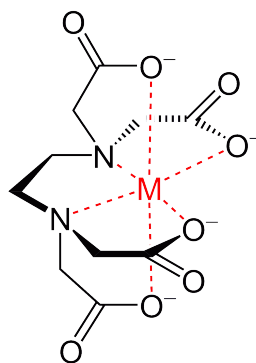


FIGURE 2.2: Chelate complex of metal and ethylenediaminetetraacetic acid (EDTA).

The second term, chelation, denotes the formation or presence of two or more separate coordinate bonds between a polydentate ligand (i.e. a ligand having more than one free electron pair) and a single central atom [38]. While the ligand, denoted as chelator, chelant or chelating agent, typically is an organic compound, the central atom is often a double positively charged metal ion (e.g. Fe_{2+} or Cu_{2+}). It is noteworthy that chelate complexes play an important role in nature: Prominent examples include heme, chlorophyll and vitamin B_{12} .

Chelate complexes are known to be more stable than similar monodentate, mutually untethered ligands. Traditionally, this “chelate effect” is attributed to two causes. First, the entropy loss upon complex formation is supposedly reduced for tethered ligands, and hence constitutes a thermodynamic advantage [39]. In the example given in Figure 2.2, the chelate complex is formed by two entities, whereas a similar complex formed by nonchelating ligands would consist of seven separate entities, each of which is subject to entropy loss upon complex formation. In addition, a chelate ligand is not able to withdraw from the central atom until all coordinate bonds are broken, so that the dissociation of chelator and metal ion is unfavorable. This renders the chelate complex very stable, and also increases the probability of immediate reassociation after bond breaking. More recent studies using isothermal titration calorimetry (ITC) and quantum chemical methods paint a more ambivalent picture of the origins of the chelate effect, and conjecture that it is either completely enthalpic in origin [40], or “a result of both enthalpy and entropy contributions that vary from one system to the other.” [41]

As will be discussed in Sections 2.2 and 2.3, many characteristics of chelate complexes are trademarks of multivalent systems in general. Although multivalent binding and chelate binding are sometimes used as synonyms, the latter is rather a special case of a more general concept. Consequently, the chelate effect, and also the concept of avidity, can be understood as examples of a more general multivalent enhancement effect. Remarkably,

this effect can be encountered in very heterogeneous systems that cover several magnitudes of length scale, ranging from sub nanoscale chelate complexes, involving small organic molecules and metal ions, to microscale lateral interactions between viruses and cell surfaces [1].

Given the somewhat conflicting results regarding the origin of the multivalent enhancement effect even for the smallest chelate complexes, it becomes clear that the quantitative theoretical description of multivalent chemical systems remains a challenge [1]. Enthalpic and entropic contributions (e.g. the reduction of entropy loss at complex formation), as well as the immediate reassociation after complex breaking (“rebinding”) have to be considered and put into a larger perspective. First, the following section will try to compile the thermodynamic aspects of multivalent binding.

2.2 Multivalent thermodynamics

As a first step, it is straightforward to look for the multivalent enhancement effect in the thermodynamics of multivalent binding. Once an interaction between a monovalent receptor and a single binding site takes place, a free energy change ΔG_{mono} will occur. Hence, for n independent monovalent ligands interacting successfully with n binding sites, the total change in free energy is $n \cdot \Delta G_{mono}$. Based on this simple observation, if a multivalent enhancement effect is at work, it must hold that for an n -valent ligand, $\Delta G_{multi} < n \cdot \Delta G_{mono}$. Using the van’t Hoff equation, it is possible to relate the Gibbs free energy of binding ΔG with the associated binding constant K ,

$$\Delta G = -RT \cdot \ln K \quad (2.1)$$

where $R = 8.3144621(75)$ J/mol K is the universal gas constant, and T is the absolute temperature. Consequently, for an interaction that benefits from a multivalent enhancement effect, it must hold that $K_{multi} > (K_{mono})^n$ [2]. Going back to the original definition of the Gibbs free energy of binding,

$$\Delta G = \Delta H - T\Delta S \quad (2.2)$$

it is straightforward to trace the multivalent effect in the enthalpic (ΔH) and the entropic term ($T\Delta S$) separately, which will be done in the following sections. Nonetheless, it shall be mentioned that recent research suggests that certain more intricate aspects of multivalent enhancement (such as an increase in the occurrence of rebinding events) are not clearly separable into enthalpic and entropic components and have to be treated differently in order to arrive at a proper thermodynamic formulation [30].

2.2.1 The role of enthalpy

The enthalpy change ΔH is equivalent to the total energy generated by a chemical or enzymatic reaction at constant temperature and pressure, and only depends on the initial (unbound) and the final (bound) state of the system under observation. As the enthalpy change is largely dependent on the nature of the ligand and the binding site of the receptor, one might assume that for an n -valent ligand, it holds that $\Delta H_{multi} = n \cdot \Delta H_{mono}$. This assumption, however, is unrealistic.

Whitesides *et al.* use a prototypical, idealized bivalent receptor to illustrate this problem, see Figure 2.3 [2]: In order to reach the theoretically possible bivalent binding enthalpy of $\Delta H_{bi} = 2 \cdot \Delta H_{mono}$, the connecting rigid spacer would have to provide perfect fit to the receptor, regarding both the spacing between the two binding moieties as well as the angle of binding moiety presentation. As soon as the spacer is less than perfect (e.g. the spacing between the binding moieties is too wide or too narrow), an enthalpic penalty due to strain in both ligand and receptor has to be added to the balance, i.e. $\Delta H_{bi} = 2 \cdot \Delta H_{mono} + \Delta H_{strain}$. Intuitively, if the mismatch between binding moiety presentation and receptor becomes too severe, bivalent binding will become impossible. In practice, designing a spacer structure with perfect fit is not feasible, therefore a certain enthalpic penalty for strain in multivalent ligand and receptor has to be anticipated.

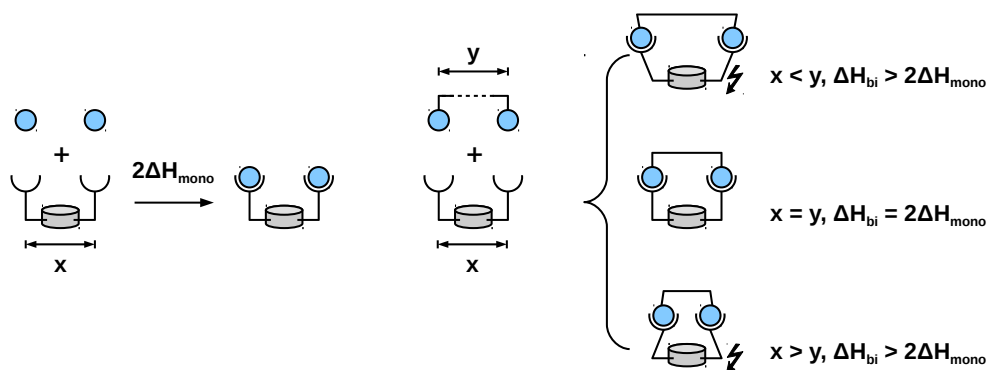


FIGURE 2.3: Simplified model of the enthalpic change upon binding for a bivalent ligand. For a bivalent receptor with a binding site distance x , the enthalpic change upon binding for two monovalent ligands is $\Delta H = 2 \cdot \Delta H_{mono}$. For a bivalent ligand with a spacer (or binding moiety spacing) of length y , $\Delta H_{bi} > 2 \cdot \Delta H_{mono}$ in cases where y does not match x , as strain energy in both bivalent ligand and receptor has to be added to the enthalpic balance [2].

Another enthalpic penalty might arise from the chemical modification of the binding moiety that is necessary in order to attach it to a spacer or linker. This is especially true in cases where the receptor binding site is very selective, and the binding moiety to be used for the multivalent ligand has already undergone several cycles of structural refinement regarding the receptor. In this situation, any modification of the original

monovalent binding moiety might lead to a less favorable enthalpy.

To be fair, one has to add that a spacer, scaffold or linker structure might be designed such that it is able contribute to a favorable enthalpic balance, i.e. $\Delta H_{bi} = 2 \cdot \Delta H_{mono} + \Delta H_{strain} + \Delta H_{spacer}$, with $\Delta H_{spacer} \leq -\Delta H_{strain}$. In this situation, the penalty ΔH_{strain} is compensated by favorable interactions between spacer/scaffold and/or the linker with the receptor surface, in sum denoted as ΔH_{spacer} . Apart from the fact that the rational design of such compounds represents a considerable challenge, it would also be hard to determine the contribution of ΔH_{spacer} by experiment.

Although not *per se* related to multivalency, the positive scenario $\Delta H_{multi} < n \cdot \Delta H_{mono}$ can be encountered as well. In these cases, the first ligand-receptor binding event at one of the binding sites of an m -valent receptor triggers a conformational change in adjacent, unoccupied receptor binding sites, which in turn leads to (increasingly) favorable enthalpy for the next $m - 1$ binding events. This effect is denoted as allosteric (“inter-site”) cooperativity and plays an important role in certain biological systems.

The best-studied biological example for allosteric cooperativity is the interaction between tetrameric hemoglobin and four oxygen molecules: Once a single hemoglobin monomer becomes oxygenated, a conformational change in the whole complex is initiated, causing the remaining three monomers to gain an increased affinity for oxygen [42]. This effect is important for the function of hemoglobin as the transporter of oxygen, as it enables the selective uptake of oxygen at high oxygen levels in the lung, and the targeted release of oxygen in tissues with a low level of oxygen. As a consequence, the oxygen binding curve of hemoglobin is sigmoidal, as opposed to the normal hyperbolic curve associated with noncooperative binding.

Another biological example for allosteric cooperativity is the binding of the oligosaccharide portion of the cell surface receptor ganglioside G_{M1} , called oligo- G_{M1} , to pentameric cholera toxin B5, which was studied by Schön and Freire using a calorimetric method [43]. They were able to show that the initial binding of a single oligo- G_{M1} molecule to one of the binding sites of B5 will lead to improved enthalpy for subsequent binding events at the adjacent binding sites.

It is important to note that this form of enthalpic cooperativity is not related to a multivalent enhancement effect, and that the above biological examples do not represent multivalent ligand-receptor interactions according to the original definition, see Section 2.1. Nonetheless, multivalent interactions are able to benefit from allosteric cooperativity, which makes it even more difficult to determine the “true” contribution of multivalent enhancement. The idea of cooperativity in multivalent interactions will be revisited in Section 2.4.

2.2.2 The role of entropy

As was shown in the previous section, there is not much room for improvement of ΔH_{multi} compared to $n \cdot \Delta H_{mono}$ in an n -valent interaction. Several factors, mainly related to the manner in which the binding moieties are presented to their receptor, may disfavor the enthalpy of the multivalent interaction. Therefore, whenever multivalent enhancement occurs ($\Delta G_{multi} < n \cdot \Delta G_{mono}$), it is traditionally attributed to the entropic term in Equation 2.2.

A short side note on the concept of entropy

The term ΔS states the difference of entropy between the unbound and the bound state. This term is often denoted a “entropy loss upon binding”, because a system is supposed to lose entropy when it undergoes the transition from the unbound state into the bound state. In this context, it is helpful to recall the general concept of entropy. Ludwig Boltzmann defined entropy as

$$S = k \log W. \quad (2.3)$$

This formula defines the macroscopic quantity entropy S in terms of the multiplicity W of the microscopic degrees of freedom of the system.

The concept of multiplicity can be illustrated with a simple lattice model for diffusion, taken from Ref. [44]: Given that one has four black particles and four white particles in a volume of eight lattice sites. There are four lattice sites on the left, and four lattice sites on the right, separated by a permeable wall. The overall volume is fixed, which means that all the lattice sites are occupied either by a black or a white particle. The degrees of freedom of the system are now defined in terms of black and white particles on each side of the permeable wall.

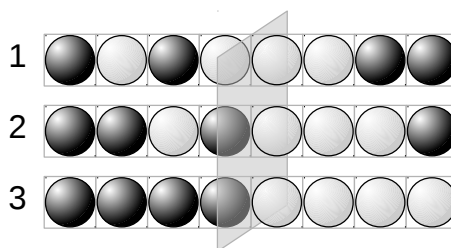


FIGURE 2.4: A simple lattice model of diffusion, adapted from Ref. [44]. The four lattice sites on the left and the four lattice sites on the right are separated by a permeable wall. The four black and the four white particles can adapt every configuration of occupying the lattice sites with equal probability. The overall volume is fixed.

Assuming that each spatial configuration (or sequence) is equally probable, the number of distinguishable arrangements of N particles in M sites is given by

$$W(N, M) = \frac{M!}{N!(M - N)!}$$

For each given value of left and right compositions, the total multiplicity is the product of the multiplicities for the left and the right side. With regard to the exemplary particle configurations shown in Figure 2.4, one can compute the following multiplicities:

1. $W = W(\text{left}) \cdot W(\text{right}) = \frac{4!}{2!2!} \frac{4!}{2!2!} = 36$
2. $W = W(\text{left}) \cdot W(\text{right}) = \frac{4!}{1!3!} \frac{4!}{3!1!} = 16$
3. $W = W(\text{left}) \cdot W(\text{right}) = \frac{4!}{0!4!} \frac{4!}{4!0!} = 1$

Systems tend to be found in the states with the highest multiplicity W , or, in other words, “a system will change its degrees of freedom to reach the microscopic arrangement with the maximum possible multiplicity” [44]. Hence, one can predict the most probable composition of black and white particles by maximizing the multiplicity of arrangements. The first case, which has the most uniform distribution of particles, has the highest multiplicity, and therefore is the most probable. The third case, which has the greatest particle segregation, is the least probable.

This leads directly to the Second Law of Thermodynamics, which states that isolated systems tend towards their states of maximum entropy. With that in mind, it is not meaningful to define entropy as a measure for the “disorder” of a system, because the notion about what is orderly and what is not is often a matter of personal preference, while the term multiplicity is well-defined.

For use in thermodynamics, k in the original definition of entropy in Equation 2.3 is the quantity of Boltzmann’s constant $k_B = 1.3806488(13) \cdot 10^{-23}$ J/K, which puts entropy into units that can be interconverted with energy. k_B leads to the entropy per particle. Entropy, however, is a concept that applies to any type of probability distribution function. For other types of probability distributions, k is chosen to suit the current purpose, e.g. $k = 1$. Finally, a simple justification for the use of the logarithm function (more precisely the natural logarithm $\log_e = \ln$) in Equation 2.3 is the need for compatibility with thermodynamics, which requires entropies to be *extensive*.

Given a system with two subsystems A and B and multiplicities W_A and W_B ($W_{total} = W_A W_B$), this means that the total entropy of the system is the sum of the subsystem entropies: $S_{total} = S_A + S_B$. This property is easily satisfied when the multiplicity of

states is counted in a logarithmic way, i.e. if $S_A = k \ln W_A$ and $S_B = k \ln W_B$, then $S_{total} = k \ln W_A W_B = k \ln W_A + k \ln W_B = S_A + S_B$ [44].

Whitesides' model of entropically enhanced bivalent binding

Considering the above, it now becomes clear that any kind of binding of a ligand to a receptor results in a loss of entropy, as the multiplicity of states in the bound state is (inevitably) less than in the unbound state. To be more precise, the entropic term from Equation 2.2 can be split up into four components, namely translational and rotational entropy change ΔS_{trans} and ΔS_{rot} , conformational entropy change ΔS_{conf} , and solvation-associated entropy change ΔS_{sol} (cp. Equation 2.4) [2].

$$\Delta S = \Delta S_{trans} + \Delta S_{rot} + \Delta S_{conf} + \Delta S_{sol} \quad (2.4)$$

In the unbound state, ligand and receptor are free to translocate in three dimensions and to rotate around their three principal axes. The same holds for attaining conformational changes. The movement and dynamics of the unbound entities is only hindered by the viscosity (and the resulting friction) of the solvent molecules. The solvent molecules, in turn, are free to form a more or less defined shell around the solute.

In the bound state, the individual translational and rotational freedom of the ligand is lost, conformational changes of both ligand and receptor are constrained, and the free arrangement of solvent molecules around the solute molecules is limited - in particular with regard to the ligand, which may be buried deep in narrow protein binding site, with only small room for solvent molecules left.

While this phenomenon *per se* is inevitable, any reduction of the entropy loss upon binding will lead to a (more or less) significant improvement of ΔG (cp. Equation 2.2). The rationale behind improving ΔG by means of multivalency is that a preorganization of the ligands in terms of multivalent arrays leads to a reduction of the entropy loss upon binding. This form of preorganization is equivalent to reducing the multiplicity of microstates in the unbound state.

The schematic depicted in Figure 2.5 summarizes this notion for different situations in a simplified manner, and is mainly based on Whitesides *et al.* [2]. For a monovalent ligand and a monovalent receptor (case a), the entropy loss upon binding is $\Delta S_{mono} = \Delta S_{trans} + \Delta S_{rot}$. Intuitively, for two monovalent ligands and two receptor binding sites (case b), the entropy loss doubles, as each ligand requires one equivalent of ΔS_{trans} and ΔS_{rot} . Using a bivalent ligand with an ideal (and completely rigid) spacer (case c), the entropy cost can again be reduced to approximately one equivalent of ΔS_{trans} and ΔS_{rot} , as the translational and rotational degrees of freedom of the two ligands are

joined together:

$$\Delta S_{bi/rigid} = \Delta S_{trans} + \Delta S_{rot}$$

Finally, when using a bivalent ligand with a flexible spacer (case d), the conformational entropy loss of the spacer has to be added to the balance. The quantity ΔS_{conf} is dependent on the structural parameters of the spacer, and can, in theory, become arbitrarily large. This implies that the – theoretically optimal – advantage of saving one equivalent of ΔS_{trans} and ΔS_{rot} each is threatened to be nullified when an inappropriately long and flexible spacer is used.

For these merely qualitative summations, the intrinsic conformational entropy change of the ligands has been omitted, as it is supposedly invariant with regard to the valency of the system. Hence, ΔS_{conf} in Figure 2.5 only refers to the conformational entropy change in the spacer. The same holds for the solvent-associated entropy change ΔS_{sol} . Any effect of the spacer on ΔS_{sol} (e.g. by interacting with the surface of the receptor) is neglected for the sake of generality.

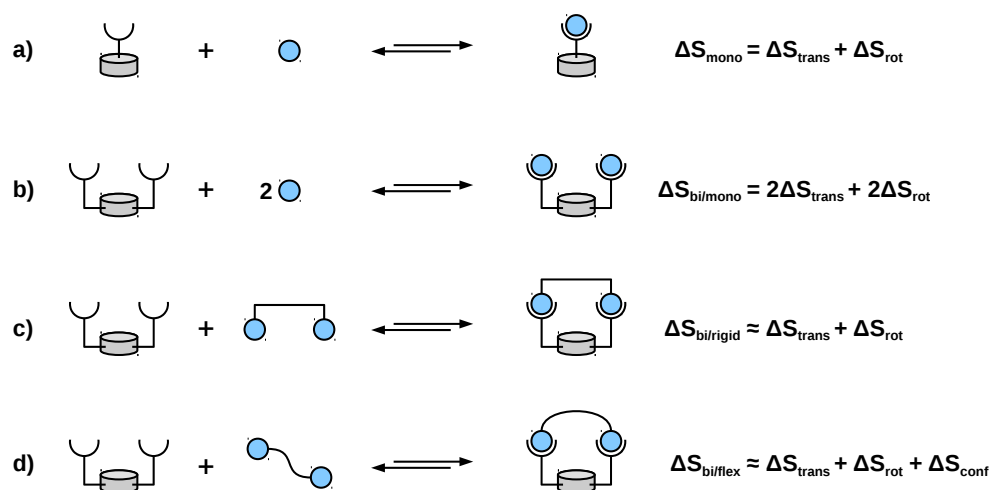


FIGURE 2.5: Simplified model of the entropy loss upon binding for a) one monovalent ligand and a monovalent receptor, b) two monovalent ligands and a bivalent receptor, c) a bivalent ligand, rigid spacer with a bivalent receptor and d) a bivalent ligand, flexible spacer with a bivalent receptor [2].

More important, the simplified scheme shown above is only valid if the quantities ΔS_{trans} and ΔS_{rot} are approximately equal for mono- and bivalent scenario.

According to the Sackur-Tetrode equation for calculating the absolute entropy of an ideal (monoatomic) gas, S_{trans} is related to the logarithm of the molecular mass m ($S_{trans} \propto \ln(m)$), and inversely to the logarithm of its concentration ($S_{trans} \propto \ln([L])^{-1}$). Similarly, S_{rot} is related to the logarithm of the product of the molecule's three principal moments of inertia, I_x , I_y and I_z ($S_{rot} \propto \ln(I_x I_y I_z)$) [2], which in turn are dependent on its shape, mass distribution and axis of gyration. Whitesides *et al.* argue that the “weak” logarithmic dependence of ΔS_{trans} and ΔS_{rot} on the mass of the dimensions

of the molecule allow for this approximation, provided the molecules under comparison are given in the same concentration. Still, assuming equivalence for mono- and bivalent ligands with regard to ΔS_{trans} and ΔS_{rot} is a very optimistic scenario, and probably should be conceived only as a theoretical upper limit for the entropy loss that can be saved by moving from mono- to bivalent ligand.

Turning again to Figure 2.5, one can now focus on the comparison of the two bivalent scenarios, case c) and case d). In case c), the two ligands are tethered by a completely rigid spacer (no torsional rotation around bonds is possible) that perfectly matches the spacing of the two receptor binding sites. The initial binding event of the first ligand to the first receptor binding site costs approximately $\Delta S_{trans} + \Delta S_{rot}$ in terms of entropy loss. As this hypothetical bivalent ligand only has a single conformation, the subsequent binding event of the second ligand to the second receptor binding site can occur without additional entropy cost. Given that no additional enthalpic cost accumulates, the free energy change for the second (intramolecular) binding event is entropically enhanced: $\Delta H \approx \Delta H_{mono}$ and $\Delta S \approx 0$, therefore $\Delta G \approx \Delta H_{mono}$.

This hypothetical scenario suffers from the fact that in reality, a completely rigid spacer cannot be implemented. All chemical linking groups introduce a certain amount of flexibility to the structure, and in many cases, structural flexibility is even wanted in order to compensate for incomplete knowledge of the target receptor’s shape and dimensions (the receptor, in turn, might also have flexibly spaced binding sites). Last but not least, structural flexibility may be required for maintaining an acceptable degree of solubility. A more realistic scenario is depicted in Figure 2.5 d). Here, the ligands are tethered by a spacer that undergoes a conformational entropy change ΔS_{conf} upon binding to the receptor. Typically, for any kind of complex formation, the conformational entropy change is bound to be unfavorable ($\Delta S_{conf} < 0$), as the multiplicity of available conformational states in the complexed state is smaller than before complexation. Intuitively, a long, chain-like spacer with great structural flexibility, as indicated in the schematic of case d), is likely to lose more conformational entropy upon binding to the receptor as a more rigid counterpart.

The quantity of ΔS_{conf} is important, as it ultimately decides over the fate of the bivalent interaction – at least within the scope of this limited model. In cases where $\Delta S_{conf} > \Delta S_{trans} + \Delta S_{rot}$, the bivalent ligand is entropically enhanced, which means that bivalent binding is favored over two monovalent binding events (cp. Figure 2.6, path A). If $\Delta S_{conf} < \Delta S_{trans} + \Delta S_{rot}$, bivalent binding is entropically diminished, and only monovalent binding events will take place (cp. Figure 2.6, path B). Finally, when $\Delta S_{conf} \approx \Delta S_{trans} + \Delta S_{rot}$, the bivalent ligand is entropically neutral, which means that both bivalent and monovalent binding of two separate entities will occur [2].

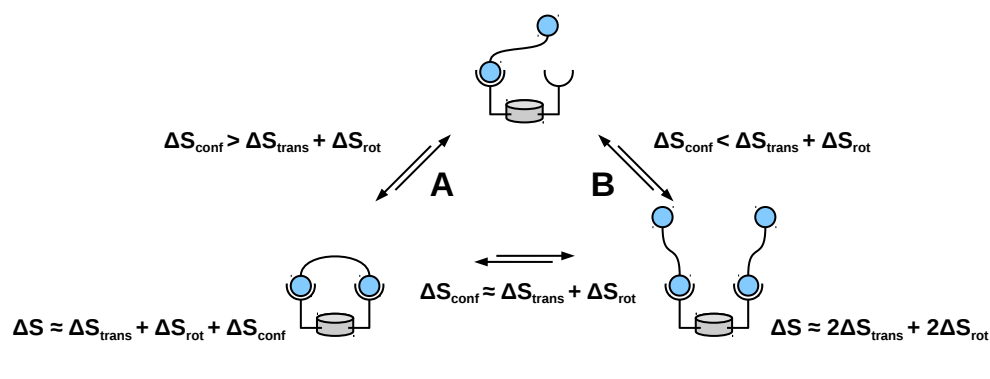


FIGURE 2.6: Entropically enhanced binding (only path A), entropically diminished binding (only path B) and entropically neutral binding (bottom, paths A and B). In the latter case, the entropic cost of bivalent binding and monovalent binding of two separate entities is approximately equal, so that both scenarios will occur with the same likelihood [2].

Whitesides *et al.* give an estimate of the entropic cost of freezing a single, rotating carbon-carbon bond at a temperature of 298 K at 0.5 kcal/mol, and 0.1 - 1.0 kcal/mol for other types of rotating bonds [45, 46]. According to this estimate, the upper bound for the conformational entropy loss of an triethylene glycol spacer is 10 kcal/mol, assuming that all rotating bonds are frozen upon complexation [2]. Although a flexible spacer is likely to retain some of its conformational flexibility even after complexation (which makes the factual entropy loss less severe), the order of this number indicates that ΔS_{conf} may become so large that a multivalent interaction will be rendered impossible due to entropic reasons, given the spacer is both long and very flexible.

Enthalpy-entropy compensation

Another aspect connected to quantifying the entropy loss upon binding is the phenomenon of enthalpy-entropy compensation [47], which refers to a seemingly linear relationship between ΔH and ΔS in many (and originally only biological) processes: a gain in enthalpy will be compensated by a loss in entropy, and vice versa. By now, enthalpy-entropy compensation has also been studied in the context of supramolecular chemistry (host-guest interactions) [48], as well as in biomolecular (small molecule/drug-receptor) [49, 50] and multivalent interactions [51]. It has to be mentioned that enthalpy-entropy compensation suffers from a somewhat dubious reputation, as the linear relation between ΔH and ΔS can also arise as a statistical artifact caused by interpolating temperature-dependent data that covers an insufficient range of measurements. However, whenever enthalpy-entropy compensation is derived from calorimetric measurements, “it cannot be easily dismissed” [52].

A simple justification for enthalpy-entropy compensation is given by the following, very

intuitive line of thought: In order to maximize the enthalpic gain (i.e. $\Delta H \ll 0$), the interacting species (e.g. a small molecule ligand and a macromolecular biological receptor) have to form a tight and stable interaction. This corresponds to giving up conformational multiplicity, and has an impact to both interaction partners. Even if the small ligand molecule is very rigid (e.g. a steroid compound), it may induce a notable loss of conformational multiplicity in the receptor binding site. This corresponds to a loss in conformational entropy, i.e. $\Delta S_{conf} \ll 0$. Accordingly, a less tight interaction with only a moderate gain in enthalpy will allow for more conformational multiplicity, so that the conformational entropy loss is less severe.

The impact of this effect is dependent on the structural properties of the system in question. A very interesting example from the literature is the bivalent binding of immunoreceptor tyrosine-based activation motif (ITAM) to the Syk tandem Src homology domain (Syk tSH2). The tSH2 domain of Syk consists of two SH2 domains that are connected by a flexible interdomain. While the SH2 domains form defined, compact structures, the interdomain is subject to notable conformational variability, which, according to Fütterer *et al.* [53], provides the structural basis for “a surprising flexibility in the relative orientation of the two SH2 domains.” Figure 2.7 shows the crystal structure of the tSH2 domain of Syk bound to dually tyrosine-phosphorylated ITAM.

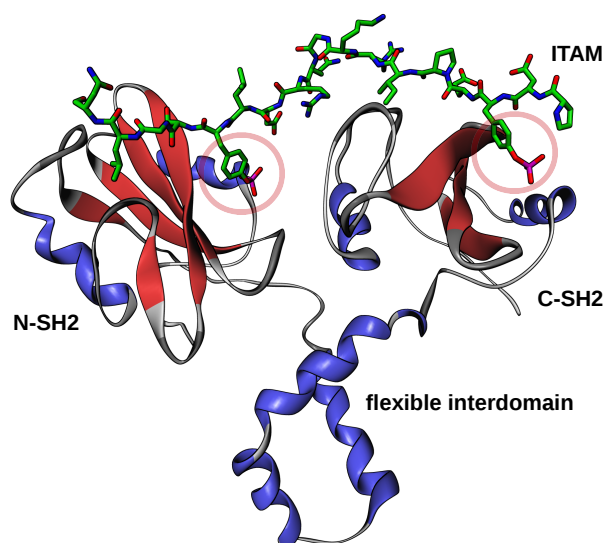


FIGURE 2.7: Dually tyrosine-phosphorylated ITAM bound to the tSH2 domain of Syk (secondary structure representation). Two phosphotyrosine side chains (highlighted) form the central element of the tSH2 domain recognizing binding motif of ITAM (PDB ID 1A81 [53]).

In their work from 2005, de Mol *et al.* [51] compare two different ligands for the tSH2 domain: **1**, a flexible peptide derived from the tSH2 recognizing ITAM sequence, and **2**, a ligand in which the amino acids between the two SH2 binding motifs in ligand **1** have

been replaced by a rigid linker of comparable length. Following the line of thought from the previous sections, the rigidified ligand should have a higher affinity for the tSH2 domain due to an advantage with regard to ΔS_{conf} . The authors found, however, that the expected entropy advantage is not realized. On the contrary, ligand **2** binds with a considerably higher entropy price of approximately -9 kcal/mol, which is attributed to a further decrease in protein flexibility upon binding. Considering the structure of Syk tSH2, it is very likely that this mainly applies to the flexible linker that connects the two SH2 domains (cp. Figure 2.7). Consequently, de Mol *et al.* [51] conclude that “the entropy price for fixing the flexible protein counteracts the entropy gain of rigidization of the ligand.” Interestingly, the rigidified ligand brought about an enthalpic enhancement in the same order of magnitude, so that both ligands bound to the Syk tSH2 domain with comparable affinity.

This example shows that enthalpy-compensation adds another unknown variable to the design of multivalent compounds. It shows that rigidization of spacer and scaffold structures can have adverse effects, in particular when the receptor is prone to losing conformational multiplicity upon binding. Similarly, using very flexible structures can be a double-edged sword, as the flexibility that allows all ligand-receptor interactions to occur without energetic strain is the same flexibility that may render a multivalent interaction unfavorable due to entropic loss.

Furthermore, depending on the nature of the system, the loss in (conformational) entropy can be “notoriously difficult to quantitate” [2] with experimental methods such as isothermal titration calorimetry (ITC). In this respect, the use of structure-based modeling and simulation methods in a predictive way may offer relief, as will be discussed in Chapter 3.3.

Entropic cost of the first binding event

The final aspect with regard to the role of entropy in multivalent interactions to be discussed here is the entropic cost of the first (or initial) binding event. Whitesides’ model suggests that a multivalent ligand benefits from the fact that following the first binding event – i.e. after the first interaction site of the multivalent ligand has bound to the first unoccupied interaction site of the multimeric receptor – subsequent ligand-receptor binding events can take place without additional (or at least significantly reduced) cost with regard to ΔS_{trans} and ΔS_{rot} . Consequently, the determinant factor for the success or failure of the multivalent interaction is the amount of ΔS_{conf} that accumulates in the process. While the latter argument holds true, it is important to consider that the occurrence of the initial binding event is in fact at an entropic disadvantage in the multivalent case.

Recently, Weber *et al.* have shown this using two simplified model systems that, again, describe the scenario of two monovalent ligands versus one bivalent ligand binding to a dimeric receptor [30]. Each of the two model systems consists of a three-dimensional box with hard walls in which a receptor with two binding sites is placed at a fixed position. Additionally, the box contains either two monovalent ligands (scenario 1) or one bivalent ligand (scenario 2). In the first case, the ligands are allowed to diffuse in the box independent of each other. In the second case, the ligands are connected by a flexible spacer that imposes a restraint on their intermolecular distance. Thermodynamic data for the two systems was generated using Markov chain Monte Carlo (MCMC) [54, 55] simulations on the basis of a simple force field implementation. Potential energy plots of two typical simulation runs are shown in Figure 2.8.

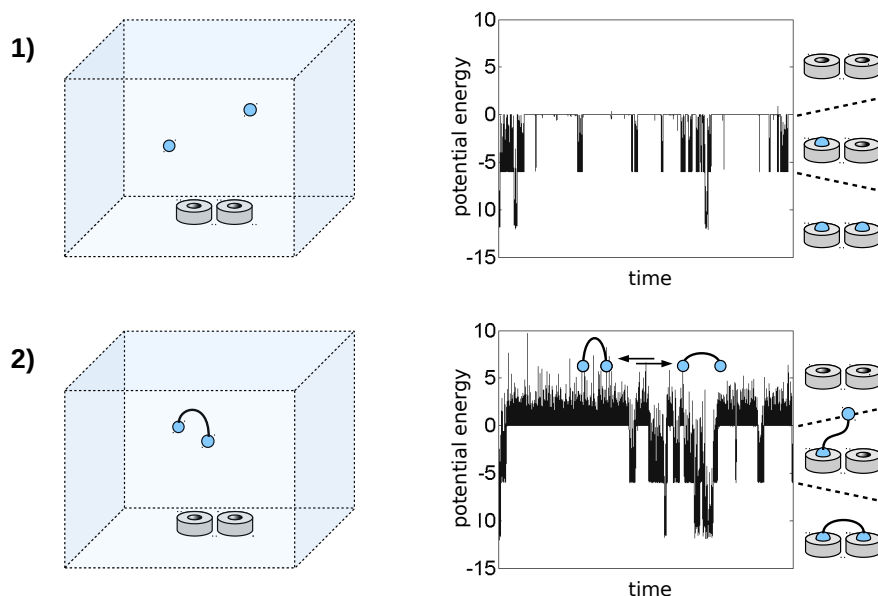


FIGURE 2.8: Potential energy plots of “monovalent” scenario 1 (top) and “bivalent” scenario 2 (bottom), derived from two exemplary Markov chain Monte Carlo (MCMC) simulations. The unit of energy is chosen arbitrarily [30].

One can clearly discriminate the three main energy levels of the systems: $V \approx 0$ is the unbound state of the system (no interaction between ligands and receptor binding sites). $V \approx -6$ is the singly bound state and $V \approx -12$ is the doubly bound state. In the “bivalent” scenario 2, vibrations of the spacer lead to oscillations in the distance restraint potential, particularly visible above the base line. The model does not include energetic interactions between spacer and receptor, which in a real system, depending on its chemical structure, can be non-negligible.

By plotting the minimal distance of the first ligand to the first receptor binding site against the distance of the second ligand to the second receptor binding site (where the pair with nearest distance is chosen as “first” pair), the conformational space of the two

systems can be projected unto two dimensions for better classification into the three macroscopic states “unbound”, “singly bound” and “doubly bound”. The effect of the spacer is clearly visible: Whereas in the absence of a spacer, almost every combination of ligand-receptor distances is possible, the attachment of a spacer leads to a strong correlation between the two collective variables. Consequently, a certain portion of conformational space remains unpopulated.

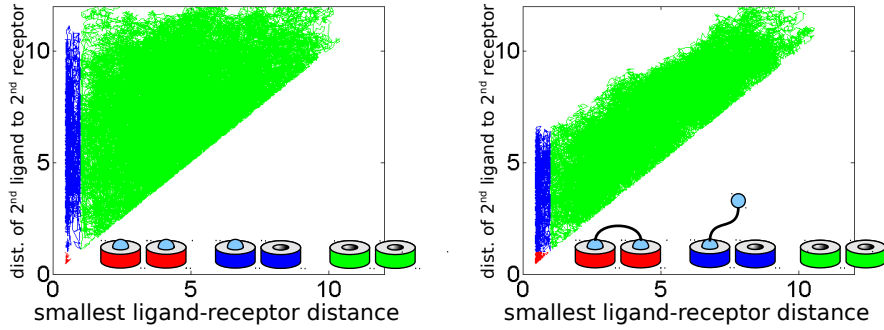


FIGURE 2.9: Conformational space representation of “monovalent” scenario **1** (left) and “bivalent” scenario **2** (right). Red: doubly bound state, blue: singly bound state, and green: unbound state [30].

Under the assumption that the potential energy values are approximately piece-wise constant within each of the three macroscopic states, the entropy differences can be calculated as ratios of the corresponding conformational space volumes. In order to compare the entropic cost² of the first binding event, one has to take the difference in entropy change upon transition from unbound to bound state for the monovalent and the bivalent system:

$$\Delta\Delta S_{1^{st}} = \Delta S_{1^{st}/bi} - \Delta S_{1^{st}/mono} = k_B \ln \left(\frac{2(a - \epsilon)^2}{(a - 2\epsilon + b)(2a - \epsilon)} \right), \quad (2.5)$$

where a is the maximum distance of a ligand to one of the receptors (depending on the box size), and ϵ is the ligand-receptor binding distance. Furthermore, $b = a - x - y$, where x is the maximum (i.e. extended or “stretched”) spacer length, and y is the distance between the receptor binding sites. Finally, k_B is Boltzmann’s constant. See Ref. [30] for the derivation of Equation 2.5.

Because of $\epsilon < b < a$, this expression is always negative: $\Delta S_{1^{st}/bi}$ is smaller than $\Delta S_{1^{st}/mono}$, which means that the entropy loss in the bivalent scenario is always larger. Due to the fact that a single bivalent ligand has a smaller “search space” than two monovalent ligands, the probability of a ligand entity encountering a receptor binding site is reduced, and the first ligand-receptor binding event in the bivalent case is entropically

²In the scope of this model, entropy changes refer exclusively to ΔS_{trans} , as the ligands are modeled as mere particles without any rotational degrees of freedom.

diminished. Consequently, in situations where, regardless of receptor valency, only one binding event per receptor is required, the use of a monovalent ligand is more promising. By contrast, in situations where the equilibrium is to be driven towards a state where all receptor binding sites are occupied, a multivalent ligand is the best choice.

2.3 Multivalent kinetics

Apart from the entropic factors discussed above, another cause for the characteristic properties of multivalent complexes stems from the underlying kinetics. The association and dissociation of monovalent ligand-receptor complexes is governed by two rate constants – k_{on} and k_{off} – that quantify the rate of transition from unbound to bound, and bound to unbound state, respectively. The ratio of these two rate constants yields K_d , the dissociation constant, $K_d = k_{off}/k_{on}$. K_d is widely used as a measure for the affinity of complexes, and can be related to the Gibbs free energy of binding by using the van't Hoff equation (cp. Equation 2.1). The rate constants, in turn, are mainly dependent on the concentrations of the two interacting species, and the environmental factors that dictate the microscopic dynamics of the system, such as temperature, pressure, and the viscosity of the solvent.

Describing the association and dissociation of multivalent complexes is intrinsically more complicated. Depending on the valency of the system, a model of multivalent binding in terms of reaction kinetics has to define additional kinetic entities in order to describe the various partially bound states that bridge the gap between the “all ligands unbound” and “all ligands bound” states, which in turn are connected by a complex network of “on” and “off” rate constants. This is feasible for small (typically synthetic) multivalent systems of defined valency, but becomes impractical for large (typically biological) multivalent systems where the valency cannot be controlled exactly. The description of multivalent binding processes in terms of reaction kinetics is complicated further by the fact that, by using current experimental methods, it is often not possible to (i) identify the partially bound states of the complex as kinetic entities of the system, supposedly due to their transient nature, and, consequently, (ii) determine the microscopic rate constants that quantify the transition rates between the kinetic entities. Therefore, experimental measurements of multivalent kinetics typically have to be interpreted as the average over all microscopic steps of the binding process.

In their prototypical study from 1998, Kramer and Karpen [56] studied the binding of PEG-linked dimers (PLDs) of cGMP to tetrameric cyclic-nucleotide-gated (CNG) channels, namely olfactory CNG and rod photoreceptor CNG. As the activation of

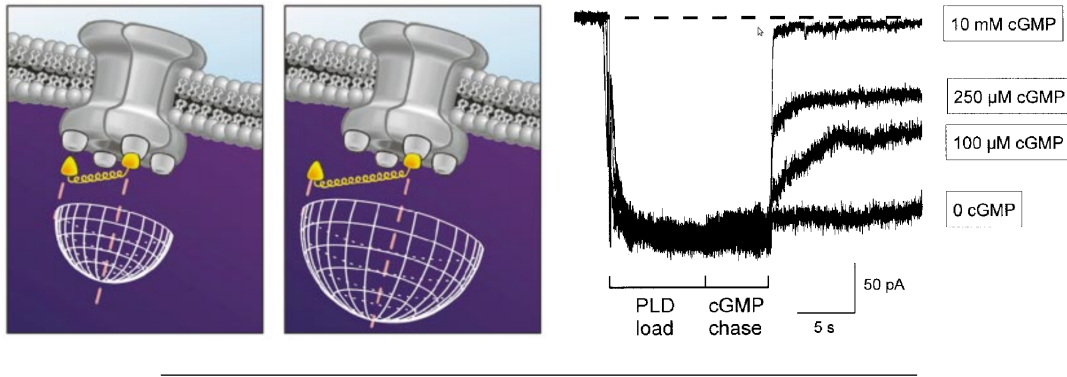


FIGURE 2.10: Left: Schematic of PEG-linked cGMP dimers binding to a tetrameric CNG channel. After the first ligand has bound, the effective concentration C_{eff} of the unbound ligand is determined by the hemisphere that is circumscribed by the tethered molecule. The length of the flexible spacer determines the radius of the hemisphere. Right: The “off” rate of the dimeric compound (“PLD”) accelerates with the amount of free, monovalent cGMP that is applied to the loaded channels. Figures taken from [56], reprinted with kind permission of Nature Publishing Group.

the channel protein by binding of cGMP induces a change of the membrane potential, the authors were able to record the averaged kinetics of the binding process with a patch clamp method. In comparison to the monovalent control, the dimeric compounds showed largely decelerated dissociation rates (up to 3400-fold slower than for monovalent cGMP), in particular when the average length of the PEG spacer presumably matched the binding site distance of the CNG channels, so that “channels remained open for minutes, even with continuous superfusion with agonist-free solution.” [56] Consequently, the K_d values for the bivalent compounds turned out to be up to 260-fold lower than for monovalent cGMP.

Based on their observations, the authors formulated an estimate for predicting the dissociation rate constant of a bivalent ligand that is doubly bound to two binding sites of a receptor, k_{off}^{bi} , as follows:

$$k_{off}^{bi} = 2k_{off} \cdot K_d / (C_{eff} + K_d), \quad (2.6)$$

where k_{off} is the intrinsic “off” rate of cGMP from a single binding site, K_d is the dissociation constant of cGMP, and C_{eff} is the effective concentration of cGMP at the second binding site. C_{eff} , in turn, is calculated from the volume of the hemisphere with radius r , the average length of the spacer that tethers the two ligand moieties. The above equation assumes that k_{off}^{bi} is equal to twice the monovalent k_{off} , times the probability that the second site is not occupied. The authors propose that, as soon that a single cGMP dissociates, it “has a high probability of rebinding before its partner can also dissociate.” [56] Unfortunately, it is not reported in how far the predictions are in accordance with the measured rates.

In a study from 2000 by Rao *et al.* [57], the authors designed and characterized a high affinity trivalent ligand-receptor system derived from vancomycin (a glycopeptide antibiotic) and the peptide L-Lys-D-Ala-D-Ala. By using high-performance liquid chromatography (HPLC) in combination with a competitive binding assay, they estimated a dissociation constant in water of approximately $4 \cdot 10^{-17}$ M, which is “one of the tightest known for low molecular weight organic species” [57], and surpasses the corresponding monovalent dissociation constant ($K_d^m \approx 1.6 \cdot 10^{-6}$ M) by several orders of magnitude. In order to explain the high affinity of their trivalent complex, they proposed a binding mechanism based on a kinetic model that consists of four main states (one unbound state and three bound states of increasing valency), and consequently requires three pairs of on- and off-rates (Figure 2.11).

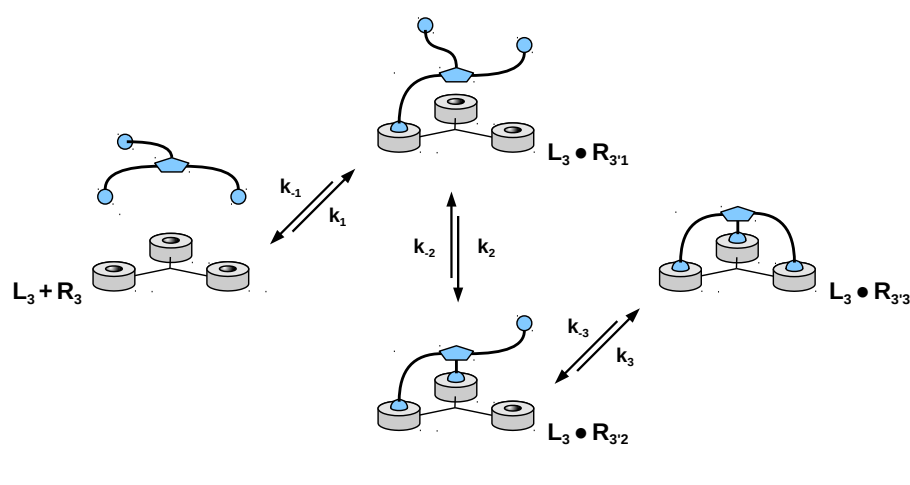


FIGURE 2.11: Kinetic model of a trivalent ligand-receptor system, adapted from [57]. The association and dissociation of complex $\mathbf{L}_3 \bullet \mathbf{R}_{3^3}$ from ligand \mathbf{L}_3 and receptor \mathbf{R}_3 is assumed to a stepwise process with clearly defined kinetic entities.

Due to the fact that the rate constants for the trivalent system could not be measured experimentally, the authors proposed an estimate of the kinetic parameters of the system that interpolates from the (measurable) k_{on} and k_{off} values of the monovalent interaction, and applies statistical factors to take into account the valency of the system. The estimate relies on three main assumptions: (i) The value of the rate constant for the first step in association of \mathbf{L}_3 with $\mathbf{R} =_3$, k_1 , is nine times that of k_{on} for monomeric \mathbf{L} and \mathbf{R} : $k_1 = 9k_{on}$. (ii) The rate constant for dissociation of each $\mathbf{L} \bullet \mathbf{R}$ pair in $\mathbf{L}_3 \bullet \mathbf{R}_{3^3}$ is approximately the same as k_{off} for monomeric $\mathbf{L} \bullet \mathbf{R}$: $k_{-1} = k_{-2}/2 = k_{-3}/3 = k_{off}$. (iii) The rate constants of the two “following” associations, after corrections for statistical factors, are approximately the same: $k_2/4 = k_3$. Combining these assumptions, k_2 and

k_3 can be calculated as follows:

$$\begin{aligned} K_d^t &= (k_{-1}k_{-2}k_{-3})/(k_1k_2k_3) \\ &= (k_{off} \cdot 2k_{off} \cdot 3k_{off})/(9k_{on} \cdot k_2k_3) \\ &\approx 4 \cdot 10^{-17} \end{aligned} \quad (2.7)$$

With $k_{on} \approx 2.8 \cdot 10^7 \text{ M}^{-1} \text{ s}^{-1}$ and $k_{off} \approx 30 \text{ s}^{-1}$, the authors estimated the rates of the intra-complex associations to be $k_2 = 8 \cdot 10^6 \text{ s}^{-1}$ and $k_3 = 2 \cdot 10^6 \text{ s}^{-1}$. They conclude that the large values of these rate constants indicate “fast intramolecular rebinding” [57] of partially dissociated $\mathbf{L} \bullet \mathbf{R}$ pairs, which in turn they suggest to be the kinetic origin of the high stability of complex $\mathbf{L}_3 \bullet \mathbf{R}_{3 \cdot 3}$.

Although the above estimates of rate constants are highly approximate (e.g. in not considering possible variations in the “off” rates), they are very insightful for pointing out the main properties of multivalent binding kinetics. Two main points can be maintained: (i) The association of multivalent complexes is discriminative with regard to the first and all subsequent binding events. (ii) The dissociation of multivalent complexes is presumably slowed down by the occurrence of rebinding events.

In answer to (i), mainly inspired by the field of supramolecular chemistry, it has indeed become common practice to distinguish between the first and possible subsequent binding events of a multivalent binding process [1]. More precisely, while the initial binding event is considered to be intermolecular binding (as in “conventional” ligand-receptor interactions), the subsequent binding events are considered to be intramolecular³ binding. This differentiation is based on the observation that, while the initial, intermolecular binding event is dependent on the concentrations of both ligand and receptor, the subsequent, intramolecular binding events are solely dependent on the concentration of the ligand. Moreover, the concentration of the ligand in intramolecular binding is bound to be (locally) increased, as the spatial proximity of ligand and receptor has already been established. This notion is explained in more detail in Section 2.4 on cooperativity.

In the kinetic model of bivalent binding proposed by Kramer and Karpen, this phenomenon is formulated in terms of the effective concentration C_{eff} . Given that the first ligand has bound to the receptor, the value of C_{eff} is calculated under the assumption that the second, unbound ligand is uniformly distributed in a hemisphere of a radius determined by the length of the spacer that is used to tether the two ligands (cp. Figure 2.10). This model of an increased local concentration was refined by Diestler and Knapp [58], who proposed that the unbound, tethered ligand is distributed according to

³In this context, despite usage of the term “intramolecular”, the ligand-receptor interactions in question are still assumed to be non-covalent and reversible.

a three-dimensional Gaussian distribution. In practice, C_{eff} is hard to predict due to its dependence on various (mainly structural) parameters of the system, and in particular the spacer used to tether the ligands, a problem that will be elaborated in Chapter 4, dealing with the conformational properties of flexible spacer structures. More severely, C_{eff} cannot be determined by experiment.

A more widespread approach of taking into account the increase of local concentration in multivalent binding is given in terms of the effective molarity (EM). The concept of effective molarity, which has the unit of a concentration (mol/l), is to weight the intramolecular equilibrium constants with a factor that “accounts for the ease of the intramolecular process” [59], and presumably corresponds to the gain in local concentration that is caused by the spatial proximity of the unbound ligands to the receptor (a more precise definition will follow in the next section). In contrast to the effective concentration, the effective molarity can be determined by experimental methods, at least for synthetic systems of defined, and typically low, valency. Hunter and coworkers [60, 61], for example, were able to quantify the effective molarity in a series of bivalent ligand-receptor systems by implementing the “double mutant cycle”, an approach that since then has been proven successful in different scenarios [11, 62].

Recently, Hogben *et al.* [63] used denaturation titrations in order to investigate the stability of the complexes of cyclic and linear zinc-porphyrin hexamers with multidentate ligands having up to six pyridyl coordination sites (Figure 2.12). The results reveal that the stepwise effective molarities for the third through sixth intramolecular coordination events with the cyclic hexamer are extremely high ($EM = 10^2$ M and above), whereas the values for the linear porphyrin oligomers are only modest ($EM \approx 0.05$ M). This example shows very impressively that the increase in effective molarity is directly related to the preorganization of the binding sites of a multivalent ligand with regard to the structure of the receptor.

Even more interesting, the effective molarity can be developed into a measure for a cooperative effect in multivalent binding processes. Consequently, it is possible to link the effective molarity (or, in other words, the increase in local concentration after the initial binding event) to the very origins of the multivalent enhancement effect. The phenomenon of cooperativity in multivalent systems (and a delineation from allosteric cooperativity) will be presented in the next section.

Point (ii), the occurrence of rebinding effects, was investigated by Kramer and Karpen by conducting a competitive binding assay. For monovalent ligands, the dissociation from a single binding site is “normally unaffected by the concentration of free ligand in solution.” [56] For a partially dissociated multivalent ligand, however, the free ligand in solution becomes a competitor for the unoccupied binding sites that are available for rebinding – an idea that is depicted in Figure 2.13. Indeed, the authors found that

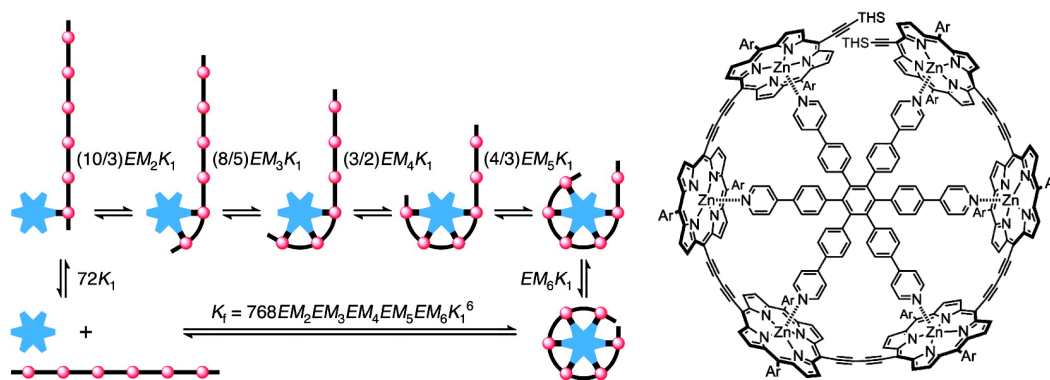


FIGURE 2.12: Stepwise equilibria defining the effective molarities $EM_2 - EM_6$ for formation of the linear zinc-porphyrin hexamer-hexapyridil complex as shown on the right. $K_1 = K_{mono}$ is the microscopic binding constant for coordination of a pyridyl group to one face of a zinc site. For information on the statistical factors, please refer to the original reference. Figures taken from [63], reprinted with kind permission of the American Chemical Society.

delayed application of cGMP to the loaded receptors accelerated the “off” rates of their PEG-linked cGMP dimers, proportional to the concentration of free ligand that was applied (also cp. Figure 2.10).

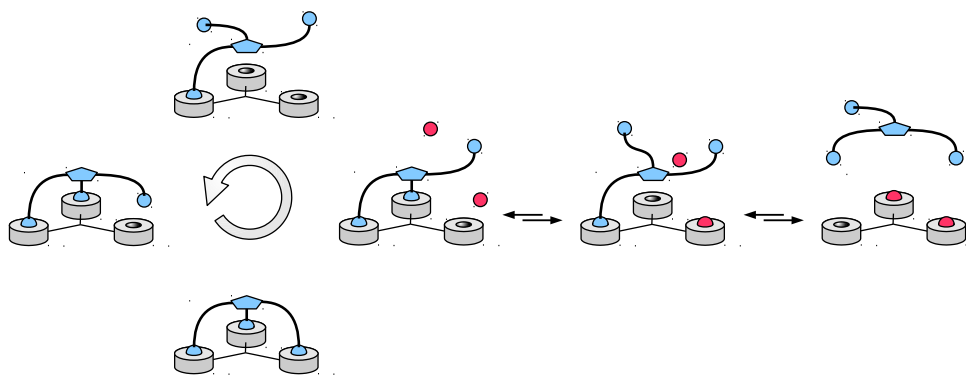


FIGURE 2.13: Schematic representation of rebinding in a trivalent complex, adapted from [57]. As soon as a single ligand dissociates, the spatial proximity to the receptor is maintained by the remaining partner(s) associated with the receptor. Consequently, it has a high probability of rebinding. The occurrence of rebinding events becomes less prevalent when another ligand (red) is competing for the receptor binding sites.

A similar competitive binding assay, quantified by means of HPLC, was conducted by Rao *et al.* [57] for their trivalent system introduced above (cp. Figure 2.7). They found that a large excess of monovalent ligand L in the millimolar range (≈ 86 mM) was necessary to induce a dissociation of the complex $L_3 \bullet R_{3 \times 3}$ (formed from an initial solution of $22 \mu\text{M } L_3$ and $3 \mu\text{M } R_{3 \times 3}$, respectively). Surprisingly, even given the high excess of L , the trivalent complex $L_3 \bullet R_{3 \times 3}$ could still be detected by HPLC in significant amounts after equilibration. This findings not only support the idea of rebinding in

general, but once more indicate that the increase in local concentration (or effective molarity) for a well-designed multivalent ligand can be very large.

While rebinding is a likely factor for the high stability of multivalent complexes, it poses certain challenges when it comes to formulating kinetic models of multivalent binding processes. The “conventional” kinetic model of ligand-receptor binding is based on the assumption that the kinetic entities – the steps of the binding process – are intrinsically Markovian, which means that the future of the system, e.g. the probability of transition from unbound to bound state, is dependent only on its present state. While this holds true for systems where the states are separated by high energy barriers, it can be shown that Markovianity is lost when the states are not well separated [64]. In a recent study, Weber *et al.* [30] argued that the transient nature of partially bound states in rebinding systems requires a mathematical reformulation of the kinetic model of multivalent binding in terms of a “soft”, function-based definition of states in order to quantify the contribution of rebinding correctly. The notion of a soft definition of kinetic entities in the formalization of (multivalent) ligand-receptor binding processes will be picked up again in Chapter 6.

2.4 Cooperativity in multivalent systems

In 1998, Whitesides *et al.* [2] first proposed a quantity for characterizing the enhancement effect in a multivalent interaction: The enhancement factor ϵ^4 is defined in terms of the ratio of the association constant of the multivalent interaction K_{multi} , and the association constant of the monovalent interaction K_{mono} (Equation 2.8).

$$\epsilon = \frac{K_{multi}}{K_{mono}} \quad (2.8)$$

Consequently, the enhancement factor ϵ comprises all parameters that distinguish the multivalent from the monovalent interaction, and can even be applied to systems where the exact number of binding sites is unknown. While this seems to be practical in many situations, it has the disadvantage that it is not possible to separate symmetry effects from other effects that enhance or diminish the binding affinity in a multivalent interaction [1]. These “other” effects can be summarized as cooperativity effects.

Cooperativity is a widely used concept to quantify the special properties of systems with coupled interactions (a feature that is not necessarily limited to multivalent systems). A system is considered cooperative if the collective system acts differently than what is to be expected from the sum of its individual interactions. More precisely, positive

⁴Whitesides’ enhancement factor ϵ is denoted as β in the original publication. This notation is not adapted in order to discriminate from the cooperativity factors α and β (Equations 2.9 and 2.13).

cooperativity describes a scenario where the interactions in the collective system are enhanced, while negative cooperativity describes a scenario where the interactions in the collective system are diminished – each compared to the sum of the individual interactions in a non-cooperative reference system [65].

Generally, one delineates two different types of cooperativity, namely allosteric cooperativity, and chelate cooperativity. The former phenomenon is well characterized and not related to multivalent interactions *per se* (cp. Section 2.2.1, where the prime example for allosteric cooperativity in biology, hemoglobin, was briefly mentioned). Chelate cooperativity, however, is less well recognized than its counterpart, and represents (to a large extent) a synonym for multivalent cooperativity. The naming presumably stems from the fact that chelate complexes were the initial objects of study in this regard (cp. Section 2.1). Due to the fact that both effects can trigger similar behavior on the macroscopic level, in spite of having a completely different origin on the molecular level, separating the two types of cooperativity has been a notoriously difficult problem [65]. Ercolani and Schiaffino define the origin of allosteric cooperativity to be the interplay of two or more intermolecular binding interactions, whereas chelate cooperativity “arises from the presence of one or more intramolecular binding interactions [...] as a consequence of the chelate effect.” [59] This will be illustrated for a generic bivalent ligand-receptor system in the following.

Allosteric cooperativity

Figure 2.14 shows the binding of a monovalent ligand **L** to a bivalent receptor **RR**. The two equilibria between the three main states of the system are described in terms of the microscopic association constants K_1 and K_2 , multiplied with the statistical factors that arise from the degeneracy of the singly bound “intermediate” state [32]. The reference constant K_{mono} can be obtained by studying the binding of **L** to a monovalent version **R** of the bivalent receptor **RR**, or, alternatively, “by directly taking the value of the constant K_1 as the reference constant” [59].

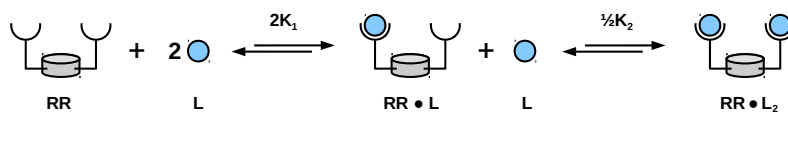


FIGURE 2.14: Binding of a monovalent ligand **L** to a bivalent receptor **RR**, adapted from [59].

The allosteric (intersite) cooperativity of the above system is evaluated by comparing the joint association constant of the (presumably) cooperative system $K_1 \cdot K_2$, with the

joint association constant of the non-cooperative reference system, K_{mono}^2 . The ratio of the two quantities yields the allosteric cooperativity factor α (Equation 2.9).

$$\alpha = \frac{K_1 \cdot K_2}{K_{mono}^2} \quad (2.9)$$

For $\alpha < 1$, the system under study exhibits negative cooperativity, while for $\alpha > 1$, it is considered to be positively cooperative. In both cases, intermolecular binding of **L** to the two binding sites of **RR** is coupled by one or more possible factors (e.g. electrostatic, or steric/conformational interactions). If α takes the value 1, the system under study is non-cooperative, and its association constants correspond to the association constant of the reference system: $K_2 = K_1 = K_{mono}$. This means that intermolecular binding of **L** to the two binding sites of **RR** is not coupled in any way.

Chelate cooperativity

The second type of cooperativity, chelate cooperativity, can occur only in multivalent systems, and characterizes cooperativity effects that stem exclusively from the multivalent properties of the system under study. Synonymously used terms for chelate cooperativity are multivalent cooperativity, multivalent enhancement, or, plainly, “multivalency” [1]. Due to the fact that the latter terms are often not well separated from other enhancing effects that may play a role in multivalent binding, it is helpful to remain with the term chelate cooperativity.

The concept of chelate cooperativity is based on regarding all but the first binding events in a multivalent ligand-receptor interaction as intramolecular binding events. The grade of coupling between the intramolecular binding interactions is quantified by the effective molarity (EM). Figure 2.14 shows the binding of a bivalent ligand **LL** to a bivalent receptor **RR**, under the assumptions that (i) the ligand is present in a large excess relative to the receptor ($[\mathbf{LL}]_0 \gg [\mathbf{RR}]_0$)⁵, and (ii) allosteric cooperativity is excluded ($\alpha = 1$). Under these conditions, the receptor can exist in only four possible states, namely free (**RR**), partially bound in a 1:1 open complex ($o - \mathbf{RR} \bullet \mathbf{LL}$), in a doubly bound cyclic complex ($c - \mathbf{RR} \bullet \mathbf{LL}$), and in a doubly bound 1:2 complex involving two ligand molecules ($\mathbf{RR} \bullet (\mathbf{LL})_2$) [59].

The formation of the doubly bound cyclic complex is determined by the intramolecular association constant K_{intra} , which is defined as the product $1/2 K_{mono} \cdot \text{EM}$. Analogous to the previous example, $1/2$ is a statistical factor for the cyclization process, and K_{mono} is the reference association constant. EM, in turn, denotes the effective molarity. Ercolani

⁵This assumption simplifies the situation in so far that complexes involving more than one receptor molecule (e.g. $(\mathbf{RR})_2 \bullet \mathbf{LL}$) can be neglected.

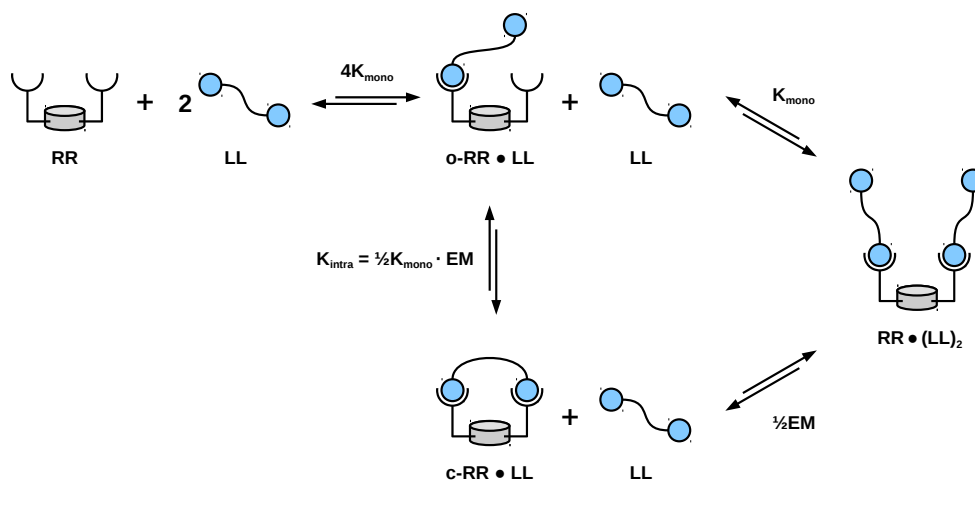


FIGURE 2.15: Binding of a bivalent ligand \mathbf{LL} to a bivalent receptor \mathbf{RR} , adapted from [59]. This simplified model assumes that $[\mathbf{LL}]_0 \gg [\mathbf{RR}]_0$ and $\alpha = 0$.

and Schiaffino [59] demonstrate the effect of chelate cooperativity on the above system by comparing the speciation profile (the fractions of the four “species”, depending on the concentration $[\mathbf{LL}]_0$) in the absence ($K_{intra} = 0$) and in the presence of the chelate interaction (e.g. $K_{intra} = 25$, corresponding to $K_{mono} \cdot EM = 50$), as shown in Figure 2.16.

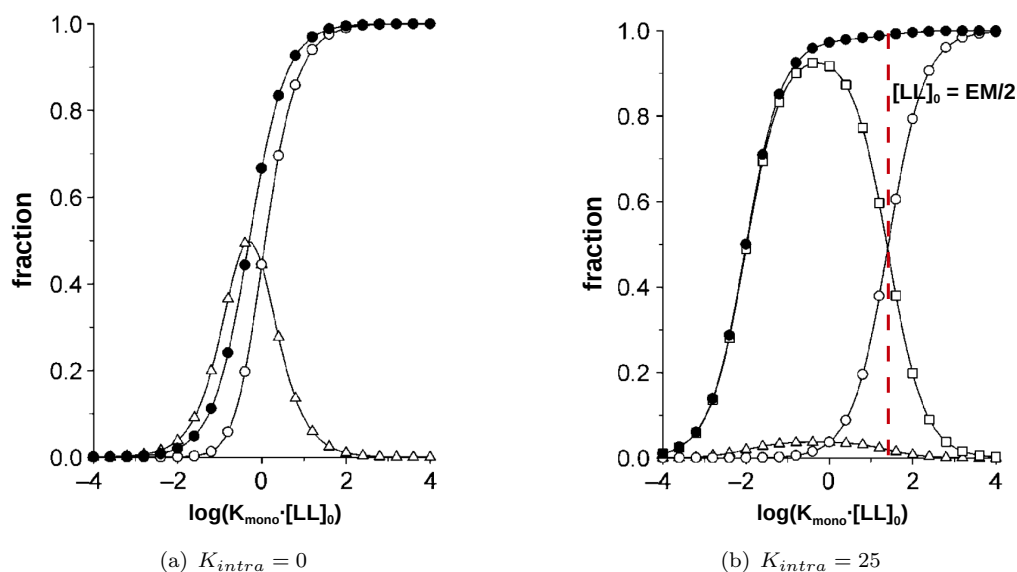


FIGURE 2.16: Speciation profiles for the equilibria shown in Figure 2.15, (a) in the absence of chelate cooperativity, and (b) in the presence of chelate cooperativity. The concentration scale is normalized by multiplying with K_{mono} . The fractions of the different species are represented by the following symbols: $\triangle \triangleq o - \mathbf{RR} \bullet \mathbf{LL}$, $\square \triangleq c - \mathbf{RR} \bullet \mathbf{LL}$, and $\bullet \triangleq$ total fraction of occupied \mathbf{RR} binding sites.

Figure taken from [59], reprinted with kind permission of John Wiley and Sons.

Intuitively, with $K_{intra} = 0$, the cyclic complex $c - \mathbf{RR} \bullet \mathbf{LL}$ does not form at all. With

increasing concentration of \mathbf{LL} , however, the receptor binding sites are progressively saturated, and eventually the partially bound open complex $o - \mathbf{RR} \bullet \mathbf{LL}$ is displaced completely by the doubly bound 1:2 complex involving two ligand molecules, $\mathbf{RR} \bullet (\mathbf{LL})_2$. By contrast, the presence of the chelate interaction leads to a sharp decrease of the partially bound complex $c - \mathbf{RR} \bullet \mathbf{LL}$, in favor of the doubly bound cyclic complex $c - \mathbf{RR} \bullet \mathbf{LL}$. This “all-or-none” behavior, the depletion of intermediate states, is characteristic of cooperative systems [65], and once more underlines that multivalency is successful in driving the system towards a state where a maximum number of ligands is bound. A second look at the speciation profile of $c - \mathbf{RR} \bullet \mathbf{LL}$ reveals another interesting property of the chelate interaction: At high concentrations of \mathbf{LL} , the cyclic complex is competing (and eventually losing) against the doubly bound 1:2 open complex $\mathbf{RR} \bullet (\mathbf{LL})_2$, which results in a bell-shaped, or “none-all-none” [59] profile.

This can be explained by the fact that the concentration of the cyclic complex $c - \mathbf{RR} \bullet \mathbf{LL}$ depends linearly on the ligand concentration (Equation 2.10), whereas the concentration of the doubly bound open complex $\mathbf{RR} \bullet (\mathbf{LL})_2$ depends on the square of the ligand concentration (Equation 2.11):

$$[c - \mathbf{RR} \bullet \mathbf{LL}] = 2K_{mono}^2 \cdot EM \cdot [\mathbf{RR}] \cdot [\mathbf{LL}] \quad (2.10)$$

$$[\mathbf{RR} \bullet (\mathbf{LL})_2] = 4K_{mono}^2 \cdot [\mathbf{RR}] \cdot [\mathbf{LL}]^2 \quad (2.11)$$

Consequently, the ligand concentration at which the switch between the cyclic and the doubly bound open complex occurs can be obtained by equating the right-hand sides of the two equations above (Equation 2.12, also cp. Figure 2.16 (b)).

$$[\mathbf{LL}]_{switch} = EM/2 \quad (2.12)$$

Thus, Ercolani and Schiaffino reason that “EM is the threshold concentration of ligand binding groups \mathbf{L} above which the intramolecular process loses the competition with the intermolecular one” [59], and conclude that chelate cooperativity, in contrast to allosteric cooperativity, is dependent on the ligand concentration. Interestingly, a similar phenomenon had already been described in 1912 by the Swiss chemist Ruggli, who found that in covalent cyclization reactions, intramolecular interactions are favored over intermolecular interactions at low concentrations (Ziegler-Ruggli dilution principle) [66].

Several approaches of quantifying the degree of chelate cooperativity for a given system have been discussed in the literature [59, 65]. Considering the effective molarity to be an immediate measure for chelate cooperativity is not consistent, as it has the unit of a concentration, whereas a “proper” cooperativity factor similar to α (being the ratio of two equilibrium constants) should be dimensionless. Hunter *et al.* [65] proposed the product $K_{mono} \cdot EM$ to be a suitable cooperativity factor, which however was challenged

by Ercolani and Schiaffino [59], who found this choice to be misleading due to the fact that (i) it coincides with the intramolecular association constant K_{intra} , save for the statistical factor $1/2$, (ii) it does not take into account the ligand concentration on which chelate cooperativity depends, and (iii) it tends to be zero in the absence of chelate cooperativity, where it should tend to 1.

The authors argue that the correct way to assess chelate cooperativity requires that the intermolecular binding leading to the doubly bound open complex $\mathbf{RR} \bullet (\mathbf{LL})_2$ (with a joint association constant of $4K_{mono}^2$) has to be used as a reference. Following this approach, one obtains two different cooperativity factors β and β' (Equations 2.13 and 2.14).

$$\beta = \frac{EM}{2 \cdot [\mathbf{LL}]} \quad (2.13)$$

The factor β (Equations 2.13) relates intramolecular binding with the intermolecular binding, and can be interpreted as the “apparent” (or virtual) equilibrium constant for the conversion of the non-cooperative doubly bound open complex $\mathbf{RR} \bullet (\mathbf{LL})_2$ into the cooperative doubly bound cyclic complex $c - \mathbf{RR} \bullet \mathbf{LL}$. The value of β is dependent on the ligand concentration. If the concentration of the bivalent ligand is equal to $EM/2$, there is no chelate cooperativity ($\beta = 1$), whereas positive chelate cooperativity ($\beta > 1$) requires $[\mathbf{LL}] < EM/2$, and, vice versa, $[\mathbf{LL}] > EM/2$ yields negative chelate cooperativity ($\beta < 1$). Consequently, the effective molarity of the intramolecular interaction has a noteworthy impact on the equilibrium of the system: “The higher the value of EM , the larger the concentration range over which the chelate interaction displays positive cooperativity.” [59]

β' , in turn, relates the overall binding (the sum of intermolecular and intramolecular binding) with “pure” intermolecular binding, or, in other words, the monovalent reference system (Equation 2.14).

$$\beta' = 1 + \frac{EM}{2 \cdot [\mathbf{LL}]} \quad (2.14)$$

In this situation, there can only be positive chelate cooperativity: The factor β' is never smaller than one. Hereby, Ercolani and Schiaffino [59] show that multivalent binding is always favored when competing with comparable monovalent binding, unless the advantage is nullified by negative allosteric cooperativity.

While the newly defined chelate cooperativity factors β and β' are theoretically rigorous in separating the multivalent cooperativity from other effects, they suffer from their applicability to real systems. The precondition that the ligand is present in a large (and invariant) excess relative to the receptor, $[\mathbf{LL}]_0 \gg [\mathbf{RR}]_0$, is not given in typical ITC titration experiments, where the concentration of the ligand is significantly smaller than the concentration of the receptor in the beginning, and larger at the end of the titration [66]. In lack of a better alternative, Hunter’s aforementioned approach of evaluating the

expression $K_{mono} \cdot EM$ remains a practical means for studying the chelate cooperativity of bivalent complexes [1].

The above example of a bivalent system paints a “minimal” (and arguably not very impressive) picture of the phenomenon of chelate cooperativity. Nonetheless, it is the same phenomenon that enables fundamental processes such as supramolecular self-organization and protein folding [65].

In summary, the theoretical concepts of describing multivalent enhancement are manifold, but also (at least partly) redundant. A common motif of all models is the idea that the quasi-intramolecular binding events (following a first binding event between multivalent receptor and multivalent ligand) are favored due to preorganization of the interacting entities. The impact of this preorganization can either be expressed in terms of reduced entropy loss upon binding, or an increase in the local (or effective) concentration of ligand (C_{eff} and EM). The concept of EM is enticing as it is, at least for some multivalent systems, derivable from experiment and can be linked directly to the degree of chelate cooperativity to be expected. Finally, as has been pointed out in Ref. [30], separating the contribution of rebinding events from other aspects of multivalent binding may pose difficulties when using standard kinetic models that rely on the presence of high energy barriers between unbound and bound states, i.e. when binding is perceived in the spirit of an activated process.

Chapter 3

Molecular simulation – a theoretical background

3.1 Basics of classical molecular simulation

Microscopic systems formed from elementary particles are governed by the laws of quantum mechanics. Due to the fact that particles can be strongly delocalized, even to the extent of having wavelike properties, the probability density of finding a particle in a given place at a given time is described in terms of a wave function. The dynamics of how the wave function evolves over time is described by the Schrödinger equation, a partial differential equation that is fundamental to quantum mechanics. Unfortunately, given the high computational cost, solving the Schrödinger equation “as is” is feasible only for the smallest (two-atomic) molecules [67].

For systems involving large masses and energies, the laws of quantum mechanics are known to give over to the laws of classical mechanics, which govern the dynamics of macroscopic systems. Molecular systems as observed in the course of this thesis are on the borderline of these two regimes. While the electrons show a distinctly quantum mechanical behavior, the atomic nuclei can in most situations be described in terms of classical mechanics, a property that is owed to their high mass. The disparity between the mass scales of electrons and nuclei gives rise to both the Born-Oppenheimer and the classical approximation [19].

The Born-Oppenheimer approximation aims at overcoming the high computational cost of solving the Schrödinger equation for coupled nuclei and electrons by splitting the computation in two consecutive steps. In the first step, the electronic Schrödinger equation is solved, yielding the wave function depending on electrons only. For this purpose, the nuclei are fixed in a certain configuration. This approach is justifiable by the fact that,

from the perspective of the fast-moving electrons, the disproportionately heavier nuclei are rendered virtually immobile [19]. In the second step, the electronic wave function serves as a potential in a Schrödinger equation that contains only the (now mobilized) nuclei: Every electronic state of the molecule creates a potential along which the nuclei are moving. The Born-Oppenheimer approximation delivers good results for molecules in the ground state, in particular for atoms with heavy nuclei. Unfortunately, the computational cost of solving the electronic Schrödinger equation *ab initio*, independent of using empirically derived parameters, remains high. The use of wave function-based *ab initio* methods¹ is typically limited to systems of no more than 20 (coupled-cluster theory [69]) or 100 atoms (Møller-Plesset perturbation theory of the second order [70]). By comparison, the relatively new approach of density functional theory (DFT), working on functionals of electron density instead of wave functions, is apt to describe systems of up to 1,000 atoms. DFT is in principle another *ab initio* method, but semi-empirical in character due to necessary approximations in the density functionals. Finally, fully semi-empirical methods are able to cope with systems of up to 10,000 atoms, but are dependent on parameters from empirical data, and choose to omit certain types of interactions completely [67].

While the ongoing improvement of quantum mechanical methods allows for the handling of increasingly large systems, these class of methods remains predominantly a tool for determining precise ground state energy, and not for studying larger dynamical processes. In order to explore the conformational space of biomolecular systems, not uncommonly consisting of hundreds of thousands of particles, a further simplification of the dynamics is necessary. This need leads to the classical approximation – the description of a molecular system in terms of classical mechanics² – which describes the motion of bodies under the action of a system of forces, and, in different (re)formulations, goes back to Newton, Lagrange, Hamilton and Liouville [19].

Any formulation of classical mechanics requires that, for every particle involved, the current position and velocity can be determined. This is directly contradictory to Heisenberg's uncertainty principle, a fundamental convention of quantum mechanics, which states that in a quantum mechanical system, two complementary properties of a particle cannot be determined with arbitrary precision at the same time. Consequently, the spatial whereabouts of electrons in quantum mechanics are described in terms of probability densities, and not in terms of positions (as defined by coordinates), velocities, and the resulting trajectories known from classical mechanics. The classical approximation

¹The foundation for this class of methods is the Hartree-Fock approach [68], which by now has been enhanced by electron correlation methods to allow for higher precision [67].

²The general study of mechanics began as early as the 4th century BC (Aristotelian physics), but the term *classical mechanics*, coined in the early 20th century, typically refers to the system of physics that was begun by Newton and his contemporaries.

acknowledges the inability to properly describe electrons in a very straightforward manner – by ignoring them. Therefore, the smallest elementary particle that is considered in classical molecular simulation is typically the atom (or ion). However, certain properties that arise from electronic effects are emulated in classical simulations, e.g. by attaching a non-integer partial charge to every classical particle, which in turn adds to the potential that acts on the particles. Despite its lack of sub-atomic precision, classical molecular simulation has been successful in reproducing the macroscopic properties of various molecular systems, in particular because the empirical potential functions that are used in classical simulation can be tailored and tuned for the application in question (this is elaborated in the following section).

One of the main limitations for the classical simulation of biomolecular systems (apart from the aforementioned sampling problem that affects very large systems, cp. Chapter 1.4) is the inability to model charge transfer reactions, and the dynamic breaking and formation of covalent bonds. Therefore, it is for instance not possible to modulate the protonation state of amino acid side chains by simulating a pH titration experiment. Currently, the hybrid quantum mechanics/molecular mechanics approach (QM/MM) [71] as well as (yet experimental) polarizable force fields [72] try to offer some relief in this respect. However, for modeling dynamic, non-covalent interactions that involve macromolecules, and in particular when explicitly modeled solvent is required, classical molecular simulation remains the one method of choice, and it is extensively used in the context of this thesis. In the following, the theory behind this method is presented in more detail.

3.1.1 Phase space and force field

In the classical approximation, a specific three-dimensional configuration of the N -particle molecular system under observation is described as a position state q , where $q \in \mathbb{R}^{3N}$. The whole molecular position space, consisting of all possible position states q , is then denoted as Ω . Analogously to q , $p \in \mathbb{R}^{3N}$ is the vector of collective momenta acting on q . By joining q and p , one obtains a complete characterization of the molecular system in terms of classical mechanics, given by the phase space vector $\gamma = (q, p)^\top \in \mathbb{R}^{6N}$.

The most commonly used formulation of classical mechanics in the context of molecular simulation is the one by Hamilton. Therein, the total energy of a single point γ of the phase space Γ is given in terms of the separable Hamiltonian, the sum of the potential energy U and the kinetic energy K (Equation 3.1).

$$H(\gamma) = H(q, p) = U(q) + K(p) = U(q) + \frac{1}{2}p^T \mathcal{M}p \quad (3.1)$$

\mathcal{M} represents the $3N \times 3N$ -mass matrix, which for every atom k contains the corresponding atom mass m_k thrice on its diagonal. While the kinetic energy K , dependent only on p , can be calculated directly from atom masses and momenta, the potential energy U , dependent only on q , has to be approximated by a molecular force field. Molecular force fields describe U as the sum of a number of energy terms that represent all bonded and non-bonded interactions in the system. In order to assemble a complete force field representation, the atom types of all particles as well as their connectivity have to be known. The energy terms, in turn, are usually very simplistic approximations of the quantum mechanical potentials that are at work in “true” molecular systems, and have to be considered as a necessary compromise between the speed of computational evaluation and physico-chemical accuracy.

For example, the bond stretching potential U_b between two covalently bonded atoms i and j is often represented by a simple harmonic potential (Figure 3.1) of the form

$$U_b(r_{ij}) = 1/2 k_{ij}^b (r_{ij} - b_{ij})^2, \quad (3.2)$$

so that the resulting force acting on atom i is

$$\mathbf{F}_i(\mathbf{r}_{ij}) = k_{ij}^b (r_{ij} - b_{ij}) \frac{\mathbf{r}_{ij}}{r_{ij}}. \quad (3.3)$$

In the above equations [73], the force constant k_{ij}^b as well as the equilibrium bond length b_{ij} are parameters that have to be derived from *ab initio* calculations beforehand. Therefore, a classical force field is typically tailored for performing a certain limited task, e.g. for reproducing peptide dynamics at a temperature of 300 K in water, and loses its validity if certain of these preconditions are not met. Consequently, in classical molecular simulation, the force field has to be chosen on the basis of the application.

The Lennard-Jones potential U_{LJ} , in turn, describes the interaction energy between two non-bonded uncharged atoms, and is given by

$$U_{LJ}(r_{ij}) = \left(\frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^6} \right), \quad (3.4)$$

with a resulting force of

$$\mathbf{F}_i(r_{ij}) = \left(12 \frac{C_{ij}^{(12)}}{r_{ij}^{13}} - 6 \frac{C_{ij}^{(6)}}{r_{ij}^7} \right) \frac{\mathbf{r}_{ij}}{r_{ij}}. \quad (3.5)$$

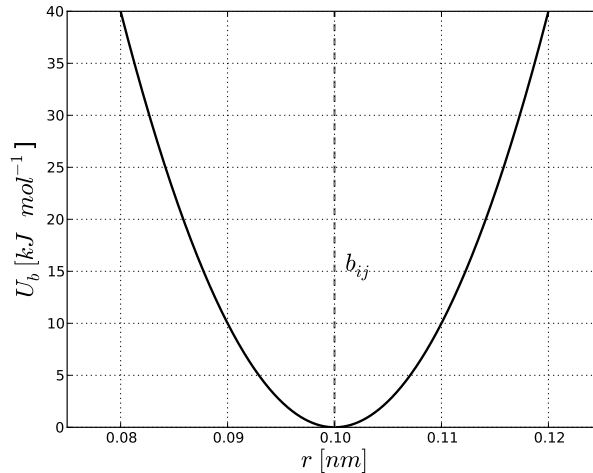


FIGURE 3.1: Plot of a harmonic bond-stretching potential as used in the molecular dynamics package GROMACS for modeling the oscillation of covalent bonds [73]. The equilibrium bond length b_{ij} as well as the force constant k_{ij}^b are chosen according to the atom types of the two atoms i and j that are involved in the covalent bond.

In the above equations [73], the parameters $C_{ij}^{(12)}$ and $C_{ij}^{(6)}$ are again atom type-dependent parameters. The Lennard-Jones potential is often rewritten in the form

$$U_{LJ}(r_{ij}) = 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right), \quad (3.6)$$

dependent on the two parameters ϵ_{ij} , the depth of the potential minimum, and σ_{ij} , the position of the intercept of zero, which again have to be fitted to the results of *ab initio* calculations (Figure 3.2).

The Lennard-Jones potential has an attractive and a repulsive component, the latter of which becomes dominant when the distance between the two interacting atoms undercuts the sum of their van der Waals radii. With increasing distance, the Lennard-Jones potential converges against zero. Therefore, unlike the second important non-bonded interaction in classical force fields, the electrostatic Coulomb interaction, the Lennard-Jones interaction can typically be cut off at a certain distance without having to correct for the long-range portions separately.³

U_b and U_{LJ} represent examples for bonded and non-bonded terms in classical force fields which are simple functions of the distance between two particles. However, not all of the energy terms in classical force fields are dependent on only two particles. The oscillation of an angle between two covalent bonds is described by a three-body potential, and the fluctuation of torsion (or dihedral) angles is computed from a four-body potential. More complicated many-body potentials are typically not considered [19].

³In certain scenarios, however, the effect of the truncation can be non-negligible, e.g. on the gas-liquid phase transition, and has to be corrected [74].

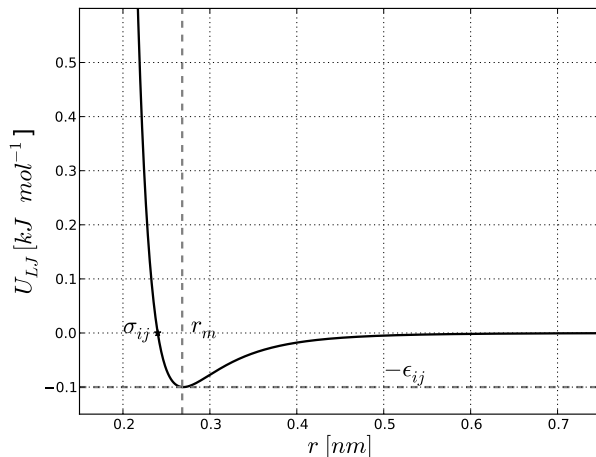


FIGURE 3.2: Plot of a Lennard-Jones potential as used in the molecular dynamics package GROMACS for modeling the interaction energy of non-bonded uncharged atoms [73]. The depth of the potential minimum ϵ_{ij} as well as the position of the intercept of zero σ_{ij} are chosen according to the atom types of the two atoms i and j that are involved in the interaction. The optimal interaction distance r_m (with an interaction energy of $-\epsilon_{ij}$) can be computed as $r_m = 2^{1/6}\sigma_{ij}$.

The computational evaluation of these simple terms appears to be trivial, but the challenge in designing classical force fields lies in the sheer number of interacting particles. While the computation of bonded interactions is typically manageable, the brute-force calculation of all non-bonded interactions in a large system is becoming an increasingly tedious – and eventually impossible – task, given that it has to be performed millions of times during a typical molecular simulation, and considering that, in principle, each particle has a non-bonded interaction with all other particles in the system. Therefore, state of the art molecular dynamics (MD) code uses an abundance of different methods in order to accelerate the evaluation of the force field, some of which are presented below (Table 3.1).

Of course, the most valuable (and by now indispensable) tool for accelerating the force field evaluation in classical molecular simulation is parallel computing. The development of efficient and scalable algorithms for this task has seen large advancements in recent years [76]. The performance of MD simulations, however, does certainly not scale linearly with the number of processors, and is ultimately dependent on the nature and modeling of the system in question.⁴

More details concerning molecular force fields can be found in references [73] and [19]. The choice of force fields for the simulations conducted in the course of this thesis is discussed in Chapters 4.5, 5.5, 6.5, and Appendix A.4.

The next section will elaborate on a crucial step in the theory of molecular simulation: Moving from the calculation of solitary phase space trajectories of molecules to

⁴A rule of thumb for parallel MD simulations is to use not more than one processor per 1,000 particles.

Method	Description
Tabulated potentials	Potential values are precomputed for the prevalent coordinate range and looked up in a table during the simulation.
Periodic boundary conditions	The use of periodic boundary conditions allows to mimic the properties of infinitely large systems (commonly solvent boxes) by creating virtual periodic images at a comparatively low cost.
Cut-off for non-bonded interactions	Only particles within a certain distance of each other are considered to interact directly. Long-range interactions (including interactions with periodic images) are treated separately, e.g. by particle-mesh Ewald [75].
Neighbor lists	The neighbor lists of directly interacting particles are updated only every n integration steps.
Constraints	By constraining the length of covalent bonds, the high-frequency bond stretching potential does not have to be calculated explicitly. This allows for a larger integration step (cp. Section 3.1.4).

TABLE 3.1: A selection of methods for reducing the computational cost of force field evaluations in classical molecular simulation [73].

the calculation of statistical ensembles, a prerequisite for the prediction of macroscopic observables from molecular simulations.

3.1.2 The canonical ensemble

In the previous section, a description for the classical state of a molecular system in terms of the separable Hamiltonian $H(q, p)$ was introduced (cp. Equation 3.1). It was elaborated how the potential energy U can be approximated by a molecular force field. Using the Hamiltonian as well as the according equations of motion

$$\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q} \quad (3.7)$$

one can now, given an initial state $(q^{(0)}, p^{(0)})$ is at hand, calculate the the dynamic evolution of the system over a certain time in terms of the phase space trajectory $(q^{(t)}, p^{(t)})$. This calculation will yield precise and deterministic information regarding the dynamics of the isolated system, however biased by a strong dependency on the chosen starting conditions $(q^{(0)}, p^{(0)})$ [19].

In practically every experimental setup, the molecular system under observation is embedded into some kind of larger environment with which it is interacting. Typically, this experimental environment dictates certain macroscopic properties, such as temperature and pressure, and thus has a direct influence on all of the microscopic states of the system. Consequently, studying isolated molecular systems in terms of deterministic phase space trajectories is often not very meaningful. Therefore, using the theoretical

framework of statistical thermodynamics, it is common practice to embed the molecular system under observation within a formulation that allows for (i) fixing certain macroscopic environmental properties in terms of constraints and (ii) predicting the values of macroscopic observables from the microscopic states in a reliable manner [19].

Thus, in contrast to monitoring a single molecular system with one possible state at a time, one now considers a large number of identical systems (all obeying the same Hamiltonian), each of which in a different state, and summarize these in terms of a statistical ensemble. A statistical ensemble can also be thought of as large – and possibly infinite – number of realizations of a random experiment [44, 77]. In this context, it is helpful to define the terms *microstate* and *macrostate*.

A microstate corresponds to a detailed description of the molecular system on the atomic level by a state γ in phase space. Due to its microscopic scale, a microstate generally cannot be determined by experiment. The term macrostate, by contrast, refers to thermodynamic variables of the molecular system that in principle are measurable, such as temperature T , volume V , pressure P , number of particles N , or inner energy E .

Typically, every macrostate of the system can be realized by a high number of microstates, which implies that, if a certain macrostate is given, the microstate is, to a large extent, undefined [19]. Therefore, a statistical ensemble is described by a probability density $\varrho(q, p, t)$ in phase space Γ ,

$$\int_{\Gamma} \varrho(q, p, t) dq dp = 1 \quad (3.8)$$

which assigns a probability measure to every microstate of the statistical ensemble. In the context of molecular simulation, one is typically interested in those ϱ where ϱ is dependent on (q, p) solely by means of H . If $\varrho = \varrho(H, t)$ with $H = H(q, p)$, it can be shown that $\partial\varrho/\partial t = 0$, which implies that the probability density ϱ is time-invariant. This corresponds to a stationary ensemble, i.e. a molecular system in thermodynamic equilibrium [78].

The quantities of interest are now the expected values of observables over the statistical ensemble. An observable is any function $A : \Gamma \rightarrow \mathbb{R}$ that assigns a real number to every microstate (q, p) of phase space. Examples for relevant observables are the potential energy U , or the kinetic energy K , but also geometric or structural properties, such as the value of a particular torsion angle, or the degree of membership with regard to a certain metastable state [77].

Finally, enforcing different macroscopic environmental constraints on ϱ leads to different types of statistical ensembles. A selection of widely used ensembles in statistical thermodynamics is listed in Table 3.2. The most commonly used statistical ensemble in classical molecular simulation is probably the canonical (or NVT) ensemble, which fixes

Name of ensemble	Constraints	Description
Microcanonical (NVE)	Constant number of particles Constant volume Constant total energy	Thermally isolated
Canonical (NVT)	Constant number of particles Constant volume Constant temperature	Thermally coupled, exchange of energy between the systems
Isobaric-isothermal (NpT)	Constant number of particles Constant pressure Constant temperature	Thermally and pressure coupled, exchange of energy between the systems
Grand canonical, or macrocanonical	Constant temperature Constant volume	Thermally coupled, exchange of energy and particles between the systems

TABLE 3.2: Different types of ensembles in statistical thermodynamics [74]. In classical molecular simulation, most experimental setups can be mimicked by either the canonical or the isobaric-isothermal ensemble.

the number of particles N , the volume V and the temperature T of the system under observation. In the canonical ensemble, one can imagine the N -particle system of volume V being connected to an infinitely large thermal bath with a constant temperature T . While the mean kinetic energy of the system remains constant in the limit, it can now constantly exchange energy with the surroundings. The canonical ensemble can also be envisioned as a distribution over microcanonical ensembles, thermally isolated systems with constant total energy E , which means that the average total energy $\langle E \rangle$ of the canonical ensemble is likewise constant.

The invariant probability density describing the distribution of phase space microstates (q, p) in the canonical ensemble is given by the Boltzmann distribution:

$$\pi(q, p) = \frac{1}{Z} \exp(-\beta H(q, p)), \quad (3.9)$$

where the *partition function*

$$Z = \int_{\Gamma} \exp(-\beta H(q, p)) dq dp \quad (3.10)$$

serves as a normalization factor that is necessary to turn π into a probability distribution. Since Z (from the German word *Zustandssumme*, “sum of states”) in this case is an integral over all possible states of a $6N$ -dimensional system, it can be calculated analytically only for the most simple molecular systems [19].

The temperature T of the thermal bath, given in Kelvin, enters the equation by means of the inverse temperature $\beta = 1/k_B T$, where k_B is Boltzmann’s constant. Due to the contact with the virtual thermal bath, there is in theory no limitation to the total energy of an individual microstate of the system. Consequently, at least in principle, every microstate in phase space is accessible. For increasing total energy, however, the

probability to actually find a certain microstate decreases exponentially. Therefore, the overwhelming majority of statistical weight is represented by regions of phase space that have a comparatively low total energy. It is these regions which almost exclusively enter into Z . Of course, the actual distribution of π on different energy levels is strongly dependent on β , and the character of the energy landscape described by H .

As the Hamiltonian H in the classical case is separable into $U(q)$ and $K(p)$, substituting Equation 3.1 into 3.9 yields

$$\begin{aligned}\pi(q, p) &= \frac{1}{Z_q \cdot Z_p} \exp(-\beta U(q)) \exp(-\beta K(p)) \\ &= \rho(q) \cdot \eta(p).\end{aligned}\tag{3.11}$$

Thus, the Boltzmann distribution can be factorized into independent distributions ρ of positions and η of momenta, where the latter corresponds to the Maxwell-Boltzmann distribution [19]. In contrast to Z_q , Z_p can generally be calculated analytically for Cartesian coordinates.

For many applications in classical molecular simulation, it is sufficient to sample from the spatial distribution ρ , as one is interested mainly in observables in q . The expected value of an observable $A : \Omega \rightarrow \mathbb{R}$ is given by the integral

$$\begin{aligned}\langle A \rangle_\rho &= \int_\Omega A(q) \rho(q) dq \\ &= \frac{1}{Z_q} \int_\Omega A(q) \exp(-\beta U(q)) dq\end{aligned}\tag{3.12}$$

over the position space Ω , which, due to the high dimensionality of Ω , cannot be computed directly for systems of relevant size [19].

Let (q_1, \dots, q_n) be an independent sequence of molecule configurations generated by drawing samples from the distribution ρ . Then it follows from the law of large numbers that the sample means

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n A(q_i)\tag{3.13}$$

converges to the expected value $\langle A \rangle_\rho$ for the number of samples $n \rightarrow \infty$. Furthermore, according to the central limit theorem, with increasing n , the sampling error, i.e. the difference between the sampled distribution and ρ , decreases asymptotically in $\mathcal{O}(1/\sqrt{n})$ almost certainly [78].

This is also connected to an important requirement for “good” statistical ensembles: ergodicity. The ergodicity requirement states that the ensemble average of an observable should coincide with its time average [77]:

$$\bar{A}_\infty = \langle A \rangle_\rho.\tag{3.14}$$

In order to implement ergodicity, the time evolution of the system needs to be mixing. A more precise definition of the term follows in the next section.

Finally, all that remains to be done for predicting the relevant observables of the system is to generate a sufficient number of samples from ρ . Unfortunately, in practical molecular simulation, the rough energy landscape inherent to most molecular system prevents the efficient generation of samples: The sampling typically remains confined to basins of low potential energy containing almost invariant microstates for a long time, while high energy transition states leading to different regions of conformational space are avoided. One way to address this problem is to use the Markov chain Monte Carlo (MCMC) approach. MCMC generates samples from a probability distribution ρ by constructing an ergodic Markov chain that has ρ as its unique stationary distribution.

3.1.3 Markov chain Monte Carlo

MCMC methods are a class of numerical methods, which, based on a stochastic process, create a dependent sequence $(q^{(n)})$ of random vectors $q^{(n)} \in \Omega$ distributed according to ρ and satisfying the Markov property:

$$P(X_k = q^{(k)} \mid X_{k-1} = q^{(k-1)}, \dots, X_0 = q^{(0)}) = P(X_k = q^{(k)} \mid X_{k-1} = q^{(k-1)}).$$

Due to the manner of its construction, the resulting Markov chain then has ρ as its unique stationary distribution for $n \rightarrow \infty$. The original MCMC algorithm, published in 1953 by Metropolis *et al.* [79], was developed for studying the diffusion of neutrons in fissionable material.

The most remarkable property of the MCMC algorithm is its ability to generate a Markov chain which has ρ as its unique stationary distribution without having to state ρ explicitly, in particular without knowledge of the partition function Z_q .

To achieve this, the transition from q to a successor state \tilde{q} is split into a trial step and an acceptance step. In the latter, the newly proposed trial state \tilde{q} is accepted with a probability of P_{acc} and rejected (in favor of remaining in state q) with $1 - P_{acc}$:

$$P(q \rightarrow \tilde{q}) = P_{try}(q \rightarrow \tilde{q}) \cdot P_{acc}(q \rightarrow \tilde{q}). \quad (3.15)$$

During the trial step, it must be ensured that every state $q \in \Omega$ is reachable within a finite number of steps. If, in addition, one chooses a symmetric trial probability $P_{try}(q \rightarrow \tilde{q}) = P_{try}(\tilde{q} \rightarrow q)$, which means that proposing state \tilde{q} in state q is equally likely to proposing state q in \tilde{q} , one obtains

$$\frac{P_{acc}(q \rightarrow \tilde{q})}{P_{acc}(\tilde{q} \rightarrow q)} = \exp(-\beta\Delta U). \quad (3.16)$$

The probability of accepting a state is then defined according to the *Metropolis criterion*

$$\begin{aligned} P_{acc}(q \rightarrow \tilde{q}) &= \min \left\{ 1, \frac{\rho(\tilde{q})}{\rho(q)} \right\} \\ &= \min \{ 1, \exp(-\beta\Delta U) \}. \end{aligned} \quad (3.17)$$

Due to the fact that the probability distribution ρ is given within a quotient, the problematic partition function Z_q cancels out in the acceptance step and leaves only a ratio of probabilities that is easy to compute.

The resulting Markov chain is *irreducible* because in the trial generation algorithm, every position state is, in principle, reachable from any other, i.e. all position states are able to communicate. Due to the possibility of rejecting a trial, the Markov chain is also *aperiodic*. A Markov chain that is both irreducible and aperiodic is ergodic. Therefore, the unique stationary distribution exists [77].

It has to be mentioned that the sequence of samples generated by MCMC methods does not represent a phase space trajectory (or even a time series), and can thus not interpreted in terms of a dynamic process.

The main problem of all MCMC methods is that the proposed trial state \tilde{q} has to be substantially different from its predecessor q , and yet has to be accepted with a high probability in order to achieve efficiency. Thus, \tilde{q} has to be as far away from q as possible for the algorithm to cover a large amount of Ω in a short time, while, simultaneously, $U(\tilde{q})$ should be smaller than or equal to $U(q)$. Therefore, the efficiency of any sampling strategy based on the Metropolis algorithm is largely dependent on two quantities, which are (i) the computational cost of the trial step, and (ii) the average acceptance probability [77].

MCMC works wonders in situations where the trial state can be generated quickly. This applies to systems that intrinsically have a low dimensionality, or are modeled in a coarse-grained fashion that allows for an easy guess of trial states. In Chapters 1.4 and 2.2, respectively, examples were presented where MCMC was applied successfully for sampling the (atomistic) conformational space of flexible PEG spacers [29] and for sampling the (non-atomistic) conformational space of a simplified model of bivalent ligand-receptor binding [30].

Obviously, for molecular systems with a large number of particles, generating a completely random (and thus almost certainly unphysical) position state for the trial step will not produce satisfying results. The so-called hybrid Monte Carlo (HMC) [80, 81] approach aims at overcoming this problem by calculating a short phase space trajectory that originates in the current state q , and ends with the trial state \tilde{q} . Thus, by combining MD (in the microcanonical ensemble) with the MCMC approach, the HMC algorithm aims at generating trial states that by design have a high probability of being accepted.

While this idea is elegant, it suffers from the fact that (i) generating the trial step is computationally expensive and potentially inefficient for technical reasons⁵ and (ii) for larger numbers of interacting particles – and certainly as soon as explicitly modeled solvent is involved – the acceptance probability deteriorates rapidly due to pronounced energy fluctuations.

Due to the limitations of Monte Carlo methods with regard to the application to large systems, a different approach is required. In the next section, it will be discussed how conventional molecular dynamics can be enhanced such that it is enabled to sample from the canonical ensemble as well.

3.1.4 Molecular dynamics

Molecular dynamics (MD), which in contrast to the MCMC class of methods is a deterministic approach, derives the motion of a molecular system over time by solving the Hamiltonian equations of motion by numerical integration. Within the limitations of classical mechanics, MD will reproduce the correct physical behavior of a molecular system, although a certain error stemming from numerical integration as well as inaccuracies in the force field has to be expected. Furthermore, one should be aware that small perturbations of the initial state can lead to dramatically different trajectories, i.e. the initial value problem in MD is ill-conditioned [99]. Nonetheless, unlike sequences of samples generated by MCMC, MD trajectories can be interpreted directly in terms of a physical process [19].

An analytic solution of the equations of motion (cp. Equation 3.7) is known only for the most basic molecular systems. Therefore, it is common practice to employ numerical integrators for this task. Intuitively, an ideal numerical integrator should conserve the total energy of the system exactly: A motion on the falling flank of the force field is accelerated, whereas a motion against the gradient $-\nabla U(q) = F$ is decelerated. This corresponds to the conversion of potential energy into kinetic energy, and vice versa, while the total energy remains constant [82].

Due to the limitations of machine precision and trade-offs made in order to achieve good performance, energy conservation in numerical integration is typically only required to be “good enough”, i.e. the total energy of the system should be conserved on average. Furthermore, for molecular simulation of Hamiltonian systems, the integrator should be symplectic (preserving the volume of phase space Γ) and time-reversible (applying the inverse impulse leads back to the previous state) [19].

A popular integrator that is both symplectic and time-reversible is the velocity Verlet

⁵Parallel computing of very short trajectories tends to be inefficient, because initiating the communication between the processors (e.g. via a message-passing interface such as MPI) creates significant computational overhead.

integrator [74], which, like the Euler integrator and other Verlet-type integrators, is derived from a Taylor expansion of the trajectory $q(t)$. Verlet integrators are based on a second-order Taylor approximation in which the third-order terms cancel, which is why one arrives at an error order of $\mathcal{O}(\tau^3)$ without having to compute any expensive third-order terms [19]. The velocity Verlet integrator updates the position q and the velocity $v = \dot{q}$ according to the equations

$$\begin{aligned} q(t + \tau) &= q(t) + \tau\dot{q}(t) + \frac{\tau^2}{2}\mathcal{M}^{-1}F(t), \\ \dot{q}(t + \tau) &= \dot{q}(t) + \frac{\tau}{2}\mathcal{M}^{-1}(F(t) + F(t + \tau)), \end{aligned} \tag{3.18}$$

where τ denotes the length of the integration step and F the force acting on q , corresponding to $-\nabla U(q)$. Given some initial state $(q^{(0)}, \dot{q}^{(0)})$, the trajectory describing the motion of the system over time is then generated by repeatedly applying Equations 3.18. The time step length τ is typically chosen on the order of 1 fs = 10^{-15} s, so as to cover even the most high-frequent events – bond length oscillations – with a sufficient number of integration steps. Constraining bonds length with the help of constraint solving algorithms such as SHAKE [83] or LINCS [84] allows for increasing τ to up to 2 fs and thus leads to an improved overall simulation performance.

Whereas MCMC methods are able to sample from the canonical ensemble, the basic form of MD introduced above only samples from the microcanonical ensemble. The microcanonical ensemble fixes the number of particles N , the volume V and the total energy $E = H$ of the system as constants (cp. Table 3.2). The latter constraint causes a problem, because, as states of constant energy are not necessarily connected, an MD trajectory in the microcanonical ensemble is not an ergodic (mixing) Markov chain [77], meaning that even in a hypothetical simulation trajectory of infinite length, only a subsection of phase space is ever sampled.

The main advantage of MD in this “pure” form is the ability to reproduce the true dynamics of a molecular system given some initial state $(q^{(0)}, p^{(0)})$, which makes it a valuable tool for calculating time-dependent quantities or transport coefficients in isolated systems [19].⁶ However, due to the high demand for predicting the properties of *coupled* molecular systems (cp. Section 3.1.2), that is dictated by the reality of most experimental setups, a lot of effort has been put into extending the MD approach so as to allow for canonical sampling.

⁶Even in this scenario, longer MD calculations suffer from a high degree of error amplification that stems from errors in numerical integration.

3.1.5 Thermostats

In 1980, Andersen [85] first suggested that MD could generate other ensembles than the microcanonical one in order to better mimic experimental conditions, in particular the canonical ensemble for thermally coupled systems.⁷ His proposal for generating the canonical ensemble, NVT , from an MD run was to randomly pick a particle during the simulation and adjust its velocity to the appropriate Maxwell-Boltzmann distribution for the desired temperature (cp. Equation 3.11). While being formally correct, Andersen's thermostat suffered from a supposedly poor efficiency, and “the fact that discontinuities in the trajectories were introduced.” [86] One of the main problems of Andersen's thermostat is that, unlike MD in the NVE ensemble, which conserves the total energy E , there is no conserved quantity for evaluating the numerical stability of the simulation. A by now more popular thermostat was introduced by Nosé [87] and reformulated by Hoover [88]. The Nosé-Hoover thermostat allows for the control of temperature without the use of random numbers. For this purpose, the Hamiltonian of the system is “extended by introducing a thermal reservoir and a friction term in the equations of motion. The friction force is proportional to the product of each particle's velocity and a friction parameter, ξ . This friction parameter (or ‘heat bath’ variable) is a fully dynamic quantity with its own momentum ($p\xi$) and equation of motion; the time derivative is calculated from the difference between the current kinetic energy and the reference temperature.” [73] In contrast to Andersen's thermostat, Nosé's dynamics has a conserved quantity that can be monitored throughout the simulation [74].

While the Nosé-Hoover thermostat was developed for sampling the canonical ensemble, it can be shown to be nonergodic in certain situations, in particular for small and stiff systems. In order to compensate for this shortcoming, the Nosé-Hoover chain approach was proposed, where each of thermostats in the “chain” has its own thermostat controlling its temperature, and in the limit of an infinite chain of thermostats, the dynamics are guaranteed to be ergodic [89]. The manual for the popular MD software GROMACS (which by default uses a chain of ten Nosé-Hoover thermostats) states that “just a few chains can greatly improve the ergodicity, but recent research has shown that the system will still be nonergodic, and it is still not entirely clear what the practical effect of this [is].” [73] In any case, the introduction of chains of thermostats requires additional tuning by the user, and further complicates this temperature coupling scheme.

Another popular thermostat is that of Berendsen *et al.* [90], in which Hamilton's equations of motion are supplemented by a first-order equation for the kinetic energy. The driving force of this equation is the difference between the instantaneous kinetic energy, and its target value. According to Bussi *et al.* [86], “Berendsen's thermostat is stable,

⁷The original article, however, also contains proposals for sampling the isoenthalpic-isobaric and the isothermal-isobaric ensemble.

simple to implement, and physically appealing; however, it has no conserved quantity and is not associated with a well defined ensemble [...]. In spite of this, it is rather widely used.” Due to the fact the Berendsen thermostat suppresses kinetic energy fluctuations, the distribution of the kinetic energy as well as dependent fluctuation properties, such as the heat capacity, cannot be determined in a rigorous manner. Other ensemble averages can be calculated with sufficient accuracy, in particular for very large system, as the error scales with $1/N$, the number of particles [73].

A similar thermostat which does produce a correct NVT ensemble is the velocity-rescaling thermostat by Bussi *et al.* [86].

The velocity-rescaling thermostat is an extension of the original Berendsen method that uses a properly constructed random force in order to enforce the correct distribution for the kinetic energy. The rescaling procedure is typically distributed over a number of time steps in order to prevent abrupt and fast fluctuations in the velocities. In general, the algorithm works as follows [86]:

- (1) Evolve the system with Hamilton’s equations of motion using a symplectic integrator such as velocity Verlet (cp. Equation 3.18).
- (2) Calculate the kinetic energy.
- (3) Evolve the kinetic energy for a time corresponding to a single time step using an auxiliary continuous stochastic dynamics.
- (4) Rescale the velocities so as to enforce the new value of the kinetic energy.

The auxiliary dynamics for the kinetic energy is given in terms of (Equation 3.19)

$$dK = (\bar{K} - K) \frac{dt}{\tau} + 2 \sqrt{\frac{K\bar{K}}{N_f}} \frac{dW}{\sqrt{\tau}}, \quad (3.19)$$

where N_f is the number of degrees of freedom, and $\bar{K} = N_f/2\beta$ is the average kinetic energy at the target temperature. dW , in turn, is a Wiener process, a stochastic process that is chosen such that it leaves the canonical distribution of the kinetic energy invariant. Without the stochastic term, Equation 3.19 is reduced to that of the standard Berendsen thermostat [86]. Save for a random seed, there are no additional parameters.

Conveniently, the velocity-rescaling thermostat also has a conserved quantity, which Bussi *et al.* name the “effective energy”, \tilde{H} . For an infinitesimal time step, one finds (Equation 3.20)

$$\tilde{H}(t) = H(t) - \int_0^t (\bar{K} - K(t')) \frac{dt'}{\tau} - 2 \int_0^t \sqrt{\frac{K(t')\bar{K}}{N_f}} \frac{dW(t')}{\sqrt{\tau}}, \quad (3.20)$$

The physical meaning of the above equation is “that the fluxes of energy between the system and the thermostat are exactly balanced.” [86] For a finite time step as used in numerical integration, this compensation of the fluxes is only approximate, but nonetheless the conservation of \tilde{H} provides a measure for the accuracy – accuracy in terms of correctly representing the desired ensemble – of the states that are generated. Therefore, in current MD software, the conserved quantity is usually stored along with the other observables throughout a simulation with the velocity-rescaling thermostat [73].

The ability to generate the NVT ensemble while preserving the dynamic properties of the system makes the velocity-rescaling thermostat very attractive for practical molecular simulation. Therefore, it has been used almost exclusively for the simulations conducted in the course of this thesis. If an isobaric-isothermal scenario is required, the thermostat can be complemented by either the Berendsen barostat [90] or the Parrinello-Rahman barostat [91, 92], which both can be used to enforce a certain reference pressure on the system. The Berendsen barostat works by rescaling the simulation box vectors and coordinates at regular intervals, while the Parrinello-Rahman barostat is an extended ensemble approach similar to the Nosé-Hoover thermostat. Whereas the latter in theory produces the “true” NpT ensemble, the Berendsen barostat suffers from the same limitations as the Berendsen thermostat [73].

Stochastic dynamics as an alternative to thermostats

A different approach of mimicking a system coupled to a heat bath without using a thermostat is given by stochastic, or Langevin, dynamics. The idea of stochastic dynamics (SD) is to add a friction and a random noise term to the Hamiltonian equations of motion, therewith turning them into stochastic differential equations (hence the name):

$$\begin{aligned}\dot{p} &= -\nabla U(q) - \zeta p + \sigma \dot{W}, \\ \dot{q} &= \mathcal{M}^{-1}p.\end{aligned}\tag{3.21}$$

In Equation 3.21, ζ is the friction constant, and $\sigma \dot{W}$ is a noise process that emulates random forces conveyed by the Brownian motion exerted by the surrounding heat bath [19]. Consequently, the dynamics itself can be employed as a thermostat for the desired temperature, which however requires system-dependent tuning of the friction constant. For instance, the GROMACS manual [73] states that “an appropriate value for ζ is 0.5 ps⁻¹, since this results in a friction that is lower than the internal friction of water, while it is high enough to remove excess heat.” Consequently, when SD is used as a thermostat in connection with other kinds of solvent, an appropriate value for ζ has to be determined beforehand.

In the case of SD, it can be shown explicitly that the resulting distribution of states corresponds to the canonical distribution, and that the ensemble average corresponds to the time average (cp. Equation 3.14), which makes it a good choice for sampling observable averages in the NVT ensemble [19].

Analogously to the velocity Verlet integrator for “pure” MD (cp. Equation 3.18), Langevin/stochastic dynamics can be discretized in a symplectic manner (Equation 3.22):

$$\begin{aligned} q(t + \tau/2) &= q(t) + \tau/2 \mathcal{M}^{-1} p(t), \\ p(t + \tau) &= p(t) + \tau \nabla U(q(t + \tau/2)) - \tau \zeta p(t) + \sqrt{\tau} \sigma \mathcal{N}(0, 1), \\ q(t + \tau) &= q(t + \tau/2) + \tau/2 \mathcal{M}^{-1} p(t + \tau), \end{aligned} \quad (3.22)$$

where $\mathcal{N}(0, 1)$ is a normally distributed random variable with mean 0 and variance 1 [19]. Depending on the choice of ζ , the resulting dynamics can differ significantly from conventional MD trajectories, but will still deliver the correct averages [73].

3.2 Conformational analysis with ZIBgridfree

This section is based on (and contains content from) the following publication:

- A. Bujotzek, O. Schütt, A. Nielsen, K. Fackeldey, M. Weber: Efficient conformational analysis by partition-of-unity coupling. Accepted for publication in *J. Math. Chem.* 2013.

In the preceding section, the focus was laid on the basics of molecular simulation, and in particular on how to generate samples from the statistical ensemble that best represents the experimental conditions. This was motivated by the desire to calculate expected values for almost arbitrary observables regarding the molecular system of interest.

While the theoretical and algorithmic machinery for generating the proper ensemble is proven and tested, the sampling process in practical molecular simulation is often plagued by the aforementioned sampling problem (cp. Chapter 1.4). The sampling problem is rooted in the fact that the dynamics of molecular systems typically exhibits a distinct metastable character: Molecular systems tend to remain within an almost invariant subset of conformational space for a long time⁸, while transitions between different almost invariant subsets (i.e. conformational changes) are rarely observed events. This characteristic is due to the rough potential energy landscape inherent to most molecular systems. Basins of low potential energy, grouped around local minima, are

⁸The definition of long in this context arises from the relation to the integration time step τ (cp. Equation 3.18), which for atomistic simulations is in the order of 1 or 2 fs.

separated by high energy barriers, corresponding to conformational changes, or changes from unbound to bound state, etc. This complicates the sampling of conformational space, as MD trajectories tend to generate states from within the basin of one local minimum for a long time, while transitions between different local minima are achieved only very seldom, or not at all. This effect, often denoted as *trapping*, can lead to incomplete coverage of conformational space, and thus to insufficient statistics. This effect is particularly severe with regard to the sampling of transient regions of conformational space, e.g. in the study of ligand-receptor binding processes, as the dynamics of the system will try to avoid the energetically unfavorable (but most interesting) transition states at any cost.⁹

While, as of yet, conventional thermostated (long-time) MD remains the predominant tool in the molecular simulation community, several successful strategies for overcoming (or rather lessening) the sampling problem have been developed, including umbrella sampling [93], essential dynamics [94] and replica exchange [95].

In the following, as one example for an enhanced sampling scheme in molecular simulation, an enhanced version of the ZIBgridfree sampling algorithm [96] is presented, as it was developed in the course of this thesis. The principle of ZIBgridfree is inspired by umbrella sampling and combines an uncoupling-coupling approach with an adaptive refinement strategy in order to enable efficient and thorough sampling, even in transient regions of conformational space. In the initial uncoupling step of the algorithm, conformational space is partitioned into subsets. Each subset is sampled independently toward convergence of the correct local distribution by means of evaluating the variance-based Gelman-Rubin convergence criterion [102]. If convergence cannot be achieved (e.g. when the sampling keeps on jumping between two local minima), a refinement of the partitioning is triggered, followed by additional sampling. In the subsequent coupling step, the weighted partial densities are joined together to yield the overall Boltzmann distribution, so that the identification of conformations is reduced to a clustering problem based on the eigenstates of the overlap matrix of the partitioning. Finally, conformational weights and inter-conformational transition probabilities can be determined. The extended version of ZIBgridfree presented here broadens the scope of this sampling scheme by combining it with a standard MD software package so as to give access to the most popular and up-to-date force fields and solvent models. More details on the implementation of this version of ZIBgridfree, including the results of a number of validation runs, can be found in Appendix A.

⁹Chapter 6 will focus exclusively on the sampling of binding processes.

3.2.1 Conformation dynamics

As partitioning methods based on meshes or grids suffer from the “curse of dimensionality”, ZIBgridfree implements a meshless, function-based partitioning approach. This is motivated by the concept of conformation dynamics [97, 98], where conformations of a molecular system are defined in terms of soft-characteristic membership functions, rather than classical sets in position space (denoted as Ω below). The primary objective of conformational dynamics is to identify a set of n_C metastable conformations, defined by membership functions $\chi_1, \dots, \chi_{n_C} : \Omega \rightarrow [0, 1]$. Each point $q \in \Omega$ is assigned to each of the conformations $i = 1, \dots, n_C$ with a certain degree of membership $\chi_i(q)$. The functions χ_i are non-negative

$$\forall q \in \Omega : \chi_i(q) \geq 0, \quad i = 1, \dots, n_C, \quad (3.23)$$

and form a partition of unity

$$\forall q \in \Omega : \sum_{i=1}^{n_C} \chi_i(q) = 1, \quad (3.24)$$

which reflects that the assignment of conformations corresponds to a soft partitioning of Ω . The conformations defined on the basis of membership functions χ_i are associated with an overlapping partial density function $\tilde{\rho}_i$:

$$\tilde{\rho}_i = \frac{\chi_i(q)\rho(q)}{\tilde{w}_i}, \quad (3.25)$$

where the partition functions

$$\tilde{w}_i = \int_{\Omega} \chi_i(q)\rho(q) dq, \quad (3.26)$$

represent the corresponding thermodynamical weights. As the membership functions χ_i are, by definition, observables over the ensemble under consideration, the thermodynamic weights \tilde{w}_i are the expected values of χ_i under the joint Boltzmann distribution of positions, ρ (cp. Equation 3.11).

A central concept in ZIBgridfree is the approximation of the unknown conformation membership functions χ_i from a function basis $\phi, \dots, \phi_s : \Omega \rightarrow [0, 1]$, where the initial number of basis functions s should be chosen larger than the anticipated number of conformations. The function basis is chosen such that it has the same properties as the membership functions $\chi_1, \dots, \chi_{n_C}$, i.e. non-negativity (cp. Equation 3.23) and partition of unity (cp. Equation 3.24). Therefore, each conformation membership function χ_j can

be constructed from a convex combination of the basis functions ϕ_i [99]:

$$\chi_j = \sum_{i=1}^s \chi_{disc}(i, j) \phi_i, \quad j = 1, \dots, n_C, \quad (3.27)$$

where χ_{disc} is a row-stochastic matrix containing the linear combination factors. Analogous to $\tilde{\rho}_i$ and \tilde{w}_i in Equations 3.25 and 3.26, each of the basis function is associated with a partial density ρ_i and a thermodynamic weight w_i . In order to calculate a specific ρ_i , it is useful that each basis function ϕ_i can be formulated in terms of a modified potential energy function \tilde{U}_i as

$$\tilde{U}_i(q) = U(q) + \hat{U}_i(q) = U(q) - \frac{1}{\beta} \ln(\phi_i(q)). \quad (3.28)$$

Based on the samplings of the partial densities ρ_i associated with the basis functions ϕ_i , the subsequent cluster analysis aims at identifying both the correct number of clusters n_C , as well as the matrix χ_{disc} of linear combination factors, so as to obtain the set of membership functions χ_j by applying Equation 3.27.

As a precondition for the partitioning discussed above, a rough scheme of the relevant position space has to be given. This can be delivered in terms of a long-time MD trajectory (possibly using elevated temperature for improved coverage of position space), a targeted MD or pulling trajectory, the output of certain tools for exploring conformational space (e.g. CONCOORD [100] for protein structures) or even by manually preparing a sequence of geometries. From this *presampling* is selected a set of nodes $\{n_1, \dots, n_s\} \in \Omega$ to each of which is attached a radial basis function W_i given by

$$W_i(q) = \exp(-\alpha \delta^2(q, n_i)), \quad i = 1, \dots, s, \quad (3.29)$$

where α is a shape parameter, and δ^2 a distance measure to be specified in the next section. The basis functions W_i take on their maximum at the defining node n_i , and decrease exponentially as the distance δ^2 of a state q to n_i increases. As Equation 3.24 is not yet satisfied, one constructs a partition of unity with basis functions ϕ_i by following Shepard's approach [101]:

$$\phi_i := \frac{W_i}{\sum_{j=1}^s W_j}, \quad i = 1, \dots, s. \quad (3.30)$$

The shape parameter α is chosen in dependence on the number of nodes s and the mean node distance θ , and defines the degree of separation of the meshless discretization. A proper choice for the value of α ensures thorough sampling in the area belonging to basis function ϕ_i , as it helps to restrain the sampling process from wandering off into a lower energy basin. For $\alpha \rightarrow \infty$, the discretization converges to a Voronoi tessellation, i.e.

the soft partitioning degenerates into a hard partitioning without overlaps between the basis functions.

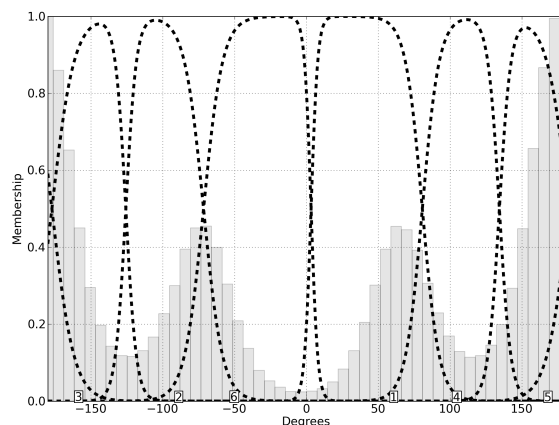


FIGURE 3.3: Soft partitioning of a single torsion angle by six overlapping basis functions (dashed black lines). The nodes (marked by white numbered labels) were picked from a high-temperature presampling distribution (light gray histogram) by using the k-means algorithm.

In practice, the sampling of the basis functions ϕ_i is run in parallel, as each \tilde{U}_i can be evaluated at every position $q \in \Omega$ independently of all \tilde{U}_j with $j \neq i$. Depending on the available resources, one can either sample several basis functions in parallel, evaluate the potential \tilde{U}_i in parallel (which in turn accelerates the sampling of the associated basis function), or combine both approaches.

3.2.2 Internal coordinates

ZIBgridfree uses internal coordinates (either torsion angles and/or distances) as collective variables in order to define the conformation of the system under observation. Prior to picking a set of nodes for discretization, a set of n_K internal coordinates has to be specified by the user. The distance $\delta^2(q, n_i)$ between state q and node n_i (Equation 3.29) is measured in the space of internal coordinates. Therefore, the outcome of the discretization is directly related to the choice of internal coordinates. Deciding on a meaningful set of internal coordinates is not always trivial. For conformational analysis of small molecules, picking all rotatable torsion angles is an obvious choice, whereas for peptides or proteins, picking only backbone torsion angles is practical. For complexes of multiple molecules, the set of torsion angles has to be complemented by a set of distances in order to describe the molecules' relative positioning to each other (Figure 3.4). Depending on the focus of the simulation, the set of internal coordinates can be reduced to a relevant subset, e.g. the part of a protein that is involved in ligand binding. In general, high-frequency torsion angles associated with very mobile atoms such as

hydrogen can be neglected, as they do not contribute to a meaningful conformational description. If only one linear coordinate is given, the ZIBgridfree discretization will resemble the discretization along a single reaction coordinate in conventional umbrella sampling.

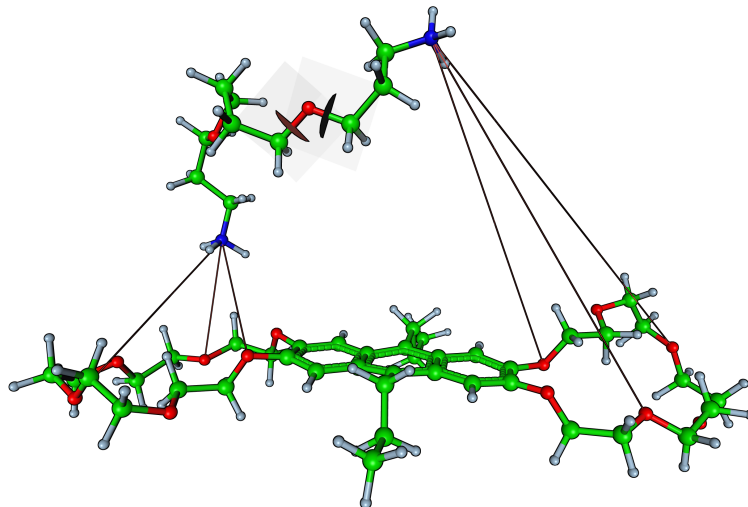


FIGURE 3.4: A set of internal coordinates for describing the state of a bivalent host-guest system. The positioning of the charged amino groups of the bivalent guest with regard to the complementary crown ethers of the host is described by a set of six linear coordinates. The linear coordinates can, for instance, be supplemented by torsion angles that describe the conformation the guest molecule.

Whereas angular internal coordinates can only take on values between $-\pi$ and $+\pi$, distance (or linear) coordinates can in principle take on any positive value. This leads to problems whenever linear coordinates with a large spread or a large absolute value are overly dominant, as other internal coordinates with more subtle changes are rendered irrelevant when the distance function δ^2 is evaluated. In order to tackle this problem, linear coordinates can be weighted and normalized automatically by calling `zgf_create_pool` with option ‘`-balance-linears`’.

Let k be a linear coordinate that corresponds to the Euclidean distance between two particles in the system under observation. The weight of this coordinate is then determined as follows:

$$\text{coord_weight}(k) = \frac{\text{coord_weight}(k)_{\text{initial}}}{\sqrt{2 * \text{var}(k)}}, \quad (3.31)$$

where $\text{coord_weight}(k)_{\text{initial}}$ is one, unless specified differently by the user. This means that coordinates with a high spread are downgraded by dividing the initial weight by the full width at half maximum. Furthermore, an offset for k is applied by subtracting its mean value in order to compensate for high absolute values. This leads to the following

weighting formula:

$$\begin{aligned} k_{balanced} &= \text{coord_weight}(k) \cdot (k - \text{offset}(k)) \\ &= \text{coord_weight}(k) \cdot (k - (\text{offset}(k)_{initial} + \text{mean}(k))), \end{aligned} \quad (3.32)$$

where $\text{offset}(k)_{initial}$ is zero, unless specified differently by the user. This approach realizes an equal weighting of all internal coordinates involved. Nonetheless, certain applications might call for biased weighting of the internal coordinates, e.g. when the distance between ligand and receptor (defined by linear internal coordinates) is to be stressed in comparison to more subtle conformational changes in the ligand molecule (defined by torsion angle internal coordinates).

3.2.3 Implementing the potential modification

Sampling the ZIBgridfree basis function ϕ_i requires a modification of the potential function $U(q)$ (Equation 3.28). The aim was to change the algorithm such that it can be run with standard force fields and unmodified MD packages such as GROMACS [76]. Treating the MD code as a black box has several advantages: The user can use readily available software (pre-compiled for many Linux distributions and pre-installed on most computing clusters), and plug in new versions as they are released. Full flexibility regarding the choice of force field and other simulation parameters is sustained. Furthermore, internal changes to the highly optimized MD code, possibly having a negative impact on the simulation performance, are evaded.

Adapting ZIBgridfree to a standard MD package is a two-step procedure. First, for each selected node n_i , the n_K -dimensional ϕ_i function is projected on a single dimension by coordinate-wise evaluation: Instead of considering the joint distance $\delta^2(q, n_i)$ involving all internal coordinates, one now considers exclusively the distance regarding coordinate k :

$$\phi_{i_k}(q) := \frac{\exp(-\alpha \delta_k^2(q, n_i))}{\sum_{j=1}^s \exp(-\alpha \delta_k^2(q, n_j))}, \quad k = 1, \dots, n_K. \quad (3.33)$$

The above expression yields the membership of state q with respect to coordinate k regarding basis function ϕ_i . The one-dimensional penalty potential acting on coordinate k of state q can simply be obtained as:

$$\hat{U}_{i_k}(q) = -\frac{1}{\beta} \ln(\phi_{i_k}(q)). \quad (3.34)$$

Finally, in order to approximate \tilde{U}_i , for every internal coordinate k , a generic cubic restraint potential (as available in many common MD packages) is fitted to the penalty potential \hat{U}_{i_k} and added to the force field representing the unmodified potential U . This approach was implemented for the GROMACS MD package, where restraint potentials of the form

$$U_{res}(\Phi') = \begin{cases} \frac{1}{2}k_{res}(\Phi' - \Delta\Phi)^2, & \text{for } \Phi' > \Delta\Phi \\ 0, & \text{for } \Phi' \leq \Delta\Phi \end{cases} \quad (3.35)$$

are readily available (given here for a torsion angle restraint on torsion angle $\Phi' = (\Phi_0 - \Phi) \bmod 2\pi$, with rest position Φ_0 and unrestrained region $\Delta\Phi$, analogous for distance restraints). The concept of fitting restraint potentials to the coordinate-wise projected basis function penalty potentials of ZIBgridfree is depicted in Figure 3.5.

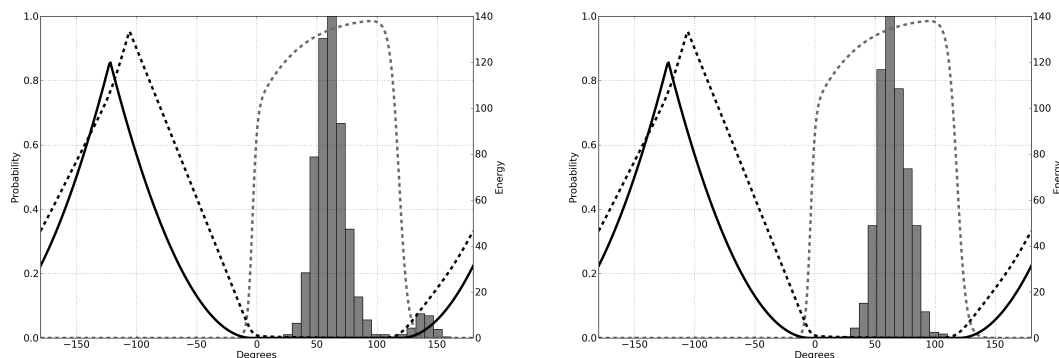


FIGURE 3.5: Sampling of a torsion angle distribution (gray histogram) with ZIBgridfree. The sampling is forced to stay within the area of an exemplary basis function (dashed gray line) by its penalty potential (dashed black line). For use with GROMACS, the penalty potential is approximated by a harmonic restraint potential (solid black line). Due to the approximation error, the sampling is not sufficiently limited to the area covered by its basis function (left). After reweighting the sampling points with regard to their basis function (right), the approximation error is removed.

The imperfect approximation of multi-dimensional basis functions by harmonic restraints introduces a certain error, as sampling points may be generated from areas of Ω that are not covered by the basis function in question. This is especially true for boundary regions, where several basis functions are overlapping. This approximation error can be removed by giving each sampling point q a weight $\mathbf{frame_weight}_i(q)$ with respect to basis function ϕ_i :

$$\mathbf{frame_weight}_i(q) = \frac{\phi_i(q)}{\exp(-\beta \cdot U_{res}(q))} \quad (3.36)$$

The effect of reweighting on the sampling distribution is depicted in Figure 3.5. Calculating the sampling point weights is inexpensive in terms of computation time. Subsequently, when checking for convergence of the sampling, or when calculating observables of any kind, only the reweighted distribution is considered.

3.2.4 Adaptive refinement of the partitioning

In order to ascertain a sufficient sampling of the partial densities ρ_i , ZIBgridfree pursues an adaptive refinement approach. After a certain number of simulation steps, convergence of the sampling is tested by evaluating the Gelman-Rubin convergence criterion [102]. If the convergence test fails, the sampling will be extended by n simulation steps (followed by another convergence test) for a maximum of m times (where n and m are user-defined settings). If convergence has not been achieved after m extensions of the original sampling length, a refinement of the partitioning in the area of the affected basis function is triggered. By default, two children nodes n_{i_1} and n_{i_2} are introduced, whereas the original parent n_i is removed from the partitioning, along with its basis function ϕ_i . This principle is illustrated in Figure 3.6.

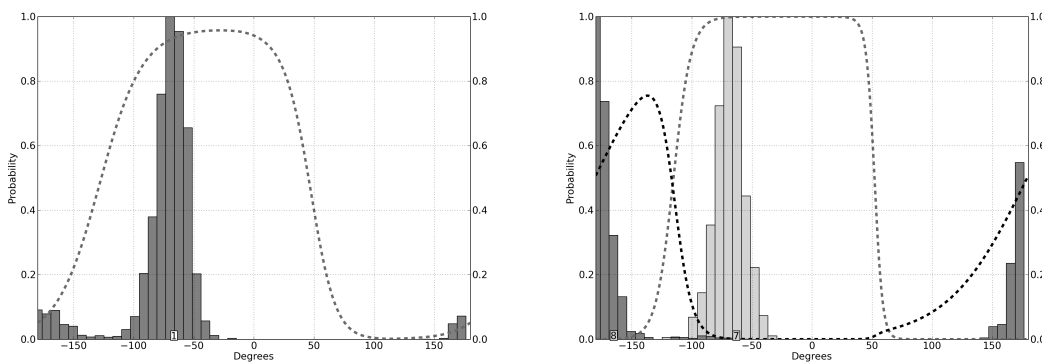


FIGURE 3.6: The sampling of basis function '1' (associated with a node at -68°) has come upon a second minimum in the region around -180° (left). In this case, convergence of the sampling is not achieved in the allocated number of sampling steps. A failed convergence test triggers an automatic refinement of the partitioning (right). The parent node '1' is removed and replaced by two children named '7' (-65°) and '8' (-167°). The samplings of the associated basis functions converge quickly, as they are now confined to a single energy minimum each.

Removal and addition of nodes have an impact on the overall partitioning, as with the number of nodes s , the mean node distance θ is bound to change. Hence, following each refinement step, the shape parameter α (Equation 3.29) is recalculated. With proceeding refinement and increasing s , α will become larger, which in turn leads to a higher degree of separation between basis functions. This mechanism leads to increased convergence rates over the course of the refinement.

Despite several cycles of refinement, the sampling of transition regions (e.g. when a node is situated on the steep flank of a potential energy barrier) may not lead to convergence according to the Gelman-Rubin criterion. In these cases, the sampling has to be discontinued as soon as a sufficient number of data points from the transition region has been collected. This can be achieved by calling `zgf_mdrun` with the option ‘`–multistart`’: Instead of prolonging the sampling in terms of trajectory extensions, this option will trigger multiple restarts from the initial position with varying starting impulse – an approach that is indispensable in the sampling of binding processes.

3.2.5 Reweighting and cluster analysis

Direct free energy reweighting

As the partial densities associated with the discretization nodes are sampled independent of each other, they may represent an accurate local distribution of states, but do not have the correct relative weighting. This leads to an error when the joint Boltzmann distribution is accumulated from the individual partial densities, as is illustrated in Figure 3.7 for the torsion angle distribution of *n*-pentane. In order to arrive at a balanced joint Boltzmann distribution, the partial densities have to be reweighted according to a free energy difference estimate implemented in the tool `zgf_reweight`. The free energy difference estimate, based on the approach of Klimm *et al.* [103], is outlined shortly in the following.

1. From each set of states $\{q_n^{(i)}\}_{n=1,\dots,N^{(i)}} \in \Omega$ representing the partial density ρ_i , $i = 1, \dots, s$, choose a set of reference points $\{q_r^{(i)}\}_{r=1,\dots,R^{(i)}}$. A reference point is characterized by having a potential energy value within the energy standard deviation of ρ_i . More precisely, with $\langle U^{(i)} \rangle$ being the mean potential energy of set $q^{(i)}$,

$$\left\| U(q_r^{(i)}) - \langle U^{(i)} \rangle \right\| \leq \sqrt{\frac{1}{N^{(i)}} \sum_n^{N^{(i)}} \left(U(q_n^{(i)}) - \langle U^{(i)} \rangle \right)^2}.$$

2. Approximate the local density of sampling points by evaluating expression D_{vol_i} , which counts the number $N_{near}^{(i)}$ of sampling points that are *near*, i.e. within a certain distance vol_i around each reference point $q_r^{(i)}$, and compute its inverse

$$\left(D_{vol_i}(q_r^{(i)}) \right)^{-1} \approx \frac{N^{(i)}}{N_{near}^{(i)} + 1}.$$

For this purpose, vol_i is chosen as large as the mean variance of the internal coordinates regarding all sets of states $q^{(i)}$, which is precomputed in a first iteration over the sampling data. The variance for each set is computed in terms of the distance function δ^2 , dependent on the type of the internal coordinates that are involved in the discretization.

3. Compute the entropy estimate

$$S_i = k_B \ln \left(\frac{1}{R^{(i)}} \sum_{l=1}^{R^{(i)}} \left(D_{vol_i}(q_r^{(i)}) \right)^{-1} \right),$$

the free energy

$$G_i = \langle U^{(i)} \rangle - T \cdot S_i,$$

and the statistical weights

$$w_i = w_{i-1} \cdot \exp(-\beta (G_i - G_{i-1})),$$

with $w_1 = 1$. The free energy values have to be ordered by size before calculating the statistical weights. Finally, the statistical weights have to be normalized so that $\sum_{i=1}^s w_i = 1$.

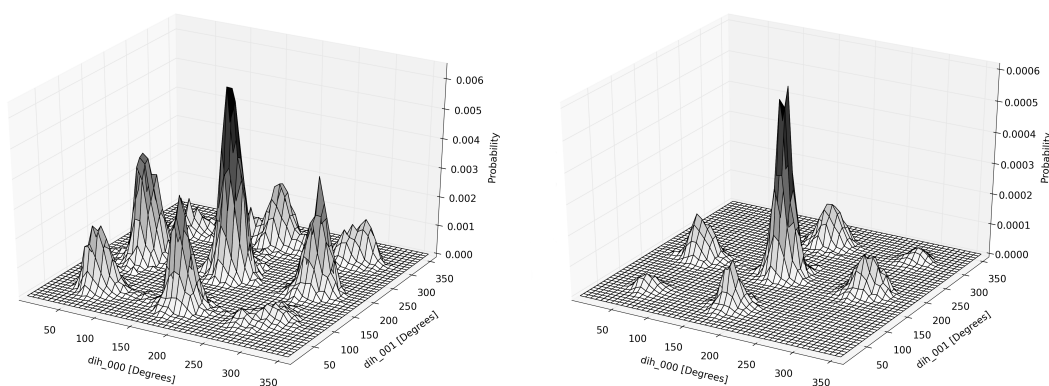


FIGURE 3.7: Torsion angle distribution of the two torsion angles of n -pentane at 300 K, assembled from 25 individual node samplings. Before reweighting, each partial density contributes equally to the joint distribution (left). This leads to disproportionately high weights of the gauche/trans, trans/gauche and gauche/gauche conformations. After thermodynamic reweighting, the correct relative weights of the partial densities are restored, which leads to an improved joint distribution (right).

Overlap weight correction

The reweighting method introduced in the previous section works best for well-separated basis functions. Depending on the given discretization and the nature of the system under observation, the basis functions in ZIBgridfree can have a more or less pronounced overlap. Therefore, in order to take basis function overlap into account, a correction of the statistical weights w is necessary. The degree of overlap between each pair of basis functions ϕ_i and ϕ_j is quantified in terms of the overlap integral matrix $S \in \mathbb{R}^{s \times s}$:

$$S_{ij} = \int_{\Omega} \phi_i(q) \phi_j(q) \rho(q) dq, \quad (3.37)$$

which for practical reasons is approximated as

$$S_{ij} = \frac{1}{N^{(i)}} \sum_{n=1}^{N^{(i)}} \phi_j(q_n^{(i)}) \cdot \mathbf{frame_weight}_i(q_n^{(i)}) \quad (3.38)$$

from the states $\{q_n^{(i)}\}_{n=1, \dots, N^{(i)}}$ that represent the partial density ρ_i . Note that the shape of S is influenced by the chosen discretization, in particular by the number of discretization nodes s . For fine discretizations (large α , cp. Equations 3.29 and 3.30), S will resemble a diagonal matrix. For very coarse discretizations and small α , it will degenerate into a full matrix.

The statistical weights w of the basis functions can be derived by solving the eigenvalue problem $w^\top S = w^\top$, which means that w corresponds to the unique, positive and normalized left eigenvector of S with regard to its eigenvalue $\lambda_1 = 1$ [99]. This eigenvector-based approach is not well-conditioned and highly dependent on sufficient sampling in the overlap regions between the basis functions [104]. In order to benefit from the advantages of both direct free energy reweighting and the eigenvector-based approach, a number of power iteration steps from the original weights w with the stochastic matrix are started, until the corrected weights (again denoted as w) are convergent.

After the correction of w , an adjustment of S is required. If the matrix S has been computed according to Equation 3.37, it is usually not symmetric. Furthermore, the row sums do not correspond to the corrected weights w . According to the method of Sinkhorn[105], an iterative rescaling of the row sums to meet w , followed by a symmetrization of S , leads to a corrected overlap integral matrix that is consistent with the precomputed statistical weights. After this correction, the vector w is a left eigenvector of $D^{-1}S$ corresponding to the eigenvalue $\lambda_1 = 1$, where $D = \text{diag}(w)$. In the following, the stochastic matrix $D^{-1}S$ is used for the identification of metastable states.

Metastability analysis with PCCA+

From the chemical perspective, metastable subsets correspond to the main conformations of the underlying molecular system. In the presence of metastable states, any matrix describing the transition behavior of the system (including the corrected overlap integral matrix $D^{-1}S$) exhibits a virtual block-diagonal structure, i.e. there exists a permutation of indices so that the metastable subsets of the system are represented by (more or less) quadratic blocks along the diagonal of the matrix (see Figure 3.8).

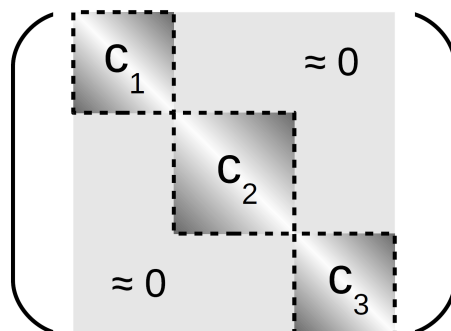


FIGURE 3.8: Schematic of a (permuted) transition matrix in the presence of metastable subsets. Within the three conformations c_1 to c_3 , states are mixing quickly. By contrast, transitions from conformation to conformation (light gray off-diagonal area) are rare events.

Every block in this matrix is associated with an eigenvector of the matrix whose eigenvalue is almost one. The set of the eigenvalues in the vicinity of one is denoted as the Perron cluster, and the size of this set corresponds to the number of chemical conformations n_C . As the eigenvector space represents a basis of the transition matrix, every conformation can be stated by a linear combination of these eigenvectors. The linear combinations of the eigenvectors associated with the eigenvalues of the Perron cluster contain, for each basis function ϕ_i , the degree of membership with regard to each of the n_C conformations. Robust Perron cluster analysis (PCCA+) [106, 107] is used to find the permutation yielding the block-diagonal structure, and hence the matrix of linear combination factors χ_{disc} (cp. Equation 3.27). The result is the matrix $\chi \in \mathbb{R}^{s \times n_C}$, where the entry $\chi(i, j) \in [0, 1]$ denotes the degree of membership of basis function ϕ_i with regard to the j -th metastable subset.

Using the weight vector w containing the thermodynamic weights of the basis functions ϕ_i , it is then possible to calculate the weights \tilde{w} of the conformations as $\tilde{w} = \chi^\top w$.

An estimate for the precision of the conformational weights that can be expected from a conformational analysis with ZIBgridfree can be made based upon the validation results presented in Appendix A.

3.3 Conformational entropy estimation

Chapter 2.2.2 introduced the concept of entropy as the logarithmic counting of the states of a given system. A high multiplicity of states $q \in \Omega$ is equivalent to a high entropy of the system. The absolute entropy of the position space Ω of a molecular system described in terms of a canonical ensemble can be expressed by the integral (Equation 3.39):

$$S_{\Omega} = -k_B \int_{\Omega} \ln(\rho(q)) \rho(q) dq, \quad (3.39)$$

where the probability that the system adopts a certain state q is given by the spatial Boltzmann distribution (Equation 3.40, also cp. Equation 3.11) [108]:

$$\rho(q) = \frac{\exp(-\beta U(q))}{\int_{\Omega} \exp(-\beta U(\bar{q})) d\bar{q}}. \quad (3.40)$$

For $q \in \mathbb{R}^{3N}$ as obtained from a molecular simulation, the integral given in Equation 3.39 does not separate the conformational entropy from the rotational and the translational component, i.e. $S_{\Omega} = S_{conf} + S_{trans} + S_{rot}$. Therefore, if only the conformational component is required, the rotational and translational degrees of freedom have to be eliminated, e.g. by shifting and aligning the principal axes of the molecule throughout the sequence of sampled states. Alternatively, it is possible to define representations of the conformational state that are inherently invariant with regard to translation and rotation (cp. Section 3.2.2 on internal coordinates/collective variables) – a thought that will be picked up later.¹⁰

Even after removing translational and rotational degrees of freedom, the position space would in principle be unlimited, if it was not constrained by the potential energy U : With increasing $U(q)$, the probability that the system can be found in the associated state q decreases exponentially. Consequently, unphysical states with a disproportionately high potential energy do not occur. Therefore, if a high number of states is accessible on a moderate energy level – with regard to the Boltzmann distribution for a given temperature – the entropy of the system (in its current macrostate) is favorable. If only a small number of states is accessible, the entropy is unfavorable. In this case, the system is driven towards a macrostate with a higher entropy, unless it is compensated by an adequate amount of enthalpy gain.

In the study of molecular systems, one is often interested in calculating entropy differences between certain macrostates of the system, i.e. in answering how many microstates are associated with each of the relevant macrostates. This knowledge is a key factor in

¹⁰The significance of conformational entropy in the context of multivalent ligand-receptor binding has been discussed in Chapter 2.2.2.

understanding binding affinities in ligand-receptor systems, and the missing link when calculating free energy differences between unbound and bound state with methods such as MMPB/SA, using the potential energy from a **M**olecular **M**echanics force fields, solvation electrostatics from the **P**oisson-**B**oltzmann equation, and an estimate of the hydrophobic effect¹¹ from the accessible **S**urface **A**rea [110]. Considering the aforementioned difficulties in sampling molecular (and in particular macromolecular) position space, calculating entropy differences, or even absolute entropy, in an efficient manner is among the most complex problems in molecular simulation. Not only computational, but also experimental methods struggle with quantifying conformational entropy.

While experimental measurements of free energy changes, e.g. by means of isothermal titration calorimetry (ITC), allow for separating entropic from enthalpic contributions (cp. Chapter 1.4, Table 1.1), Numata [110] states that “the separation of the total entropy change into solvent and solute components is in general not straightforward.” Nonetheless, in 2002 the conformational entropy loss of a protein backbone upon forced mechanical unfolding could be estimated by means of atomic force microscopy measurements [111], and advancements in nuclear magnetic resonance (NMR) spectroscopy allow for resolving separate microscopic states of solute molecules. Frederick *et al.* [112] and Marlow *et al.* [113], for instance, were able to estimate the change in conformational entropy of the protein calmodulin upon binding of different target domains by means of NMR relaxation analysis, which in turn could be related to the overall entropy change that was obtained by ITC measurements. Recently, a similar approach has been applied for studying conformational entropy changes in the carbohydrate recognition domain of galectin-3 upon binding of different small-molecule ligands [114]. However, determining the conformational entropy change for small solute molecules (or even for separate parts of such, e.g. for a given flexible spacer component) remains a notable challenge, and therefore, at least for now, stays a domain of simulation methods.

The most commonly used method for calculating *absolute* conformational entropy from simulation trajectories of macromolecules (such as proteins) is the quasi-harmonic approximation (QHA) [115–117]. QHA, in turn, is based on principle-component analysis (PCA), or covariance analysis, and can be used to identify linear correlations between pairs of coordinates. For this purpose, it uses the covariance matrix C of the atomic coordinates (Equation 3.41) [116]:

$$C_{ij} = \mathcal{M}_i^{\frac{1}{2}}(q_i - \langle q_i \rangle) \mathcal{M}_j^{\frac{1}{2}}(q_j - \langle q_j \rangle), \quad (3.41)$$

where \mathcal{M} is the diagonal atom mass matrix for mass-weighted analysis, or the unit matrix for non-mass weighted analysis. C is a symmetric $3N \times 3N$ matrix, which can

¹¹The hydrophobic effect is in truth also a largely entropic phenomenon, connected to the rearrangement of networks of water molecules [109].

be diagonalized with an orthonormal transformation matrix R (Equation 3.42) [73]:

$$R^{\top} C R = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{3N}), \text{ with } \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{3N}. \quad (3.42)$$

The eigenvectors of R are called principal (or essential) modes. The eigenvalue λ_i , in turn, is the mean square fluctuation of the i -th principal component. The matrix R defines a transformation to a new coordinate system, which can be used to project the trajectory unto the principal modes in order to obtain the corresponding principal components, $p_i(t)$ (Equation 3.43) :

$$p(t) = R^{\top} \mathcal{M}^{\frac{1}{2}}(q(t) - \langle q \rangle) \quad (3.43)$$

The first few principal modes often describe collective, global motions in the system, as they are constructed such that they contain the largest variance in the data. Thus, PCA is useful for filtering the large-scale dynamics of a macromolecule (which may have chemical/functional relevance) from the “noise” of small, local (and likely irrelevant) dynamics. The trajectory can be filtered along a certain principal mode i as follows (Equation 3.44) [73]:

$$q^{f_i}(t) = \langle q \rangle + \mathcal{M}^{-\frac{1}{2}} R_{*i} p_i(t) \quad (3.44)$$

Again, if Cartesian coordinates are used, the trajectory should be fitted to a reference structure beforehand in order to remove translational and rotational motion.

The main concept of QHA is to fit, for each principal mode, the associated observed probability density to a harmonic oscillator model for which the entropy, among other thermodynamic properties, can be calculated analytically [110]. The variances of this multi-dimensional Gaussian distribution are assumed to be equal to the eigenvalues λ of the diagonalized covariance matrix. For more complex molecular systems, this step is highly approximative, as it is based on the assumption that the (possibly very intricate) dynamics of the principal mode can be emulated by a harmonic potential.

The second significant approximation is made in limiting covariance analysis to linear correlation, although coordinates in molecular simulation, depending on the nature of the system under observation, were shown to correlate in a supralinear manner [110]. Consequently, QHA tends to substantially overestimate the entropy for systems with multimodal conformational densities, i.e. as soon as multiple energy minima are occupied [118]. In these cases, QHA only delivers an upper bound on conformational entropy, but not an exact value. By now, some of these shortcomings have been addressed in terms of different extensions of the original QHA algorithm in order to correct for anharmonic behavior and supralinear correlation, so that the upper bound on the exact conformational entropy value can at least be tightened. For a compact, but very rich overview of these methods, refer to Ref. [110].

Fortunately, for many applications in the context of molecular simulation, calculating *absolute* conformational entropy values can be bypassed by considering *differences* in conformational entropy instead, e.g. for thermodynamic reweighting of different regions in conformational space (cp. Section 3.2.5) [103], for quantifying the entropy loss between unbound and bound state in a ligand-receptor pair [16], or for comparing the conformational entropy of different (but structurally similar) spacer variants in a multivalent compound [119].

For instance, in order to estimate conformational entropy differences from simulations of different polymer variants presented on anti-fouling surfaces, Weber and Andrae presented a Monte Carlo-inspired approach that is based on comparing densities in conformational space, and relating the observed differences in conformational denseness to the conformational entropy difference [108]. While the original publication uses (subsets of) Cartesian coordinates, the approach was later extended to work with different types of collective variables (cp. Section 3.2.2) in order to simplify the application to solute molecules [103].

Weber and Andrae calculate the entropy difference between two canonical ensemble systems of interest at a given temperature T (Equation 3.45)

$$\Delta S = \frac{\Delta U - \Delta A}{T} \quad (3.45)$$

from estimates of the inner energy difference ΔU and the Helmholtz free energy difference ΔA . The computation of ΔU is the trivial part as it can be obtained from a Boltzmann-distributed sampling of the two position spaces Ω_1 and Ω_2 , and the associated potential energy functions $U_1(q)$ and $U_2(q)$ (Equation 3.46):

$$\Delta U = \int_{\Omega_2} U_2(q)\rho(q) dq - \int_{\Omega_1} U_1(q)\rho(q) dq = \langle U_2 \rangle - \langle U_1 \rangle. \quad (3.46)$$

What remains is to find an estimate for ΔA . The Helmholtz free energy A , the thermodynamic potential at work in the canonical ensemble, is defined as the logarithm of the (unknown) partition function Z_q (Equation 3.47, also cp. Equation 3.11):

$$A := -\beta^{-1} \ln(Z_q) = -\beta^{-1} \ln \left(\int_{\Omega} \exp(-\beta U(q)) dq \right). \quad (3.47)$$

The authors use a Monte Carlo quadrature approach in order to avoid the need to evaluate the partition function explicitly, which leads to the following expression for the

Helmholtz free energy difference (Equation 3.48):

$$\begin{aligned}
\Delta A &= -\beta^{-1} \ln \left(\frac{\int_{\Omega_2} \exp(-\beta U_2(q)) dq}{\int_{\Omega_1} \exp(-\beta U_1(q)) dq} \right) \\
&\approx -\beta^{-1} \ln \left(\frac{\exp(-\beta U_2(q_r^{(2)})) \text{vol}(\Omega_2) \frac{D_{vol_2}^{equi}(q_r^{(2)})}{D_{vol_2}(q_r^{(2)})}}{\exp(-\beta U_1(q_r^{(1)})) \text{vol}(\Omega_1) \frac{D_{vol_1}^{equi}(q_r^{(1)})}{D_{vol_1}(q_r^{(1)})}} \right) \\
&\approx -\beta^{-1} \ln \left(\frac{D_{vol_1}(q_r^{(1)})}{D_{vol_2}(q_r^{(2)})} \right) + U_2(q_r^{(2)}) - U_1(q_r^{(1)}).
\end{aligned} \tag{3.48}$$

The fraction $D_{vol_i}^{equi}(q_r^{(i)})/D_{vol_i}(q_r^{(i)})$ with $i = 1, 2$ is an estimator for the density of conformational states around a reference state $q_r^{(i)}$ that is meant to represent the conformational space Ω_i of volume $\text{vol}(\Omega_i)$.

The final step in Equation 3.48 requires that the two position spaces Ω_1 and Ω_2 are comparable (if not equal) in structure, e.g. by having an equal number of degrees of freedom. This requirement enforces an equal position space volume $\text{vol}(\Omega_i)$, and thus leads to the same number of equally distributed states to be expected, $D_{vol_i}^{equi}$. Consequently, both terms cancel out. Together with the estimate for ΔU (cp. Equation 3.46), one arrives at the following estimate for the entropy difference (Equation 3.49):

$$\Delta S \approx \frac{[\langle U_2 \rangle - U_2(q_r^{(2)})] - [\langle U_1 \rangle - U_1(q_r^{(1)})]}{T} + k_B \ln \left(\frac{D_{vol_1}(q_r^{(1)})}{D_{vol_2}(q_r^{(2)})} \right), \tag{3.49}$$

which, if one selects reference states with a potential energy that corresponds to the observed mean potential energy, i.e. $U_1(q_r^{(1)}) = \langle U_1 \rangle$ and $U_2(q_r^{(2)}) = \langle U_2 \rangle$, can be simplified further to (Equation 3.50):

$$\Delta S \approx k_B \ln \left(\frac{D_{vol_1}(q_r^{(1)})}{D_{vol_2}(q_r^{(2)})} \right). \tag{3.50}$$

In practice, it is necessary to consider significantly more than one reference state $q_r^{(i)}$ in order to converge the entropy difference estimate. The result is then based on a mean over a number of M_1 and M_2 density estimates taken at various reference positions in Ω_1 and Ω_2 (Equation 3.51):

$$\Delta S \approx k_B \ln \left(\frac{M_1 \sum_{r=1}^{M_2} \left(D_{vol_2}(q_r^{(2)}) \right)^{-1}}{M_2 \sum_{r=1}^{M_1} \left(D_{vol_1}(q_r^{(1)}) \right)^{-1}} \right). \tag{3.51}$$

Algorithmic details on the choice of reference states and the choice of volume vol_i for

evaluating the density estimate $D_{vol_i}(q_r^{(i)})$ can be found in Section 3.2.5. For a mathematical background on the Monte Carlo quadrature approach used in Equation 3.48, please refer to the original Ref. [108].

Chapter 4

Flexible spacer systems

4.1 Introduction

Chapters 1 and 2 have introduced the concept of multivalency and the idea of a multivalent enhancement effect in chemical systems, an effect that, based on the preorganization of complementary binding partners in multivalent arrays, induces a cooperative effect that helps to enhance binding affinities and allows to assemble complex geometries and architectures. Along with this introduction, certain topics of interest were pointed out as possible starting points for an investigation with theoretical methods, investigations that might help to better understand the different factors that are at work in multivalent systems. Among these topics were the question in how far conformational entropy can have an impact on multivalent interactions (a question that has, in turn, an impact on choices for the design of spacers and scaffolds for multivalent compounds), but also the desire to monitor an actual multivalent binding process on the molecular level, e.g. for looking into the occurrence of rebinding events. Accordingly, Chapter 3 focused on how chemical systems in general can be studied on the atomistic level, efficiently and purely computational, by means of classical molecular simulation.

This chapter constitutes the first “applied” part of this thesis, and as such presents the results from two studies that involve the application of different computational tools, mainly related to conformational analysis and conformational entropy estimation, to realistic multivalent systems. The first study (Section 4.2, based on Ref. [119]) is a purely computational one, however complemented with a chemical context provided by Min Shan and his supervisor Prof. Dr. Rainer Haag (for affiliations, see footnote 1), and addresses the problem of designing “successful” flexible and semi-flexible spacers for tethering two ligand moieties in different bivalent compounds meant for binding to

human estrogen receptor (ER), a known drug target that was shown to have a homodimeric receptor layout with two identical steroid binding sites (Section 4.1.1). The focus of this first study is on predicting spacer conformations and conformational entropy differences in water. The second study (Section 4.3, based on parts of Ref. [16]) represents a collaborative work of mainly experimental nature, conducted by Min Shan¹ at Freie Universität Berlin, Frank Abendroth² at Humboldt-Universität zu Berlin and Anja Wellner³ at Universität Innsbruck, accompanied by a computational part that was conducted by the author of this thesis. It deals with a binding study of a series of novel bivalent ER ligands involving flexible spacers, with the focus on predicting the conformational entropy loss upon binding, as well as the impact of spacer-protein interactions, again in the presence of explicitly modeled water. The findings suggest that flexible spacers based on oligo(ethylene glycol) introduce a number of effects that may interfere with ligand binding and can possibly be connected to the low binding affinities that have been reported in the literature (Section 4.1.2). Together with the results presented in Chapter 5, which is dealing with multivalent systems involving rigid and semi-rigid spacers, this chapter aims at deriving certain guidelines for the rational design of spacers and scaffolds for multivalent displays.

4.1.1 Estrogen receptor as a drug target

Estrogen receptors (ERs) are DNA-binding proteins in the cytoplasm belonging to the superfamily of nuclear receptors. As ligand-dependent transcription factors, they are involved in a variety of processes that regulate gene expression. ER is activated by the estrogen estradiol (E_2), which, upon binding to the ER monomer, triggers its release from heat shock protein and thus induces receptor dimerization. Thereupon, the ER dimer is able to recruit coactivator proteins, bind DNA and regulate gene transcription. E_2 not only regulates female sexual reproduction, but is also involved in regulating tissue growth, development and other physiological processes. According to its manifold functions, it can be found in different types of tissues such as brain, liver, endometrium, and bones [120, 121]. It is known that ERs are over-expressed in the majority of breast cancer cases, and have also been linked to other kinds of cancer. Ever since, drugs have been developed that bind to ER to prevent the conformational change that is crucial for

¹responsible for concept, synthesis, NMR experiments, and relative binding affinity measurements, supervised by Prof. Dr. Rainer Haag (also concept) at Institut für Chemie und Biochemie, Freie Universität Berlin, Takustraße 3, 14195 Berlin (Germany)

²responsible for synthesis and relative binding affinity measurements, supervised by Prof. Dr. Oliver Seitz at Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin (Germany)

³responsible for cellular assays, supervised by Prof. Dr. Ronald Gust at Institut für Pharmazie - Pharmazeutische Chemische, Universität Innsbruck, Innrain 52a, A-6020 Innsbruck (Austria)

the recruitment of coactivators, which, in turn, is supposed to inhibit the transcription process by preventing ER from establishing its DNA binding abilities [122, 123].

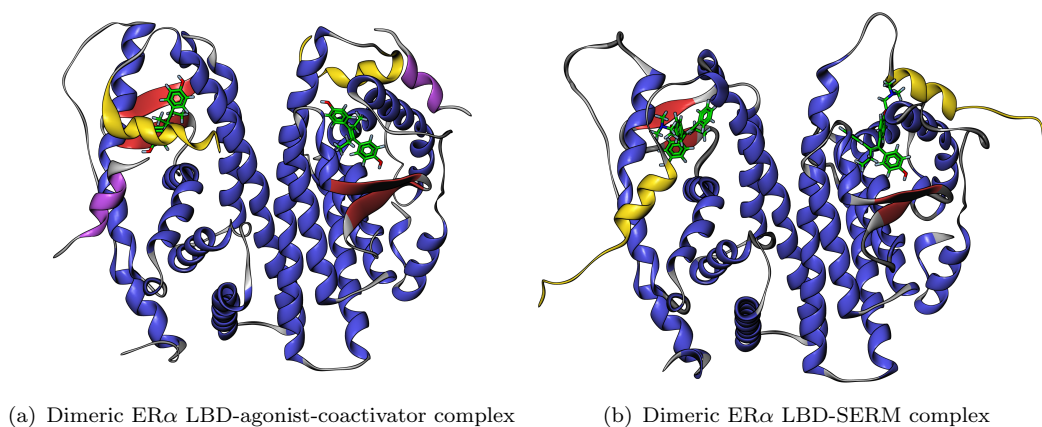


FIGURE 4.1: Secondary structure representations of the crystal structures of the ER α ligand binding domain (LBD) dimer in complex with (a) diethylstilbestrol (DES), a synthetic non-steroidal estrogen that acts as an ER agonist (PDB ID 3ERD), and (b) 4-hydroxytamoxifen (OHT), the active metabolite of the SERM tamoxifen (PDB ID 3ERT). The comparison of the two crystal structures reveals the structural mechanism of coactivator recognition and its antagonism based on the ligand-dependent positioning of Helix 12 (yellow). In (a), a coactivator protein (purple) is able to bind to the open coactivator recognition site, whereas in (b), the same site is obscured by Helix 12 and thus prevents coactivator recruitment [124].

Interestingly, ER binding ligands cannot simply be divided into agonists (such as the native hormone E₂) and antagonists (such as the breast cancer drug fulvestrant), but have to be supplemented by the class of selective estrogen receptor modulators (SERMs). SERMs can act as either agonist or antagonist, depending on the type of tissue. According to Riggs and Hartmann [125], this can be explained by three main factors, namely (i) “differential estrogen-receptor expression in a given target tissue”, (ii) “differential estrogen-receptor conformation on ligand binding”, and (iii) “differential expression and binding to the estrogen receptor of coregulator proteins.”

The first factor (i) is indeed based on the uneven distribution of the receptor subtypes ER α and ER β in different tissue types. Given that, for instance, the SERMs raloxifene and tamoxifen function as pure agonists when binding to ER β , but work as partial agonists when binding to ER α [125], a tissue-specific effect dependent on a differential specificity for ER α and ER β is to be expected. The latter points (ii) and (iii) are largely related to a ligand-dependent conformational change in ER that induces a repositioning of its C-terminal Helix 12 (Figure 4.1): Upon binding of an agonistic ligand, Helix 12 will rearrange and thus seal the steroid binding site so that the coactivator binding site, a hydrophobic groove on the receptor surface, is revealed [122, 123]. Depending on their structure, SERMs typically only allow for an imperfect repositioning of Helix 12, so that the binding of known coactivators is (at least partly) obstructed. It was also reported

that some ER-SERM complexes favor the recruitment of corepressors instead, which, in turn, leads to an amplification of the antagonistic effect. Further degrees of freedom in the modulation of ER activity by SERMs are introduced by the different tissue-specific distributions of coactivators and corepressors [125]. To make things even more complicated, it could be shown that certain SERMs that were tailored for the steroid binding site of ER are able to bind to a part of the coactivator binding site as well, at least in the case of ER β [126] (Figure 4.2).

From the perspective of drug discovery, the (somewhat unpredictable) tissue-specific character of SERMs can be both a blessing – as it opens the potential to create tissue-specific agents with reduced side effects – and a curse. The SERM and breast cancer drug tamoxifen, for instance, acts as an antagonist in breast tissue, and thus inhibits tumor growth. In endometrial tissue, however, it is a partial agonist of ER, and leads to an increased risk of developing uterine cancer. Finally, in bone tamoxifen acts as a full agonist of ER that helps to prevent osteoporosis [127].

4.1.2 Bivalent inhibition of estrogen receptor

In 1997, Brzozowski *et al.* first reported the crystal structures of the ER (subtype ER α) LBD in complex with endogenous E₂ and the SERM raloxifene [128]. This breakthrough provided not only a straightforward method to study the mechanisms of ER coactivator recruitment and its antagonism, but also substantial information about the structural parameters of the ER dimer, such as binding site distance, the size of the dimer interface, and the fact that it has a C₂-symmetric layout. The crystal structures of subtype ER β , complexed to E₂ and genistein (a partial ER β agonist), were resolved shortly after, in 1999 [129] (Figure 4.2). Moreover, evidence was found that the dimerized form of ER exists even in the absence of the ligand [130]. The sum of these findings sparked the idea to exploit the concept of multivalent enhancement by complementing the dimeric receptor layout with a bivalent ligand design, not only in order to develop more potent and stable inhibitors of ER, but also to screen for yet undiscovered ways to modulate ER activity.

A number of motives for pursuing the concept of multivalency have been discussed at length in Chapters 1 and 2. In short, the special properties of a multivalent ligands can induce a cooperative effect that is based on the preorganization of the binding partners partaking in the multivalent interaction, which in turn can lead to the formation of very stable complexes, even (and particularly) for individually weak interacting moieties.

The effects of multivalent preorganization on the interaction of the binding partners can be expressed in terms of different theoretical concepts. Most importantly, the interaction between multivalent ligand and multivalent receptor turns from an intermolecular

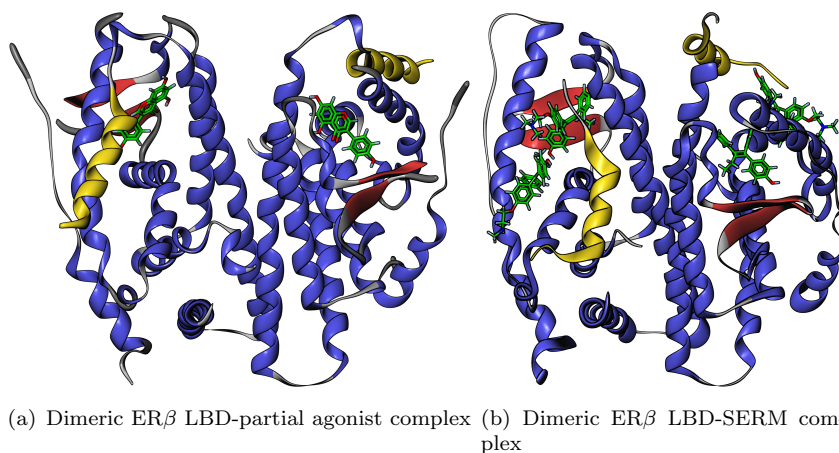


FIGURE 4.2: Secondary structure representations of the crystal structures of the ER β ligand binding domain (LBD) dimer in complex with (a) genistein, a phytoestrogen that acts as a partial ER β agonist (PDB ID 1QKM) [129], and (b) OHT (PDB ID 2FSZ) [126]. The latter structure represents a noteworthy case, as it proves that small hydrophobic molecules can bind to the coactivator binding site as well, which leads to a total of four OHT molecules in the complex. The comparison of the two crystal structures once more illustrates the conformational diversity of ER that is related to the flexible positioning of Helix 12 (yellow).

into an intramolecular one after a binding at the first ligand-receptor interaction site has been established. Thereupon, all subsequent intramolecular binding events can benefit from a significant increase in effective molarity. The extent of the increase in effective molarity, in turn, is dependent on the degree of preorganization and thus directly related to the fitness of the spacer or scaffold that is used for presenting the binding partners in the multivalent array (cp. Chapter 2.3).

The same effect can be formalized in terms of thermodynamics as a reduction of entropy loss upon binding: Due to the fact that subsequent (i.e. intramolecular according to the above definition) binding events in a multivalent binding process do not generate rotational and translational entropy loss in the extent as the binding of additional monovalent ligands would, the overall entropy loss upon binding with regard to all binding sites is favorable for the multivalent ligand. Again, the extent of this entropy “gain” by preorganization is dependent on the structure of the spacer or scaffold, as the auxiliary structure generates additional conformational entropy that is bound to be lost upon binding to the receptor, and thus has to be considered in the overall thermodynamic balance (cp. Chapter 2.2, in particular Figures 2.5 and 2.6).

Finally, once a multivalent binding has been achieved, the stability of the complex can benefit from the occurrence of rebinding events, a subject that has been discussed in Chapter 2.3 (cp., e.g., Figure 2.13).

In summary, the promise of the above advantages made implementing the multivalent concept for novel, bivalent ligands of ER worthwhile. In 1994, Bergmann *et al.* first

investigated the effects of a series of bivalent hexestrol (a non-steroidal ER agonist) linked with oligomethylene and oligo(ethylene glycol) (OEG) spacers of varying lengths [131], even before the crystal structure of ER had been resolved. During the last two decades, several bivalent ER ligands, consisting of two individual ligand moieties linked by flexible tethers, have been designed and synthesized [132–134]. In 2010, Wendlandt *et al.* showed the biological activity of synthetic bivalent estrogen dimers linked by aliphatic spacers to be higher than that of E₂ [135]. Unfortunately, whenever receptor binding affinity was reported, the bivalent compounds mostly performed poorly, significantly below the level of their monovalent counterparts. When comparing different spacer lengths, however, the binding affinities were often peaking at a distance where the occurrence of a bivalent interaction could be expected, i.e. close to the distance of the two ER steroid binding sites [134]. Based on the available theory and models on multivalent binding (cp. Chapter 2), it can be assumed that the regression in binding affinity can be attributed to the additional entropic cost intrinsic to flexible spacers, or a lack of structural fitness of the bivalent compounds that is related to the spacer, or (more likely) a combination of both phenomena.

A better understanding of these factors is a prerequisite for the development of potent bivalent ligands that can actually fulfill the promise of multivalent enhancement. In the following study (Section 4.2), the structural and entropic properties of different flexible and semi-flexible spacer types based on OEG are assessed by means of molecular simulation and conformational entropy calculations. Furthermore, it is investigated how ligand moieties and spacer may interact in aqueous solution. A second study (Section 4.3), involving a series of novel bivalent ligands using flexible EG spacers of different lengths, deals with the conformational entropy loss upon binding that is associated with flexible spacers, as well as the possibility of spacer folding and protein-spacer interactions in the bound state. In the end, certain guidelines for the design of bivalent ER ligands are derived.

4.2 Study F1: Towards a rational spacer design for bivalent inhibition of estrogen receptor

This section is based on the full text of the following publication, reprinted with kind permission of Springer (including figures and tables):

- A. Bujotzek, M. Shan, R. Haag, M. Weber: Towards a rational spacer design for bivalent inhibition of estrogen receptor. *J. Comput. Aided Mol. Des.*, 25(2):253–262, 2011. URL <http://www.springerlink.com/content/m26v953551660127>.

4.2.1 Looking for spacer design paradigms

The importance of the spacer (or scaffold⁴) used for tethering and possibly preorganizing the individual ligands cannot be emphasized enough. The fitness of the spacer w.r.t. the target receptor has a direct impact on the effective molarity of the intramolecular binding events, and thus the degree of cooperativity of the overall interaction. It has an impact on both enthalpy (an unfit spacer can render the interaction unfavorable for enthalpic reasons) and entropy (an overly flexible spacer can render the interaction unfavorable for entropic reasons) of the ligand-receptor interaction.

Arguably, the most important factor for spacer design is the effective distance that has to be bridged, i.e. the distance between the target binding sites. In case of ER α , crystal structures reveal the binding site centers to be approximately 31 Å apart, while the straight-line distance between the binding site openings of the SERM binding conformation is approximately 34.5 Å (measured in the crystal structure with PDB ID 2R6W [136]). The ER dimer is a relatively rigid structure, so that one can assume the binding site distance to remain approximately constant. If the binding sites on a multivalent target are connected by flexible interdomains (one such case is the tSH2 domain of Syk introduced in Chapter 2.2.2, see Figure 2.7), this assumption cannot be made.

Another determining factor for spacer design is spacer flexibility. The spacer has to be flexible enough to permit the attached ligand moieties to enter the target binding sites. The spacer should not create additional energy barriers that impede the positioning of the ligand moieties. In this respect, flexible spacers such as OEG appear to be most adequate, as they provide more room for sterical adjustments compared to a hypothetical rigid spacer structure. More flexibility of the spacer comes at the cost of increased conformational entropy, which in turn threatens to penalize the binding affinity.

A third factor to be considered is related to interactions of the spacer with itself and the attached ligands, which again is determined by spacer flexibility (as such interactions can only occur if spacer flexibility allows for it). Ligand-ligand interactions may occur if the spacer folds over so that the attached ligands can come into interaction range. Ligand-spacer interactions may occur if, for instance, both ligand and spacer possess hydrophobic motifs, and, finally, the spacer may form superstructures such as helices by self-interaction. All of these effects are likely to penalize binding affinity.

There is an abundance of other questions related to the design of molecular spacers, such as uptake in body and/or cell, processing in biological systems, and, of course, chemical synthesis. These, however, are beyond the focus of this study.

⁴The following statements apply to both spacers and (typically more complex) scaffolds, but the term scaffold is abandoned because the focus of this chapter is on rather simplistic spacers without higher levels of organization (cp. Chapter 1.3).

4.2.2 OEG spacers and novel semi-flexible spacer designs

Traditionally, the most commonly used spacer for bivalent targeting of ER has been OEG or poly(ethylene glycol) (PEG), which is water-soluble, has a low toxicity and is commercially available in different chain lengths. However, as was briefly discussed in the introduction, the performance of OEG-linked bivalent ER ligands still leaves room for improvement. It was shown that OEG has a flexible structure that is prone to form coiled and helix-like conformations in aqueous solution [137]. This property may be disadvantageous for the applicability of OEG/PEG as a molecular spacer. LaFrata *et al.* proposed the use of less polar compounds (namely poly(propylene) and poly(butylene glycol) ethers) to address the challenge of coil and helix formation [134]. This, too, did not immediately lead to improved binding affinities.

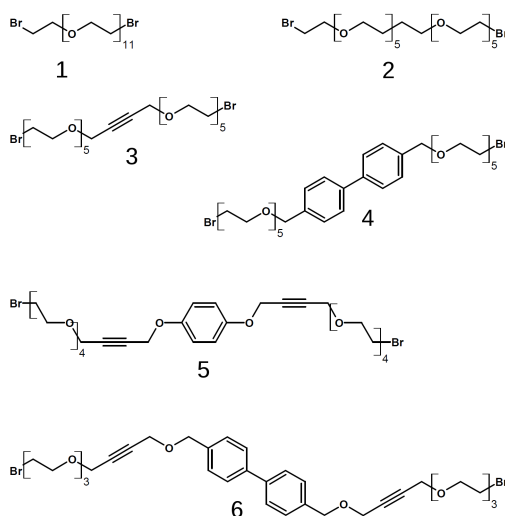


FIGURE 4.3: Structures of compounds **1** (bromine-capped EG11) to **6**, the spacer designs involved in the computational study.

In this computational study, the basic EG11 spacer (spacer **1**) is compared to five OEG-based spacer designs (Figure 4.3), that, in the fully extended conformation, are bridging approximately the same distance as EG11 (Table 4.1). The new designs introduce different structural motifs to the OEG backbone, meant to break helical folding patterns and provide rigidity for the extended, distance-spanning conformations one is typically looking for. For rigid elements, 2-butyne (spacers **3**, **5** and **6**), phenyl (spacer **5**) and biphenyl (spacers **4** and **6**) are used. Regarding spacer **2**, no additional rigidity is added to the structure, but the central EG subunit is replaced by a single carbon-carbon bond, again meant to suppress helix formation. A conformational analysis of all compounds is performed to assess the effect of the proposed structural elements. Furthermore, a method for conformational entropy estimation [108] is applied in order to look into the impact of rigid elements on the entropic properties of the structures. Retaining OEG as

structural element in all of the proposed spacer designs aims at maintaining a sufficient solubility in water.

Compound	Structure	Length [Å]
1	EG11	43.72
2	EG5-C ₂ -EG5	42.59
3	EG5-(2-butyne)-EG5	42.60
4	EG5-C-biphenyl-C-EG5	48.59
5	EG4-(2-butyne)-O-phenyl-O-(2-butyne)-EG4	44.84
6	EG3-(2-butyne)-O-biphenyl-O-(2-butyne)-EG3	45.31

TABLE 4.1: Maximum spanned distance (bromine-bromine) of spacers **1–6** in the fully extended conformation.

Attaching ligands to a flexible or semi-flexible spacer is likely to have an impact on its conformational profile. Unwanted interactions between ligand and ligand as well as ligand and spacer may be related to the low binding affinities that have been reported for bivalent ER ligands. To look into this, each of the proposed spacer designs was modeled twice, once capped only with bromine, i.e. without actual ligand moieties attached (cp. Figure 4.3), and once coupled via amide linkage groups to two copies of DES (Figure 4.4, with bivalent DES-coupled EG11 serving as an example). Note that in the context of this study, DES serves merely as a prominent (but arbitrarily chosen) representative for any ER binding ligand.

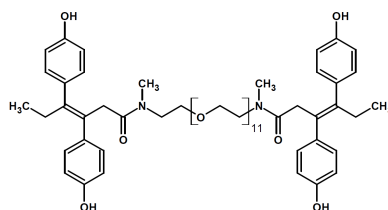


FIGURE 4.4: Two copies of DES coupled via amide to EG11, forming a bivalent ligand for ER. For comparison, each spacer design shown in Figure 4.3 was modeled with and without ligands attached (i.e. amide-coupled to DES or merely bromine-capped).

4.2.3 Results

End-to-end distances

As Table 4.2 shows, mean end-to-end distances over 100 ns simulation time in water are less than half of what had been measured for the extended conformations. This indicates that folded spacer conformations are highly favored over linear, extended conformations. The minimum distance observed throughout the trajectories is virtually equal for all

Compound	Min. [\AA]	Max. [\AA]	μ [\AA]	σ [\AA]
1	3.45	34.06	14.99	5.84
2	3.51	32.37	12.25	5.19
3	3.48	36.44	13.14	5.78
4	3.48	37.54	14.50	6.30
5	3.40	38.89	13.05	5.88
6	3.45	39.01	17.69	7.36

TABLE 4.2: Bromine-bromine distance data measured over 100 ns for the bromine-capped spacers (mean μ , standard deviation σ).

spacer designs, which implies that all structures are flexible enough to fold over, i.e. the minimum end-to-end distance is limited only by the terminal bromines' van der Waals radii. The highest mean end-to-end distance is observed for spacer **6**, with 17.69 \AA , which also has the highest distance standard deviation, followed by EG11 (spacer **1**), with a mean of 14.99 \AA . Maximum end-to-end distances reach more than 30 \AA for all spacer designs, theoretically sufficient for bridging the two binding sites of ER. However, the occurrence of extended conformations is very low (cp. Figure 4.5). The exception is spacer **6**, which also has the widest distribution of distances in general.

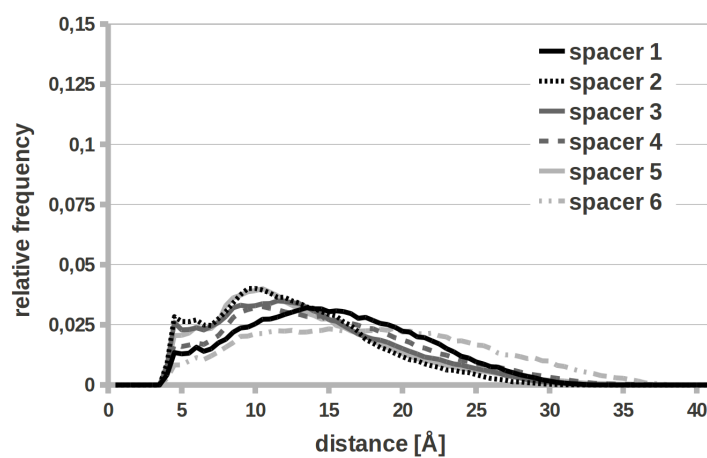


FIGURE 4.5: Histogram of bromine-bromine distances measured over 100 ns for the bromine-capped spacers.

End-to-end distances decrease further as soon as DES ligands are attached to the spacer (cp. Table 4.3). Mean nitrogen-nitrogen distances drop below 10 \AA for all compounds, with spacer **6** again performing best with a mean distance of 9.84 \AA . The decrease in mean end-to-end distances is accompanied by a general decrease of the standard deviation w.r.t. the distribution of distances. Minimum and maximum distances remain in a similar order.

Compound	Min. [\AA]	Max. [\AA]	μ [\AA]	σ [\AA]
1	3.61	31.19	9.53	3.75
2	4.06	29.81	9.72	2.27
3	3.11	31.52	8.11	2.86
4	3.71	33.03	8.55	3.34
5	3.48	40.84	9.33	3.82
6	3.36	35.07	9.84	3.44

TABLE 4.3: Nitrogen-nitrogen distance data measured over 100 ns for the DES-coupled spacers (mean μ , standard deviation σ).

Figure 4.6 shows that the end-to-end distance distributions for the ligand-coupled spacers become significantly narrower. The emergence of distinct peaks suggest that each compound adopts a relatively limited number of stable conformations. This seems to apply in particular to spacers **1–4**.

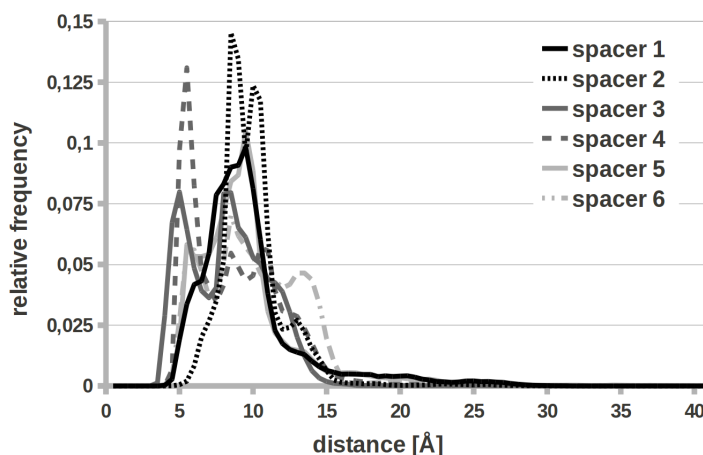


FIGURE 4.6: Histogram of nitrogen-nitrogen distances measured over 100 ns for the DES-coupled spacers.

The effects of ligand attachment on end-to-end distances mean and standard deviation for all compounds are summed up in Figure 4.7. The largest drop of both mean and standard deviation on ligand attachment occurs for spacer **6**, followed by spacer **4**. In order to investigate if the decrease in end-to-end distances is exclusively related to bivalent ligands coupled with DES, the structure of a bivalent steroidal ligand published by LaFrate *et al.* [134] was modeled and simulated analogously. This structure, consisting of an EG5 spacer that tethers two E₂ residues modified at the 17 α position, also suffered from a drop in end-to-end distances (measured at the C17 positions) to a mean value distinctly below 10 \AA (data not shown). This suggests that the phenomenon is not limited to bivalent ligands involving DES. It may rather be connected to more general

structural motifs of the coupled ligand moieties, such as the presence of conjugated systems.

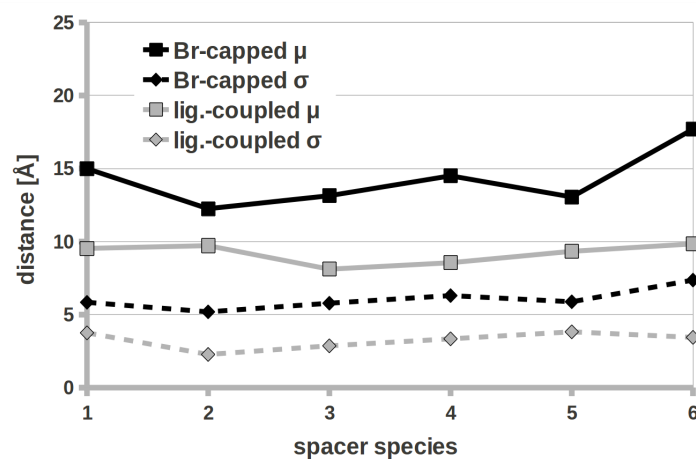


FIGURE 4.7: Impact of ligand attachment on spacer end-to-end distance mean (μ , full line) and standard deviation (σ , dotted line) for bromine-capped compounds (black) and ligand-coupled compounds (light gray).

Conformational entropy estimation

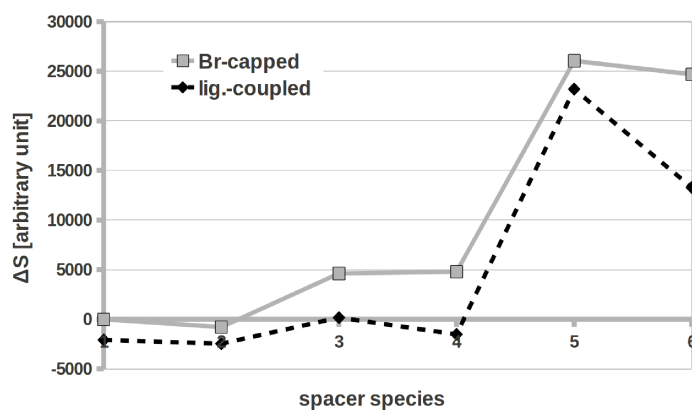


FIGURE 4.8: Conformational entropy comparison of bromine-capped spacers (full line) and DES-coupled spacers (dotted line). Bromine-capped EG11 (spacer 1) was used as the reference compound.

Conformational entropy follows an upward trend for spacers 1–6, where the lowest conformational entropy is measured for spacers 1 and 2 (EG11, and its closest derivative), followed by spacers 3 and 4 (both containing either one 2-butyne or one biphenyl element). Spacers 5 and 6, which both contain combinations of multiple 2-butyne and phenyl elements, are ranking with the highest conformational entropy (Figure 4.8).

All DES-coupled compounds exhibit a lower conformational entropy than their bromine-capped counterparts. This decrease is expected, as the results shown in the previous section suggest that ligand-attachment constraints the conformations of the spacer. The extent of the entropy drop is most significant for spacers **4** and **6**, which both contain biphenyl residues. The formation of stable hydrophobic interactions between biphenyl and DES ligands may explain this effect. By contrast, the smallest drop in conformational entropy is observed for spacers **1** and **2**. Here, either spacer structure may prohibit ligand-spacer and ligand-ligand interactions (which, however, would contradict end-to-end distance data), or the spacer intrinsically may have a folding pattern with low conformational entropy that is not dramatically altered as soon as ligands are attached. The latter seems to be the case for spacers **1** and **2**, which in aqueous solution are prone to form low entropy helix-like structures.

The distinct rise of conformational entropy from spacers **1–6** appears to be counterintuitive, as an increasing number of rigid, seemingly low entropy elements had been incorporated into the structures. In order to explain the above entropy ranking (cp. Figure 4.8), it was looked into which types of structural elements (Figure 4.9) contribute most to the spacers' conformational entropy balance.

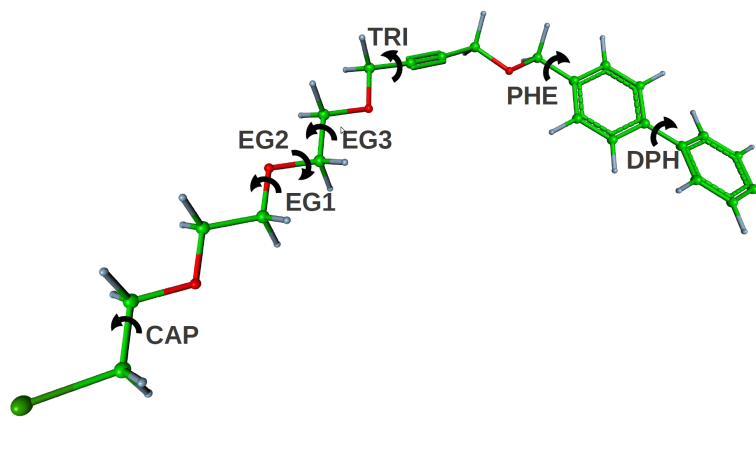


FIGURE 4.9: Torsion angle types with different degrees of conformational entropy contribution: Torsions at spacer caps or ends (CAP), torsions of EG units (EG1, EG2, EG3), torsions neighboring triple bonds (TRI), torsions neighboring phenyl rings (PHE), and torsions connecting two phenyl rings (DPH).

The highest conformational entropy contribution is found for the terminal torsion angle of the spacer (CAP), where either bromine or the ligand is attached (Figure 4.10). One EG subunit consisting of three bonds has a relatively low entropy contribution, and shows a regular pattern of two low entropy torsions (EG1 and EG2) and one high entropy torsion (EG3). The second-highest entropy contribution can be found for torsions neighboring triple bonds in 2-butyne (TRI), as they allow for unhindered rotation, resulting in an uniform distribution over the whole torsion angle range. One can also

find a high entropy contribution for torsions neighboring phenyl rings (PHE) and torsions connecting two phenyl rings (DPH). Therefore, although phenyl rings and triple bonds intrinsically are rigid, low entropy structures, they introduce high conformational entropy to the system by enabling free rotation of the attached groups.

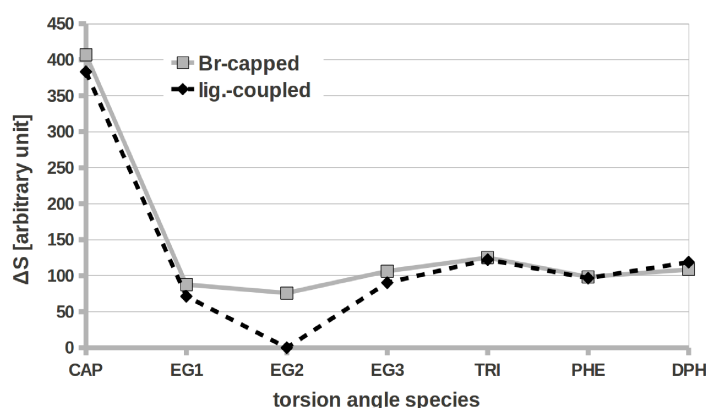


FIGURE 4.10: Conformational entropy contribution of different types of torsion angles in bromine-capped spacers (full line) and DES-couple spacers (dotted line), measured on the example of spacer **6**.

The above results suggest a relatively mild effect of ligand-attachment on individual torsion angle types, as in all cases except for EG2 they remain more or less invariant (cp. Figure 4.10). The significant drop in entropy for EG2 indicates that in spacer **6** (which was used for the torsion angle comparison (cp. Figure 4.10) the largest entropic penalty caused by ligand-spacer and ligand-ligand interactions has to be borne by the short EG3 chains, which otherwise would be able to move in an unhindered fashion. These findings help to explain the conformational entropy ranking shown in Figure 4.8: As soon as an EG subunit is replaced by a triple bond (e.g. spacer **3**), two high entropy TRI torsions are introduced to the structure. This results in an overall higher conformational entropy, especially when the TRI motif interrupts a longer EG chain that otherwise would be prone to fold into a low entropy structure such as a helical loop. The same is true for structures containing phenyl (two PHE torsions) and biphenyl (two PHE torsions and one DPH torsion).

Conformational analysis

As the end-to-end distance results presented above have indicated, one can observe a folding of spacer structures in water when looking at the actual MD trajectories. Predominant conformations for bromine-capped EG11 are either loop or U-shaped. Helical conformations are also present in the mix, but the formation of stable or even rigid helices that persist for more than a few picoseconds cannot be confirmed. The overall

flexibility of EG seems to be high. For short EG chains, folding into complex coils or loops is not possible, which adds to the high flexibility.

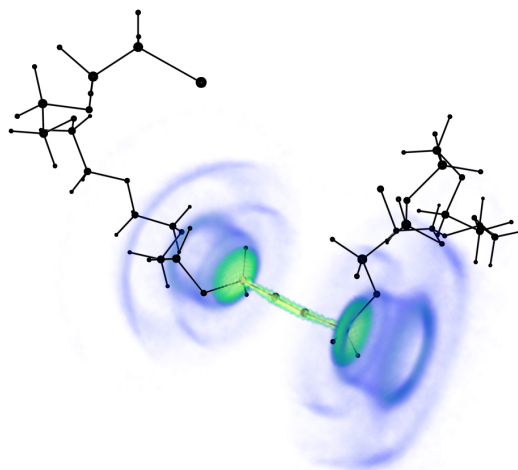


FIGURE 4.11: Conformation density of bromine-capped spacer **3**. Coloring indicates the presence of atoms with at least 10 % probability. While the central 2-butyne fragment remains linear, the attached EG chains rotate heavily so that no distinct positions are visible.

Rigid elements such as 2-butyne (Figure 4.11) and biphenyl (data now shown) retain their linear conformation, but introduce a great amount of rotational entropy to the remaining part of the molecule. It might be argued that this free rotation can have a beneficial effect as it prevents trapping in a certain conformation that may be adverse to binding. However, it leads to an increase of the entropy penalty that has to be suffered upon binding to the receptor.

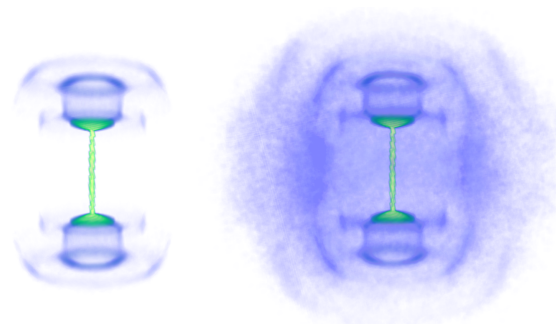


FIGURE 4.12: Conformation density plot of spacer **3**. Coloring indicates the presence of atoms with at least 15 % probability. Without ligands attached, the EG chains can move freely (left). As soon as DES is coupled to the spacer, the structure becomes more compact and chain flexibility is restricted. The DES ligands align in parallel to the 2-butyne part of the spacer, forming a barrel-shaped structure (right).

Depending on the spacer design, the attachment of DES moieties can have more or less pronounced effect. For spacer **3**, a relatively large drop in conformational entropy

upon coupling with DES had been calculated (cp. Figure 4.8). This is supported by the conformation density plot shown in Figure 4.12. While for the bromine-capped spacer, EG chains can rotate freely, attaching DES moieties leads to a barrel-shaped structure, where DES aligns parallelly to the 2-butyne part of the spacer. The molecule thus becomes more compact and less flexible. By contrast, for spacer **5**, only a relatively minor drop in conformational entropy upon coupling with DES had been found (cp. Figure 4.8). The according conformation density plot (Figure 4.13) supports this finding, as with DES moieties attached, the densities appear to be only slightly more defined than for the bromine-capped spacer. Largely unhindered rotation of the EG chains and the presence of only one aromatic ring in the structure of spacer **5** may prevent the formation of more pronounced hydrophobic stacking interactions as in the case of biphenyl-containing spacers **4** and **6**.

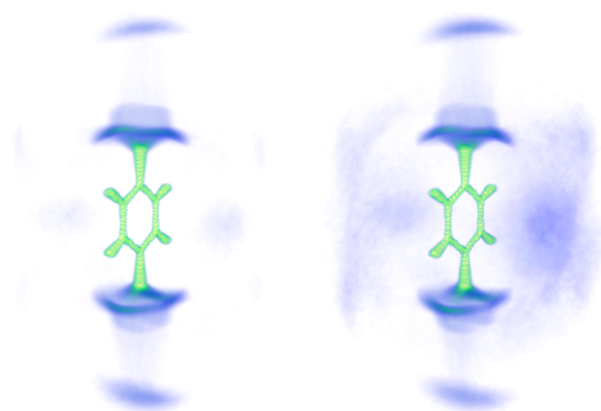


FIGURE 4.13: Conformation density plot of spacer **5**. Coloring indicates the presence of atoms with at least 15 % probability. For bromine-capped (left) and ligand-coupled spacers (right) a similar density plot is obtained, i.e. the impact of ligand-attachment is relatively small. The slight shadows appearing parallel to the phenyl ring indicate the presence of the ligands or parts of the spacer interacting with the phenyl ring. The fact that the density cloud in this area is not very defined suggests that these interactions are comparatively weak, i.e. conformational flexibility remains. This is in agreement with the conformational entropy estimation of spacer **5**, where only a minor entropy drop after ligand-attachment was found (cp. Figure 4.8).

For the spacer designs containing biphenyl, a variety of stacked conformations that maximize hydrophobic interactions involving the DES moieties was found (Figure 4.14). This provides a justification for the low end-to-end distances that were observed after coupling the spacers with DES (cp. Figure 4.5). The formation of stacked conformations not only neutralizes the distance spanned by the linear elements, but may also have adverse effects on water solubility. Although not taken into account within the course of this purely computational study, it is possible that multiple structures can be involved in such interactions to form larger aggregates.

4.2.4 Discussion

The design of potent bivalent ligands for ER remains a challenging task. The ligand binding sites are buried deeply within the receptor, and between the binding sites a distance of more than 30 Å has to be bridged. The molecular spacer used to connect the ligands has to provide the right interplay between flexibility and rigidity.

The results of this computational study suggest that an optimal spacer should be as rigid as possible, while still providing the flexibility that is required for allowing access to the binding pockets. Furthermore, in order to provide reliable design guidelines for chemists, the spacer should have a defined conformation and bridging distance when exposed to solvent, and the spacer parameters should be invariant w.r.t. the attachment of ligand moieties.

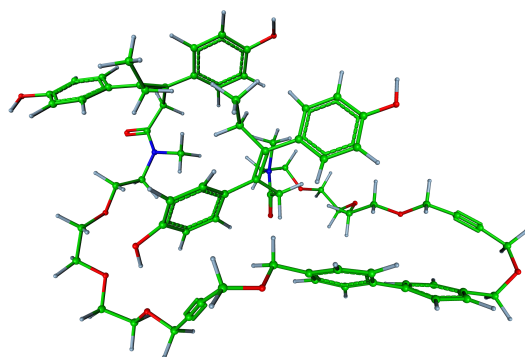


FIGURE 4.14: In water, hydrophobic interactions promote stacked, sandwich-like conformations, shown here for DES-coupled spacer 6.

The flexible and semi-flexible spacer designs under investigation in this article reveal certain weaknesses. Due to their high structural properties, the effective bridging distance in water is diminished even after introducing rigid segments. As OEG/PEG shows very variable folding patterns, a mere prolonging of the chain is only of limited use. The prediction of an optimal chain length is not possible without elaborate case-specific analysis. Furthermore, the attachment of ligand moieties leads to distinct changes of the spacers' parameters, which complicates the rational design of bivalent ligands. Conformation density plots show that the use of phenyl and 2-butyne as linear, rigid elements leads to more defined spacer structures and linear conformations. The positive effects are neutralized, however, by the flexibility of the attached OEG chains. Even short chains of three EG units allow for enough flexibility to permit interactions between the ligand and hydrophobic parts of the spacer and thus promote a hydrophobic collapse of the structure.

Consequently, there is a need for novel compounds, hydrophilic and yet rigid, that can be assembled in a modular fashion in order to provide spacer structures with defined

conformation and bridging distance. This notion is summarized in Figure 4.15. The results of this study suggest that the probability to find flexible spacers based on OEG in an extended conformation is low (cp. Figure 4.15 a)), whereas folded spacer conformations (cp. Figure 4.15 b)), occur with high probability. The use of rigid, modular spacer or scaffolds with defined ligand presentation distance and short flexible linkers (cp. Figure 4.15 c)) may offer the best compromise between flexibility and rigidity.

In order to prohibit ligand-spacer interactions that might hinder the free presentation of the ligand moieties, a prerequisite for achieving stable binding to the target, the flexible linkers used to connect spacer and ligand moieties ought to be rather short, just enough to allow for access into the binding pocket without causing steric hindrances with the receptor surface. This parameter, in turn, is dependent on both the depth of the receptor binding pocket as well as the structure of the ligand moiety. The exemplary ligand DES used in this study is an ER agonist⁵ and thus buried deeply in the ER structure. Therefore, it would require a linker of several Å length to enable the correct positioning when attached to a larger spacer or scaffold structure. SERMs such as raloxifene, in contrast, typically protrude from the binding pocket so that only a comparatively short linker would be required.

One of the main promises of using bivalent (and multivalent ligands in general) is the entropic advantage gained over their monovalent counterparts, based on the opportunity to save translational and rotational entropy for subsequent, intramolecular binding events (cp. Section 4.1.2). This advantage is opposed by the intrinsic conformational entropy contribution of the spacer, or, more precisely, the conformational entropy loss of the spacer upon binding.

Conformational entropy differences between different spacer types and elements were determined. For OEG chains, a regular pattern of triplets consisting of two low entropy torsions and one high entropy torsion was found. For phenyl and biphenyl, hardly any conformational entropy contribution of torsions involved in the conjugated systems was measurable. However, bonds neighboring phenyl groups allow for unhindered rotation and were found to introduce a notable amount of conformational entropy to the system. The same applies to bonds neighboring triple bonds. Although 2-butyne and phenyl/biphenyl parts help to construct rigid, linear structures that maintain defined distances for ligand presentation, their high contribution in terms of conformational entropy appears to be a drawback. In general, values for spacers containing 2-butyne and phenyl elements generated higher conformational entropy values.

⁵Considering the structural basis of coactivator recruitment by the ER dimer (cp. Section 4.1.1), it can be expected that the attachment of any type of spacer or scaffold via a linkage group will turn an originally agonistic ER ligand such as DES into an (at least partially) antagonistic one, as the linker/spacer/scaffold protruding from the steroid binding site is likely to prohibit an adequate repositioning of Helix 12.

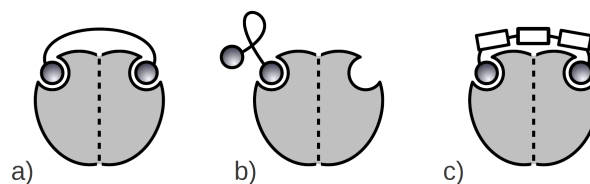


FIGURE 4.15: Different scenarios of bivalent binding that can be modulated by the structure of spacer or scaffold: a) Bivalent binding is achieved with a flexible spacer that adapts a linear conformation. b) Bivalent binding is rendered impossible due to folding of the spacer and stacking of hydrophobic ligand moieties. c) Bivalent binding is achieved with modular rigid spacer and short flexible linkers.

The entropic advantage that is connected to the cooperative effect in multivalent systems can only be exploited if the spacer does not interfere with the enthalpy of the interaction, i.e. the ligand moieties have to be presented in a way that allows for unhindered access to the target binding sites. Therefore, maintaining a defined ligand-ligand distance (corresponding to the approximate binding site displacement) and avoiding ligand-ligand and ligand-spacer interactions that interfere with ligand presentation (such as hydrophobic stacking) is more fundamental than entropic optimization of the spacer. The author's opinion is that an adequate preorganization of the ligand moieties should be the prime interest when designing multivalent ligands, only then followed by entropic considerations. An issue not addressed in this study, and yet to be discussed in the next sections, is the interplay of protein and spacer structure. For instance, one might imagine a scenario where contact between spacer and protein, following the initial binding event, induces spacer-protein interactions that promote unfolding of the flexible spacer, thus leading to linear spacer conformations that encourage bivalent binding. Furthermore, these interactions might even improve the enthalpy of the interaction so that $\Delta H_{bi} \approx 2 \times \Delta H_{mono} + \Delta H_{spacer} < 2 \times \Delta H_{mono}$. Regarding the use of OEG/PEG spacers this scenario can be challenged, as PEG, when attached to surfaces, is known to rather have protein-repellent properties [138]. These questions have to be the subject of further investigations.

The results presented in this study may be helpful in order to avoid certain traps connected with the design of bivalent ligands for ER, or similar bivalent targets. The findings may serve as starting points for research aimed at providing chemists with new and improved molecular spacer and scaffold structures.

4.3 Study F2: Conformational analysis of bivalent estrogen receptor-ligands – From intramolecular to intermolecular binding

This section is based on (and contains content from) the following publication:

- M. Shan, A. Bujotzek, F. Abendroth, O. Seitz, M. Weber, R. Haag: Conformational analysis of bivalent estrogen receptor-ligands: From intramolecular to intermolecular binding. *ChemBioChem*, 12(17):2587–2598, 2011. URL <http://onlinelibrary.wiley.com/doi/10.1002/cbic.201100529/abstract>.

The second study, presented in the following, aims at investigating the binding behavior of bivalent ER ligands tethered by flexible spacers in dependence on the spacer length. For this purpose, a series of bivalent ligands based on the SERM raloxifene (RAL), tethered by flexible OEG spacers of varying length, was designed and synthesized, and the relative binding affinities (RBA) for ER α were determined *in vitro*, accompanied by cellular assays for determining biological activity w.r.t to ER α and ER β . In order to examine why the bivalent RAL ligands (and bivalent ER ligands tethered by flexible spacers in general) are suffering from low binding potencies when compared their monovalent counterparts, a conformational study in aqueous solution based on the analysis of NMR and UV absorption spectra, as well as a computational study on thermodynamic factors and binding modes was incorporated into the project. The author of this thesis is responsible only for the computational part of this study (Section 4.3.3), but the experimental results as contributed by the collaborating scientists (see remarks in Section 4.1) are summarized in order to provide the necessary context.

4.3.1 Bivalent ligand design

In the literature, the most successful bivalent estrogen ligand tethered by flexible spacers study was reported by LaFrata *et al.* [134]. They prepared their bivalent estrogen ligands based on the crystal structure of the ER α LBD in complex with E₂ (PDB ID 1ERE) and obtained a peak in ER α binding affinity when the ligand moieties were tethered by a spacer of 35 Å length (linear extended conformation), which correlates nicely with the approximate distance of the steroid binding sites in the crystal structure (≈ 31 Å).

The design of the bivalent RAL ligands prepared in the course of this study was based on the crystal structure of the ER α LBD in complex with LY156681 (PDB ID 2R6W) [136], a derivative of RAL (Figure 4.16). LY156681, being a SERM, binds to the same steroid binding site as the ER agonist E₂, but leads to a displacement of Helix 12

when compared to the closed, agonist-binding conformation that can be found in crystal structure 1ERE (cp. Section 4.1.1). Due to fact that the basic side chain of LY156681 protrudes from the steroid binding site of ER (the structural property that prohibits the reorientation of Helix 12), the conformation of ER α in crystal structure 2R6W offers an excellent starting point for envisioning the binding mode of a bivalent ligand: Any kind of spacer used to tether two ER ligands would have to protrude from the binding pocket in a similar fashion, and thus lead to a similar arrangement of Helix 12. At the same time, the side chain of SERMs like RAL appears to provide a promising linkage site for attaching the spacer, as it leaves the main portion of the ligand moiety untouched, and consequently should retain a high binding affinity for ER – a hypothesis that is put to the test in the course of this study. The series of monovalent RAL derivatives investigated by Dai *et al.*, including LY156681, were shown to have high affinities for ER α , with binding constants of less than 10 nM [136].

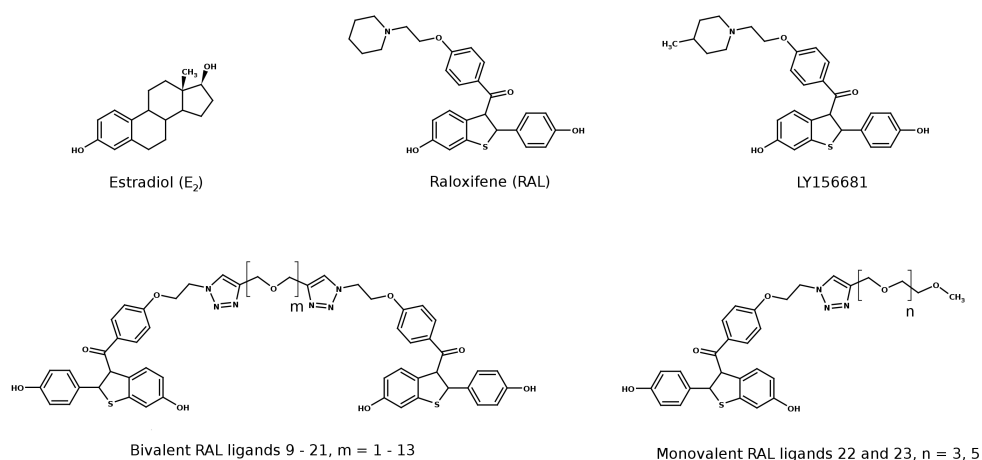


FIGURE 4.16: Structures of the ER agonist estradiol (E₂), the SERM raloxifene (RAL), its derivative LY156681, and the novel bivalent (**9–21**) and monovalent RAL ligands (**22, 23**) as designed and synthesized for this study.

Using crystal structure 2R6W and the modeling and visualization software Amira [139], the distance between the two piperidine nitrogen atoms at the side chain of the two LY156681 molecules bound to the ER α LBD dimer was determined to be approximately 34.5 Å, which again matches the optimal spacer length for bivalent ER ligands observed by LaFrata *et al.* [134]. Considering the results of the *in silico* Study F1 presented in Section 4.2, which suggested that the length of the extended, linear conformation of a flexible spacer is not a proper measure of its effective bridging distance, it seemed worthwhile to examine the influence of the spacer length of bivalent RAL ligands on the associated binding affinities. Therefore, a systematic spacer length study, covering spacer lengths from 4.71 to 47.7 Å (corresponding to EG1–EG13 in the extended linear

conformation), was carried out, acknowledging that the shorter spacer variants are factually unable to realize bivalent binding to both steroid binding sites of the ER dimer. Despite the somewhat worrisome results of Study F1 w.r.t. to the properties of OEG spacers (in particular concerning folding and the possibility of hydrophobic collapse in water), OEG was chosen as spacer structure for tethering the RAL moieties because of its favorable qualities for the application in biological and pharmaceutical applications, e.g. for being non-toxic, biocompatible and water-soluble [56]. Furthermore, OEGs have been reported to be resistant to the non-specific adsorption of proteins [140], which would exclude unwanted interactions between protein and spacer, and thus should promote an bivalent interaction only between the ligand moieties and the steroid binding sites of ER. In order to compensate for the possibility of OEG folding into helical or coiled conformations, and thus a reduction of the effective spacer bridging distance, OEG chains of significantly more than 34.5 Å length (extended) were incorporated into the study. With increasing spacer length of the flexible spacer, an increase in the conformational entropy loss upon binding has to be expected, along with an enthalpic penalty that may arise from hydrophobic interactions between the tethered ligands in aqueous solution (cp. Section 4.2.3). These questions were therefore made subject of the computational part of this study (Section 4.3.3).

The structure of the novel bivalent ligands based on RAL, 9–21⁶, is depicted in Figure 4.16, along with two monovalent RAL ligands with monomethylated OEG side chains that were prepared as monovalent control (compounds **22** and **23**). Instead of the 4-methylpiperidin that defines the side chain of LY156681, the novel compounds feature a 1,2,3-triazole at the same position.

4.3.2 Biological evaluation

The affinities of the 15 mono- and bivalent RAL ligands (**9–23**) were evaluated with the HitHunter enzyme fragment complementation (EFC) estrogen chemiluminescence assay (DiscoverRx) [141]. All binding affinities are expressed as RBA values (Table 4.4), that is, relative to the IC₅₀ values of the reference, E₂ (3.11 nM, RBA = 100%). Figure 4.17 shows the relationship between the RBA values and the maximum spacer lengths of the bivalent ligands. In general, all of the mono- and bivalent ligands had lower affinity for ER α than the reference E₂, and with increasing spacer length, the binding affinities of the bivalent ligands decrease further.

Due to the fact that RBA values do not directly reflect biological effects, such as estrogenic or antiestrogenic activity, compounds **9**, **12**, **15**, **18**, **21**, **23**, and raloxifene

⁶Compounds **1–8** are related to the synthesis of compounds **9–21** only, and do not appear in the binding study. The numbering was maintained in order to keep consistency with the original Ref. [16].

Compound	Structure	Length ^[a] [Å]	RBA ^[b] [%]	$C_{eff}^{[c]}$ [nM]
9	RAL2EG1	4.71	$67.7 \pm 27.5^{[d]}$	7.59×10^6
10	RAL2EG2	8.35	18.6 ± 8.10	1.36×10^6
11	RAL2EG3	11.9	25.6 ± 11.1	4.73×10^5
12	RAL2EG4	15.5	14.2 ± 5.41	2.13×10^5
13	RAL2EG5	19.1	24.3 ± 9.28	1.15×10^5
14	RAL2EG6	22.7	3.73 ± 2.90	6.81×10^4
15	RAL2EG7	26.2	$10.2 \pm 2.15^{[d]}$	4.40×10^4
16	RAL2EG8	29.8	4.51 ± 3.28	2.99×10^4
17	RAL2EG9	33.4	5.29 ± 1.47	2.13×10^4
18	RAL2EG10	37.0	$4.68 \pm 1.02^{[d]}$	1.57×10^4
19	RAL2EG11	40.6	3.51 ± 0.66	1.19×10^4
20	RAL2EG12	44.2	4.41 ± 0.70	9.20×10^3
21	RAL2EG13	47.7	$4.64 \pm 0.63^{[d]}$	7.29×10^3
22	RALEG4Me	14.3	$10.0 \pm 6.58^{[d]}$	-
23	RALEG6Me	21.5	$3.42 \pm 2.65^{[d]}$	-
raloxifene	-	-	$30.0^{[d]}$	-

TABLE 4.4: Structure, maximum spacer length, relative binding affinity (RBA) and effective concentration (C_{eff}) of bivalent (9-21) and monovalent ligands (22, 23). [a] Distances between the carbons at 4-position of the 1,2,3-triazole in the extended linear conformation were measured with Amira and considered as the maximum spacer length. [b] RBA values were determined by competitive bindings assays and expressed as $[IC_{50}(E_2)/IC_{50}(\text{compound})] \times 100$ % (RBA for $E_2 = 100$ %, $IC_{50}(E_2) = 3.11$ nM for $ER\alpha$). The K_d of E_2 is 0.3 nM for $ER\alpha$. [c] C_{eff} was calculated as the second RAL moiety in a hemisphere of a radius equal to the spacer length of the bivalent ligand [56]. [d] Ref. [17]. Table taken from [16], reprinted with kind permission of John Wiley and Sons.

were chosen to evaluate hormonal activity in a transactivation assay using U2OS cells transfected with plasmids for either $ER\alpha$ or $ER\beta$, as well as a luciferase reporter gene. As expected, raloxifene only showed antagonistic effects w.r.t. to $ER\alpha$ and $ER\beta$, and the antiestrogenic activity was about 20 times higher for $ER\alpha$ than for $ER\beta$. Similarly, monovalent ligand **23** did not show estrogenic effects, but weak antagonistic effects. The antagonistic effects were much greater in the case of the bivalent ligands (which turned out to be pure antiestrogens), with a 5–20 times higher selectivity for $ER\alpha$. Furthermore, the antagonistic effect on $ER\alpha$ was dependent on the spacer length and decreased in the order: **9** ($IC_{50} = 11$ nM) > **12** (19 nM) > **15** (50 nM) > **18** (92 nM). Compound **21** showed nearly the same antagonistic effect as **9**.⁷

Given that the closest distance between the two LY156681 molecules in crystal structure 2R6W is 34.5 Å, bivalent ligands tethered by shorter OEG spacers are theoretically not capable of forming a simultaneous binding between the two $ER\alpha$ steroid binding sites.

⁷More details on the cellular assays can be found in Ref. [16], cp. in particular in Tables 2 and 3.

The question from which spacer length onwards bivalent binding is feasible will be addressed in the computational section of this study. Certainly, bivalent binding to the steroid binding sites is not realizable with a spacer length of less than 20 Å, which is why the high RBA values of compounds **9–13** in comparison to the monovalent control, compounds **22** and **23**, are rather striking.

The higher binding affinity of the bivalent ligands with short spacers therefore needs to be attributed to the second, tethered RAL moiety. In principle, two scenarios can be envisioned: (i) The second “dangling” RAL moiety tethered to the one that is already bound leads to an increase in effective concentration in the vicinity of the steroid binding site, which in turn would promote the enhanced occurrence of rebinding events [56] (cp. Chapter 2.2.2). In fact, the decrease in effective concentration with an increase in spacer length follows the same trend as the decrease in binding affinities (Table 4.4, column C_{eff}). On the other hand, recent theoretical results suggest that the occurrence of rebinding events in this particular situation – low ligand valency, high individual ligand binding affinity, gated binding sites – is rendered relatively unlikely [30]. Therefore, a second scenario appears to be more probable: (ii) The second RAL moiety undergoes additional interactions with the surface of the ER α LBD, most likely on the same monomer [142, 143]. There have been several reports about a secondary estrogen binding site that lies very close to or within the coactivator binding groove of the ER LBD [144, 145], and in fact a crystal structure of ER β in complex with the SERM 4-hydroxytamoxifen binding to this particular site exists [126] (cp. Sections 4.1.1 and 4.1.2). Given the large degree of homology between ER α and ER β , a similar non-specific (or rather “semi-specific”) binding mode on the surface of ER α can be envisioned as well, and would explain the increase in binding affinity that is brought about by the second RAL moiety in conjunction with a short spacer.

Bivalent ligand **14** marks a special case, as it achieves almost the similar RBA as its monovalent counterpart, **23** (both **14** and the monovalent control **22** feature an EG6 spacer). This suggests that the binding mode of **14** is truly monovalent, and does not involve additional interactions of the second RAL moiety. Considering previous findings about the structure of ER, one can argue that the spacer length of bivalent ligand **14** is a critical length for distinguishing between two different ER binding modes for bivalent RAL ligands: The first one is an intramolecular bivalent binding mode that involves simultaneous binding to the steroid binding site of ER and a secondary site (or “subsite”) on the same ER α LBD monomer, while the second one is an intermolecular bivalent binding mode that is characterized by simultaneous binding to the two steroid binding sites on the neighboring sides of an ER α LBD dimer (Figure 4.18). The EG6 spacer of bivalent ligand **14** may represent a particularly unfortunate choice of spacer,

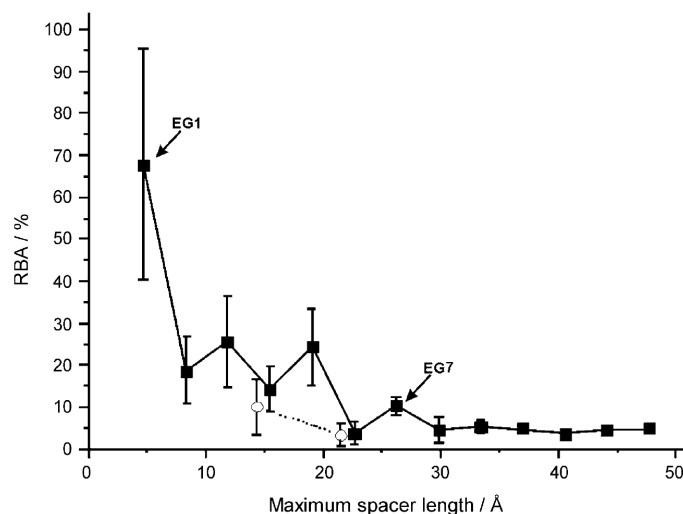


FIGURE 4.17: Relationship between the relative binding affinity and maximum spacer length of bivalent (**9–21**, ■) and monovalent ligands (**22**, **23**, ○); 100 % binding affinity for E₂. Figure taken from [16], reprinted with kind permission of John Wiley and Sons.

as it appears to be too long for intramolecular bivalent binding, but also too short for realizing intermolecular bivalent binding.

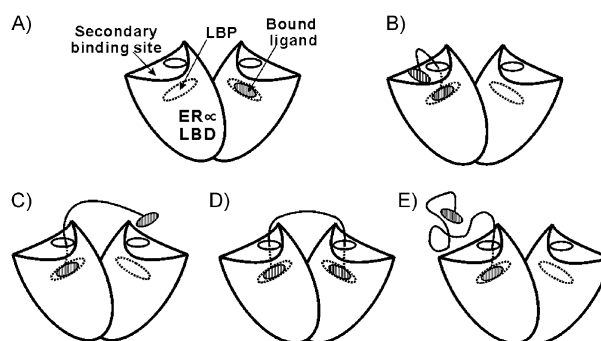


FIGURE 4.18: A) Schematic of the ER α LBD dimer in which each ER α LBD monomer features a primary (native) steroid binding site/ligand binding pocket (LPB) and a secondary site on the ER α LBD surface. B)–E) Four different binding modes for bivalent ligands binding to ER α LBD that demonstrate the progress from intramolecular to intermolecular bivalent binding. Figure taken from [16], reprinted with kind permission of John Wiley and Sons.

This notion is presented in Figure 4.18, B–E: Bivalent ligands tethered by short spacers (**9–13**) form intramolecular bivalent binding on the same ER α LBD monomer [17] (Figure 4.18 B). With increasing spacer length, binding affinity decreases as (i) the conformational entropy loss upon binding to the subsite accumulates and/or (ii) binding is hindered by a spacer-induced lack of fit to the subsite, and/or (iii) binding is rendered unfavorable by an enthalpic penalty on spacer unfolding. In the limit case (bivalent ligand **14**), no bivalent binding in either mode is achieved (Figure 4.18 C).

Consequently, the RBA of bivalent ligand **14** is in the order of the monovalent control, ligand **23**. Finally, bivalent ligands tethered by longer spacers (**15–21**) should in theory be able to achieve intermolecular bivalent binding (Figure 4.18 D), which would explain the local peak in RBA marked by bivalent ligand **15** (using an EG7 spacer). However, somewhat surprisingly, the RBA of all bivalent ligands using longer spacers than EG7 (**16–21**) drops almost to the unsatisfying level of bivalent ligand **14**. Here, again, (i) the high conformational entropy of longer OEG chains and/or (ii) spacer folding (and thus the enthalpic penalty on spacer unfolding) renders intermolecular bivalent binding unfavorable (Figure 4.18 E). By contrast, the theoretical benefit of longer spacers w.r.t. to allowing for bridging the steroid binding sites without causing structural strain (and thus without diminishing the enthalpy of the interaction) cannot be exploited in terms of RBA.

Under the premise that bivalent ligand **15** is able to realize intermolecular bivalent binding according to the above model of two different bivalent binding modes, there remain the questions (1) how this can be achieved, given that the EG7 spacer used in bivalent ligand **15** spans only 26.2 Å, in contrast to the requirement of 34.5 Å that was predicted based on crystal structure 2R6W (and supported by a previous study [134]), and (2) why the RBA values do not match the expectations sparked by high-affinity bivalent ligands that were previously reported for other multivalent target receptors [6, 56].

The first question is addressed in terms of a computational study on the bivalent intermolecular ER α binding mode of compound **15** in comparison to the binding mode of LY156681, presented in the next section. The answer to question (2), in turn, may be related to issues with the flexible spacer. Interestingly, in 2006 Kim *et al.* reported a study where E₂ was conjugated to PAMAM dendrimers through short (17 α -phenylethynyl) and long (EG6) tethers. They found that for the short tether, E₂ is “free and solvent exposed, while the long tether [...] wraps around the hormone and the dendrimer and markedly reduces access to the receptor.” [146] Furthermore, computational Study F1 on the conformations of flexible and semi-flexible spacers in water, involving a series of bivalent DES ligands (Section 4.2), had pointed out the risks of a hydrophobic collapse of the structures that arises from the interplay of spacer folding and hydrophobic (π - π stacking) interactions related to the tethered DES moieties. In order to validate this result, and thus to find a possible answer to question (2), solvent-dependent conformational changes of bivalent ligands **9**, **11** and **20** (representing short and long spacer variants) as well as monovalent ligand **22** were studied by NMR spectroscopy and UV-absorption experiments.

The ROESY spectra (ROESY = rotating-frame nuclear Overhauser effect correlation spectroscopy) reveal that (Figure 4.19 A), with an increasing proportion of deuterium water (0–40% D₂O), the intramolecular interaction between the RAL moieties (aromatic

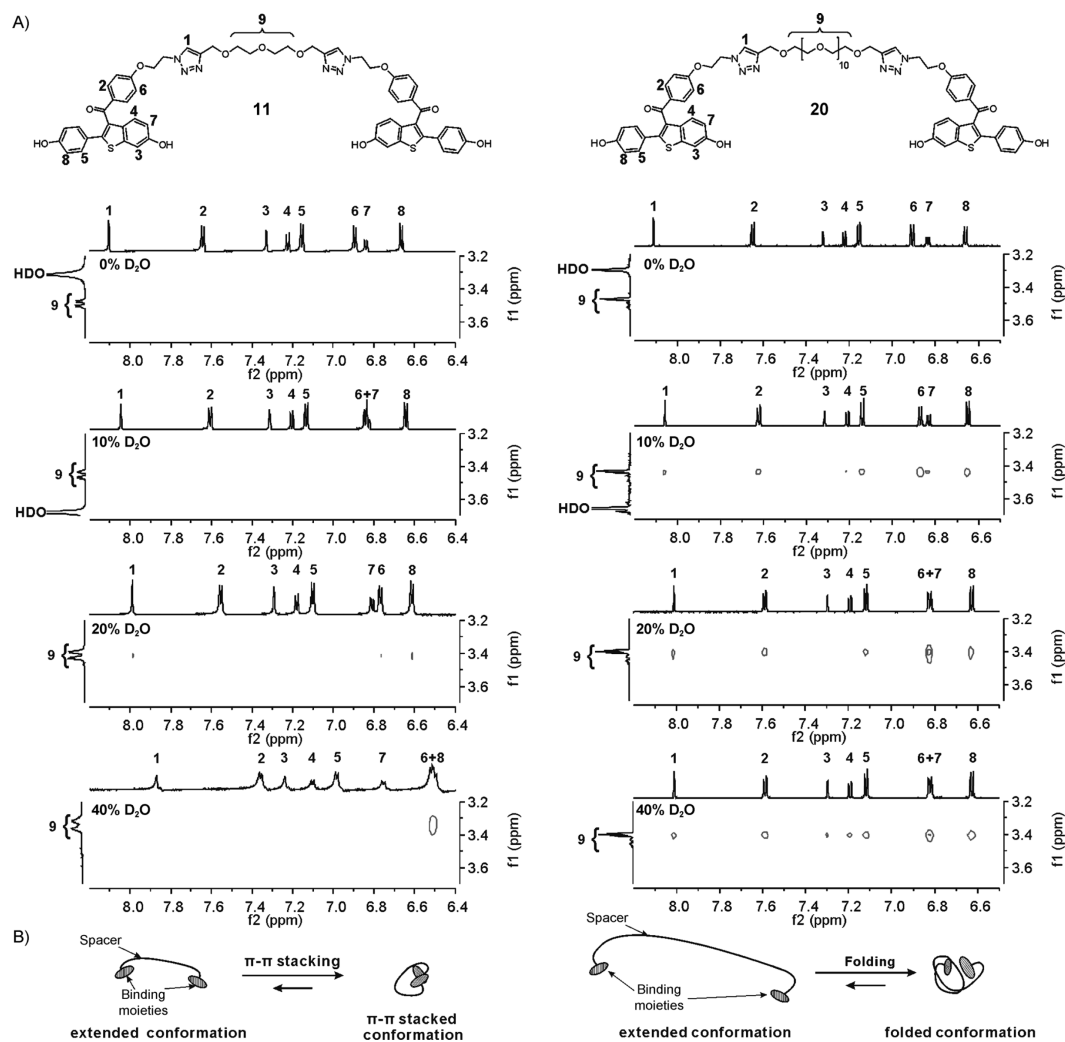


FIGURE 4.19: A) Aromatic (protons 1–8) and OEG (proton 9) regions of the ^1H NMR and ROESY spectra of OEG-tethered bivalent ligands **11** (left) and **20** (right) in a solution of $[\text{D}_6]\text{DMSO}$ with 0, 10, 20, and 40 % D_2O (top to bottom). B) Two proposed conformations of **11** (left) and **20** (right) in an aqueous environment. Figure taken from [16], reprinted with kind permission of John Wiley and Sons.

protons 1–8) and the tethered OEG spacer (alkyl proton 9) is becoming much stronger for bivalent ligand **20** than for bivalent ligand **11**. This is not surprising, as the spacer of **20** is long enough to wrap around its RAL moieties, while the spacer of **11** is not. In the case of bivalent ligand **11**, the significant upfield shift of the aromatic protons – in particular w.r.t. aromatic proton 6 – gives evidence for the prevalence of π - π stacked conformations. By contrast, the aromatic proton 6 of bivalent ligand **20** only undergoes a small upfield shift. This indicates that an intramolecular π - π stacking conformation between two RAL moieties dominates in the case of bivalent ligands tethered by short spacers, such as **11**, whereas spacer-wrapped conformations, which leave room only for weak π - π stacking interactions, are prevalent in the case of bivalent ligands tethered with longer spacers, such as **20** (Figure 4.19 B). Thus, one can argue that the ligand

moieties tethered by short spacers are more exposed to the solvent, and consequently, to the receptor as well. In order to confirm the solvent-dependent effect, the UV absorption of bivalent ligands **9** and **21** and monovalent ligand **22** was studied in different solvent systems.⁸ It was found that the UV-absorption maximum for each bivalent ligand in the aqueous solvent (3.85% DMSO in water) underwent a bathochromic shift ($\lambda_{max} = 290$ and 293 nm, for **9** and **21**, respectively) compared to the monovalent ligand ($\lambda_{max} = 287$ nm for **22**). In contrast, no bathochromic shift was observed in organic solvent (3.85% DMSO in chloroform). This confirms that the interaction between hydrophobic RAL moieties and the OEG spacer only dominates in aqueous solvent, and not in organic solvent, which is consistent with the observations in the NMR spectra.

In this respect, the computational predictions w.r.t. to the properties of flexible spacers presented in Study F1 (Section 4.2) are confirmed by the experimental results of this study. The fact that, upon binding to the target receptor, both π - π stacking as well as spacer folding/wrapping interactions have to be overcome before a bivalent interaction can be realized would account for the low RBA values. The findings are also in line with the results from the study by Kim *et al.* regarding the relationship of ER binding affinity and tether structure in E₂-PAMAM conjugates [146], and a report of similar hydrophobic ligand-ligand interactions observed in low-temperature NMR and X-ray analyses of bivalent GABA_A-benzodiazepine receptor ligands tethered by short spacers that resulted in low binding affinities [147]. In the following, the results of the computational part of this study are presented. Strategies to counter the problems that are associated with the use of flexible spacers are discussed in the chapter on rigid and semi-rigid spacers, Chapter 5.

4.3.3 Results and discussion (computational part)

Binding mode comparisons

In order to evaluate the steric fit w.r.t. to the target receptor, compounds **9** to **15** (covering spacer variants EG1 to EG7) were modeled and parametrized for MD simulations with the Amber-99SB force field. The unknown bivalent binding modes were modeled by first performing a monovalent alignment of the RAL moieties to the LY156681 ligands from the crystal structure with PDB ID 2R6W, followed by energy minimization in the presence of protein structure and explicit solvent in order to relax the OEG spacer portion of the bivalent ligand. The data suggests that bivalent binding to the ER steroid binding sites is not possible for compounds **9** to **14** (EG1 to EG6 spacers), as the insufficient length of the spacer prohibits the simultaneous servicing of both binding sites.

⁸More details on this experiment can be found in the Supporting Information of Ref. [16], along with a plot of the UV spectra (Figure S4).

Compound **15** (EG7 spacer) allows for bivalent binding to the steroid binding sites while leaving some room for conformational oscillations and binding mode adjustment (Figure 4.20). Given this result, one can assume that bivalent binding to the steroid binding sites is also possible for compounds **16–21**, which feature longer OEG spacers.

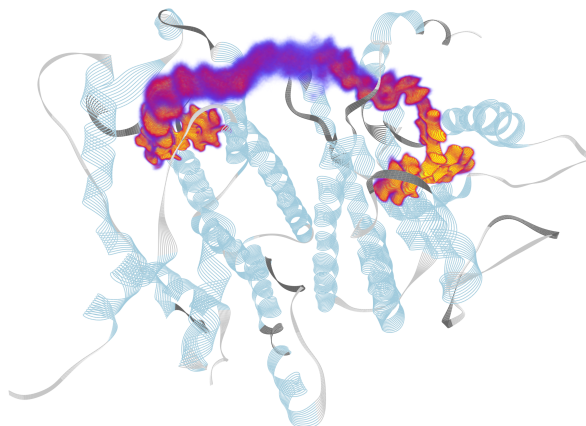


FIGURE 4.20: Conformational density of compound **15** bound to ER α , obtained from a 10 ns MD simulation in water. The conformational density of the RAL moieties (orange) is relatively compact, indicating tight fit to the steroid binding sites of ER. In contrast, the conformational density of the spacer portion (purple) turns out to be more diffuse, indicating that flexibility, especially in the central part of the chain, is retained. Figure taken from [16], reprinted with kind permission of John Wiley and Sons.

Results from a 10 ns MD simulation of compound **15** in the bivalently bound state reveal some differences to the monovalent binding mode (Figure 4.21). A hydrogen bonding interaction between the 4-methylpiperidin in the side chain of ligand LY156681 and residue ASP351 of ER α is characteristic for the monovalent binding mode of this SERM. The novel bivalent compounds presented in this study feature a 1,2,3-triazole ring structure at the same position, which is further connected to the OEG spacer. Simulations of compound **15** show that the 1,2,3-triazole ring is hindered to interact with residue ASP351 because the connecting EG7 spacer forces the linker groups to arrange towards each other.

In this respect, the spacer appears to have an disadvantageous impact on the enthalpy of the ligand-receptor interaction. To look deeper into the different binding modes for mono- and bivalent case, the positioning of two binding moieties of compound **15** bound to the ER α dimer (as obtained during MD simulation) was compared to the positioning of the two LY156681 ligands (as found in the crystal structure with PDB ID 2R6W) in terms of a set of distances measured at different positions, as indicated in Figure 4.22. While the distances between the major parts of the binding moieties of compound **15** (positions 1 to 4) change only slightly (0.01 – 0.24 nm), the distances between the ligand side chains of LY156681 and the linker groups of **15** (positions 5 to 7) show a

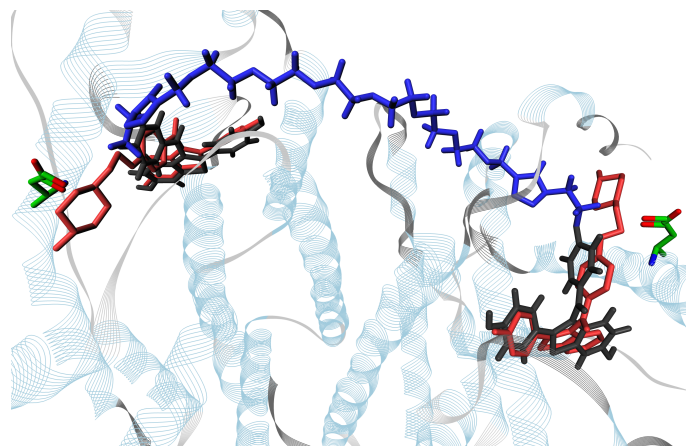


FIGURE 4.21: Overlay of bivalent ligand **15** (RAL moieties black, 1,2,3-triazole linker part and spacer in blue) with LY156681 (red), bound to the steroid binding pockets of ER α . The side chain of residue ASP351 is highlighted (PDB ID 2R6W). The snapshot was obtained after energy minimization and short MD simulation in water.

notable deviation (0.35 – 1.45 nm). In the latter case, a closer arrangement especially for the 1,2,3-triazole parts becomes apparent, allowing for some relaxation of the attached spacer.

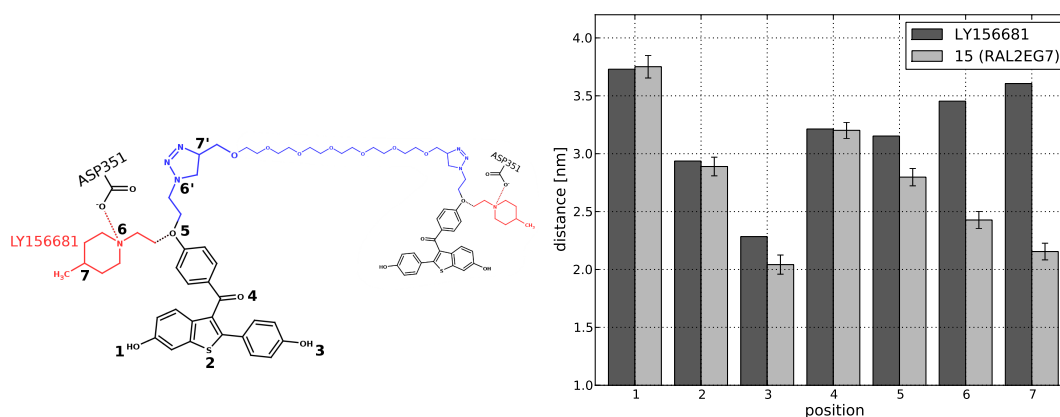


FIGURE 4.22: Distance comparison between two LY156681 ligands and two RAL moieties as part of bivalent ligand **15** while bound to the ER α dimer. Values obtained by 10 ns of MD simulation in water (blue) are plotted against values taken from the crystal structure (red).

The result shows that, despite its short spacer length of only 26.2 Å in the linear conformation, compound **15** is able to achieve bivalent binding to both steroid binding sites, however at the sacrifice of not being able to fully interact with residue ASP351. One might argue that this effect should be somewhat lessened when longer spacers are applied (compounds **16–21**).

Conformational entropy estimation

Following the sterical considerations, it was looked into how the length of the applied OEG spacers has an impact onto the loss of conformational entropy upon binding to the receptor. For this purpose, MD simulations of compounds **9–15** in the unbound state as well as in the receptor bound state were performed, starting with one ligand bound and the second ligand dangling in the solution. Based on the cumulative sampling data of 100 ns of MD per system (i.e. 200 ns of MD per unbound/bound pair), an estimate of the conformational entropy loss upon binding, regarding both RAL moieties and spacer part, was conducted (Figure 4.23).

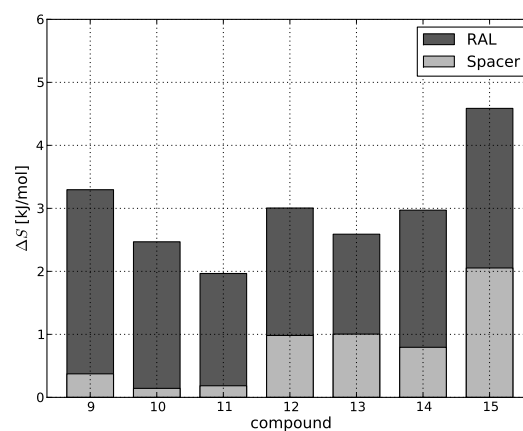


FIGURE 4.23: Conformational entropy loss of bivalent ligands **9–15** upon binding to $ER\alpha$, subdivided into the contribution of spacer part (blue) and RAL moiety (red).

The above results indicate that the major part of conformational entropy loss upon binding (ΔS_{conf}) stems from the RAL parts (including the 1,2,3-triazole ring) with a mean value of 2.19 kJ/mol. Not surprisingly, compound **9**, which has the highest RBA, also has the highest ΔS_{conf} value in the RAL moieties (2.92 kJ/mol), as it achieves intramolecular bivalent binding to the steroid subsite and a proximate hydrophobic subsite near the coactivator binding region with tight fit.⁹ Compound **15** shows the second highest ΔS_{conf} value of 2.53 kJ/mol for the RAL moieties.

The spacers' contribution to conformational entropy is comparatively low (mean value 0.79 kJ/mol), but has a higher spread than the conformational entropy of the RAL parts (standard deviation 0.66 kJ/mol for the spacers, compared to 0.45 kJ/mol for RALs). The data strongly suggests that conformational entropy loss increases with spacer length, with compound **15** (EG7 spacer) having the highest ΔS_{conf} value of 2.05 kJ/mol (spacer only). Hence, for longer spacers, conformational entropy loss of the spacer will equalize

⁹Subsite binding for this compound is demonstrated by means of molecular modeling in Study R2 (Chapter 5.3), which was published shortly before Study F2.

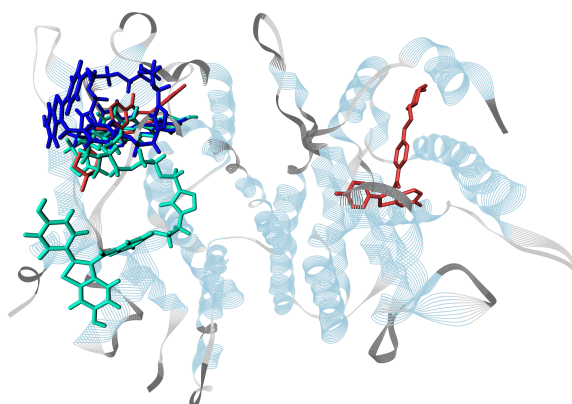


FIGURE 4.24: Different sampling outcomes for compounds with short spacers: Bivalent ligand **12** (EG4 spacer, blue) folds into a compact conformation over the steroid binding site. Bivalent ligand **13** (EG5 spacer, cyan) achieves intramolecular bivalent binding to steroid binding site and the hydrophobic subsite. An alignment with the LY156681 ligands from the crystal structure is shown in red. Figure taken from [16], reprinted with kind permission of John Wiley and Sons.

the improved ability to reach both steroid binding sites. This is also supported by the RBA measurements, where no distinct peak appears for spacers longer than EG7. Despite its short spacer length, compound **9** also shows a relatively large ΔS_{conf} value (0.37 kJ/mol), again due to its tight fit to both steroid binding site and subsite region. In contrast to bivalent ligands **9** and **15**, compounds **11** and **13**, which also show elevated RBA, only have relatively small ΔS_{conf} values. This might be either due to incomplete conformational sampling (i.e. the simulation did not capture realistic binding minima) or the predicted binding modes indeed allow for increased conformational entropy and hence lead to comparatively high RBA values.

In addition to differences in ΔS_{conf} values, other factors may be even more decisive for achieving multivalent binding. The simulations confirmed the propensity of OEG spacers to fold into helical loops, even with one ligand bound to the receptor binding site (Figure 4.24). In the Study F1 presented earlier (cp. Section 4.2), it was shown that hydrophobic ligands tend to stabilize compact, folded conformations of OEG spacers. The enthalpic penalty of unfolding in order let the unbound ligand sense for a second binding site on the receptor surface might be the limiting factor of the binding process, with the entropy loss of unbound ligand compared to bound ligand being only an additional (but minor) factor. Hence, bivalent ligand **9** may be the most successful compound w.r.t. RBA as its spacer is too short for adapting a folded conformation, but long enough to bridge steroid binding site and hydrophobic subsite. This line of thought is picked up in Study R2 (Chapter 5.3).

4.4 Conclusion

In conclusion, the cumulative results of the binding study confirm the doubts that were raised by the first study: Flexible spacers such as OEG are a problematic choice for tethering ER ligands (as is mirrored in the low RBA values), and quite possibly for tethering hydrophobic ligands in general. The general suggestions for spacer design that were formulated in Section 4.2.4 can therefore be maintained. An additional – and somewhat surprising – result of the second study is the finding that the conformational entropy loss in the OEG spacers upon binding to the receptor is comparatively low (albeit increasing with spacer length). The main burden of conformational entropy loss upon binding is suffered by the RAL moieties, and not by the OEG chain. Most likely, this can be attributed to two main factors, namely (i) the conformational entropy of OEG in solvent is not as high as expected, due to the fact that its structure undergoes folding and thus limits the fluctuation of the torsion angles, and (ii) the conformational entropy loss of OEG upon binding is not as severe as apprehended. As soon as one RAL moiety is bound to the steroid binding site (and thus buried within the receptor structure), it is no longer available for hydrophobic π - π stacking, which in turn might restore a certain degree of flexibility to the OEG chain. Even for the intermolecular bivalent binding mode of compound **15**, which uses a conceivably short EG7 spacer that is likely to undergo structural strain in the process, the simulation results indicate that a notable amount of flexibility in the OEG chain is retained – in particular when compared to the RAL moieties, which are locked tightly in the steroid binding sites. Consequently, the enthalpic penalty of “unwrapping” a bivalent ligand tethered by a long spacer, or of “unstacking” the ligand moieties tethered by a short spacer appears to be the major obstacle in achieving higher RBA values, and not the conformational entropy loss upon binding. It appears that here a higher degree of preorganization is required.

The differentiation of two different bivalent ER binding modes, i.e. intramolecular (involving the steroid binding site and the coactivator binding site on the same ER LBD monomer) and intermolecular bivalent binding (involving the steroid binding sites on the neighboring sides of an ER dimer) that can be modulated by varying the length of the tethering spacer was one of the main results of the experimental binding study. The suggested mechanism could be supported by the findings of the computational study, which provided evidence for both the sterical availability of the coactivator binding site for bivalent ligands tethered by short spacers (EG1–EG5), as well as the possibility to bridge the steroid binding sites in terms of intermolecular bivalent binding with longer spacers (\geq EG7). Furthermore, the effects of attaching a spacer on the binding mode of the tethered RAL moieties could be illustrated, and differences to the binding mode of the unmodified monovalent RAL derivative LY156681 were pointed out.

One can also conclude that the ER dimer is not a particularly “low-hanging fruit” when it comes to the task of designing potent bivalent ligands. First of all, the known monovalent ER binders already are characterized by high binding affinities in the nanomolar regime, which does not leave plenty of room for improvement. In cases where evolution has resorted to the multivalent display of ligands and receptors, it mostly involves individually weak interactions. In addition, the buried steroid binding site of ER (in combination with the low receptor valency) is a disadvantageous environment for the occurrence of rebinding events, arguably the main source of the high stability of multivalent compounds (cp. Ref. [30]). Finally, the complex structural mechanism of ER activation, related to the flexible positioning of Helix 12, complicates the interpretation of binding studies.

This also applies to computational binding studies, as, for instance, the availability of the coactivator binding site for the binding of small molecules (e.g. the second ligand moiety of a bivalent ligand) is dependent on the positioning of Helix 12, which in turn is dependent on the ligand that is bound to the steroid binding site. Monitoring macroscopic structural changes in protein structures in the course of atomistic molecular simulations is still beyond reach (at least for screening and comparing larger series of ligands), so that a good choice of the initial protein crystal structure is of great importance. Last but not least, the high number of additional degrees of freedom that are intrinsic to flexible spacers adds to the overall complexity. In this case, it appears that the shortcomings of flexible spacers render the process of rational molecular design almost impossible. It is, for example, very difficult to screen for the optimal distance of bivalent ligand presentation without a strong bias caused by different effects such as spacer folding or conformational entropy loss.

The next chapter will present a different concept of displaying ligands in multivalent arrays that is based on the use of rigid spacers and scaffold structures that can be complemented by optional flexible elements in a modular fashion. Again, molecular simulation is used to provide an insight into these systems on the atomistic level.

4.5 Experimental setup

4.5.1 Modeling and simulation

The structures of proteins, spacers and bivalent ligands were modeled using the visualization software Amira [139]. All simulations were performed using the software GROMACS, version 4.07 [76] and the according port of the Amber force fields, *ffamber* [148], more precisely the Amber-99SB force field [149] in combination with either the TIP4P-Ew water model [150, 151] (Study F1) or the TIP3P water model [152] (Study

F2), respectively. All novel structures were parametrized using the software Antechamber [153, 154] from AmberTools 1.2 [155], with charges calculated by the AM1-BCC method [156, 157]. For Study F1, the structures were put into equally sized rhombic dodecahedron solvent boxes of about 9 nm side length. For Study F2, the non-complexed structures were placed into equally sized rhombic dodecahedron solvent boxes of about 6 nm side length. The complexed structures were prepared by aligning a pre-minimized conformation of the novel bivalent ligand to the SERM ligands as found in ER α crystal structure with PDB ID 2R6W [136]. For the starting configurations of the complexes, the raloxifene binding mode was modeled based on the known monovalent binding mode, with spacer and second ligand protruding into the solvent. The protein-ligand complexes were placed into equally sized rhombic dodecahedron solvent boxes of about 9 nm side length. The energy of the systems was minimized with the steepest descent algorithm, and afterwards 200 ps simulations were performed during which the positions of all heavy (non-hydrogen) atoms of ligand and protein (if any) were restrained in order to settle the solvent molecules. These systems were subsequently used as starting conformations for either one MD run of 100 ns length per structure (Study F1) or five MD runs of 20 ns length per structure (Study F2), the latter using a different initial impulse vector on every restart, leading to a cumulative sampling data of 100 ns of MD per ligand or complex. In order to maintain a constant temperature of 300 K and a pressure of 1 bar, velocity rescaling [86] and Berendsen weak coupling [90] were applied. A twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm [75] was used for Coulomb interactions, with a switching distance of 1.0 nm. Bond lengths were constrained using the LINCS algorithm [84], allowing for an integration step of 2 fs.

4.5.2 Conformational entropy estimation

Conformational entropy differences were estimated using the density-based approach of Weber and Andrae [108] (cp. Chapter ??), based on the cumulative sampling data of 100 ns per system that was obtained from thermostated MD simulations as described in the previous section. In contrast to using (subsets of) Cartesian coordinates for describing conformational states as suggested in Ref. [108], a definition in terms of internal coordinates was preferred in order to avoid translational and rotational bias (cp. Chapter 3.2.2). The set of internal coordinates consisted of all rotatable torsion angles of the spacer as well as the spacer end-to-end distance (Study F1) or all rotatable torsion angles of the bivalent ligand, subdivided again into spacer and RAL moieties (Study F2). Consequently, conformational distances between structures were calculated with a metric in internal coordinate space, and not in terms of Cartesian root-mean-square

deviation (RMSD) values as in the original Ref. [108]. The size of the evaluation region used for density estimation was chosen according to the variance of the internal coordinates w.r.t. the cumulative sampling data. For Study F2, the first two nanoseconds of each MD run were discarded in order to allow for relaxation of the systems. Per data set, 500 sampling points with approximately mean potential energy were chosen as reference points. Triple bond torsion angles (in 2-butyne moieties) as well as delocalized bond torsion angles (in phenyl groups) were accounted for with an entropy contribution of zero.

In order to obtain a meaningful estimate of conformational entropy differences with the above method, the volume of the conformational spaces under observation has to have a comparable, if not equal, size. This notion is best implemented by picking an equal number of (internal) coordinates for the density estimate. Despite the structural similarity of the compounds under observation in Study F1, a slightly different number of rotatable torsion angles per structure was obtained (ranging from 33 to 35). To improve the comparability of results, the number of torsion angles for the density estimate was padded up by counting either one or two standard EG backbone torsion angles twice, where necessary, therewith artificially prolonging the EG portion of the spacer in a relatively minor way. In Study F2, the systems under comparison were structurally equal by definition (unbound state compared to bound state w.r.t. the same bivalent ligand), so that no padding of the internal coordinates was necessary.

4.5.3 Visualization of conformational density

Conformation density plots were computed with the approach of Schmidt-Ehrenberg *et al.* [158] which is available as part of the visualization software Amira [139]. Rather than visualizing conformations in terms of a single representative, probability densities over all molecular geometries in the sampling data are accumulated and visualized by volume rendering. This requires the application of a suitable alignment of the geometries in advance, preferably along a rigid part of the molecule. Visualizing molecular conformations in terms of density plots helps to incorporate the notion of a diverse statistical ensemble of states into a single static representation. It is helpful for discriminating flexible and rigid regions in conformational space, but also meets the idea of conformation dynamics (cp. Chapter 3.2.1), which perceives (and identifies) chemical conformations in terms of softly overlapping metastable regions in conformational space.

Chapter 5

Rigid spacer systems

5.1 Introduction

One of the main conclusions of the previous chapter was that, in order to facilitate the design of successful synthetic multivalent compounds, a modular approach of spacer and scaffold design would be the most promising: For bridging distances and pre-organizing ligands, the spacer or scaffold ought to be as structurally defined as possible, only to be complemented with flexible elements if necessary, e.g. with short flexible linkers for allowing access to the binding pockets of the target. The multivalent enhancement effect stems from the preorganization of ligands for their receptors, and, given the structure of the multivalent target is known, this idea should be exploited. One of the main arguments for using completely flexible spacers is that they compensate for structural uncertainty w.r.t. to the target, e.g. if the distance between binding sites is unknown, or prone to change for structural reasons such as flexible interdomains. One can turn around this argument by asserting that in some cases, flexible spacers can even add to the uncertainty due to their rather unpredictable behavior – such as solvent-dependent folding and wrapping of the attached ligand moieties (cp. Chapter 4.3).

This chapter presents two studies that involve spacers that are constructed from nucleic acids, a material that implements the concepts of modularity and rigidity in an elegant fashion. This is particularly interesting in the respect that nucleic acids themselves may have been the the first self-assembling structures that exploited the concept of multivalency (more precisely, multivalent base pairing and stacking) and, given the fact they are conserved throughout evolution, are a living proof of its success.

Due to the fact that nucleic acids assemble from small nucleotide subunits that can be associated with a defined bridging distance, they allow for a higher degree of preorganization as it can be implemented by the use of completely flexible spacer structures.

The proof that nucleic acids can be arranged into complex two and three-dimensional shapes on the nanoscale was provided in 2006 by the article on DNA origami by Rothemund [159]. Nowadays, it is possible not only to synthesize native nucleic acids at high rates, or to create synthetic variants of nucleic acids with certain select properties, such as peptide nucleic acids (PNA) [160], but also to modify nucleic acids such that various kinds of functional groups can be attached. This opens up the possibility to use molecular architectures based on nucleic acid for the multivalent presentation of ligands, a notion that according to Scheibe *et al.* [161] offers intriguing opportunities, “because (a) oligonucleotide synthesis provides monodisperse material, (b) the valency of the ligand display can readily be controlled by programmed self-assembly based on nucleic acid hybridization and (c) the well-known base-pairing rules allow Ångström-scale positioning of functional groups along the rigid nucleic acid helix.”

Both studies presented in this chapter are collaborative works of mainly experimental nature that were conceptualized by chemists and accompanied by a computational part in order to facilitate the interpretation of the results. The experimental part of the first study (Section 5.2, based on Ref. [161]), was conducted by Christian Scheibe¹ at Humboldt-Universität zu Berlin, and focuses on the design, synthesis and application of multivalent scaffolds assembled from DNA and PNA templates for the multivalent presentation of glycoligands to lectin binding sites. The main idea here is to show that linear PNA-DNA complexes with glycoligands attached at varying distances can be used as “molecular rulers” for the spatial screening of binding sites on targets of possibly yet unknown structure.

The experimental part of the second study (Section 5.3, based on Ref. [17]), was conducted by Frank Abendroth² at Humboldt-Universität zu Berlin and Min Shan³ at Freie Universität Berlin, and deals with the DNA-controlled bivalent presentation of ligands for estrogen receptor (ER). Due to the fact that Study R2 includes the same ligand moieties as the ER binding study involving bivalent ligands tethered by flexible spacers (Study F1, cp. Chapter 4.3), it offers the welcome opportunity of conducting a direct comparison between flexible and rigid spacers w.r.t. to ligand-receptor binding affinities and binding modes obtained under the same conditions.

¹responsible for concept, synthesis and SPR measurements, supervised by Prof. Dr. Oliver Seitz (also concept) at Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin (Germany), co-supervised by Dr. Jens Darnedde at Zentralinstitut für Laboratoriumsmedizin und Pathobiochemie, Charité Universitätsmedizin Berlin (CBF), Hindenburgdamm 30, 12203 Berlin (Germany)

²responsible for concept, synthesis and relative binding affinity measurements, supervised by Prof. Dr. Oliver Seitz (also concept) at Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin (Germany)

³responsible for synthesis and relative binding affinity measurements, supervised by Prof. Dr. Rainer Haag at Institut für Chemie und Biochemie, Freie Universität Berlin, Takustraße 3, 14195 Berlin (Germany)

Given the fact that, for instance, DNA double helices are relatively rigid structure with well-defined conformational properties, the need for predictive molecular simulation in the context of spacers and scaffolds based on nucleic acids can be questioned. Parameters that are crucial for ligand presentation, such as the canonical raise per base pair (base-to-base distance), helical twist per base pair, or helix diameter, can readily be found in the literature. Furthermore, the explicit solvent simulation of DNA macromolecules is an expensive task in terms of the required computing power, as the large linear molecules require spacious solvent boxes.

Still, molecular simulation of nucleic acids becomes worthwhile as soon as flexibility is introduced to the systems. As will be pointed out in this chapter, one of the main advantages of using nucleic acids as molecular spacers/scaffolds is the possibility that structural flexibility can be reintroduced into the architecture “in small doses” where needed, e.g. for wrapping around a spherical target protein in order to gain access to its binding sites (Study R1, cp. Section 5.2) or in order to avoid sterical clashes with the protein surface in the bound state (Study R2, cp. Section 5.3). Therefore, one of the main objectives of the computational studies presented in this chapter is to evaluate the influence of structural elements such as backbone nick sites or unpaired regions of varying length on the properties of the spacer/scaffold and thus to facilitate the rational design of such compounds in the future.

A second objective of molecular simulation in this context is related to the modified nucleotides that are used for tethering the ligands to the spacer/scaffold. Typically, the structural properties of the modified nucleotide and linker are less well studied, so that a binding simulation to the actual protein surface or binding pocket can provide additional insights. This might, for instance, reveal unwanted effects such as ligand moiety-spacer/scaffold interactions as discussed at length in the previous chapter. Again, the author of this thesis is responsible only for the computational part of these studies (Sections 5.2.3 and 5.3.2), but the experimental results as contributed by the collaborating scientists (see remarks above) are summarized in order to provide the necessary context.

5.2 Study R1: DNA-programmed spatial screening of carbohydrate-lectin interactions

This section is based on (and contains content from) the following publication:

- Ch. Scheibe, A. Bujotzek, J. Dervede, M. Weber, O. Seitz: DNA-programmed spatial screening of carbohydrate-lectin interactions. *Chem. Sci.*, 2:770-775, 2011. URL <http://pubs.rsc.org/en/content/articlelanding/2011/sc/c0sc00565g>.

5.2.1 Carbohydrate interactions benefit from multivalent presentation

Interactions between proteins and carbohydrate ligands are often characterized by low binding affinities, which, in the cellular context, are typically compensated by the multivalent presentation of both carbohydrate ligands and carbohydrate binding sites [162, 163]. Apart from the valency of the multivalent interaction, the distance and orientation at which the carbohydrate (glyco) ligands are presented, and thus preorganized, are factors that determine the resulting overall affinity and specificity of the interaction. Arguably, for interactions of low valency – involving only a limited numbers of glycoligands – the role of preorganization that is needed for achieving a sufficient binding affinity becomes increasingly important. Significant efforts have been put into the synthesis of low-valency glyco structures that allow for the precise control of the presentation of the individual ligand moieties, which in turn creates the possibility to interrogate the spatial arrangement of the binding sites of the targeted lectin [164]. However, highest binding affinities are usually achieved by using polyvalent glyco clusters based on multivalent scaffolds such as polymers [165–167], dendrimers [168–170], and nanoparticles [171–173]. A precise control of number, orientation and distance between the presented ligand moieties is often difficult to realize in these glyco clusters.

Consequently, the use of well-defined and highly preorganized molecular architectures based on nucleic acids seems to be a promising alternative for the multivalent display of glycoligands. Kobayashi explored self-organized, high molecular weight DNA-galactose clusters of high periodicity and studied the effect of the helical torsion of glycan display on cooperative lectin recognition [174, 175]. More recently, Winssinger described the hybridization of end-labeled PNA carrying flexibly tethered bivalent hexamannose units [176], a study that suggested that nucleic acid hybridization can be used to mimic complex carbohydrate epitopes. The study that is presented in the following extends the concept by using DNA-programmed multivalency as a tool for the interrogation of the spatial arrangement of lectin binding sites.

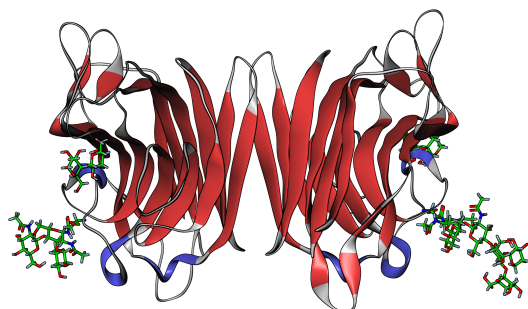


FIGURE 5.1: Crystal structure of ECL with *N*-linked oligosaccharide and lactose, bound at opposite sides of the dimer at a distance of approximately 65 Å [177]. This arrangement of binding sites is not optimally bridged with a rigid linear spacer.

One problem that has to be overcome in this context is that lectins, being roughly spherical proteins, have their binding sites arranged over a convex surface – a fact that calls for ligand displays from either concave, a structural feature that can be difficult to implement, or flexible scaffolds. Previous systematic screenings of distance dependencies in multivalent systems have relied on flexible scaffolds [5, 178–180], cyclodextrins [181], and cyclopeptides [6, 182–184]. Due to the fact that the rigid linear conformation of the double helix (by design an almost perfect “molecular ruler”) is a disadvantage in this situation, the nucleic acid architectures in this study were designed to contain semi-rigid regions in order to avoid torsional and bending stress upon multivalent binding to lectins with convex surfaces. The main objective of the computational part of this study, complementing the experimental section, is to evaluate the effectiveness of these designs. The approach is based on the use of PNA conjugates (Figure 5.2 I) that contain *O*-linked *N*-acetyl-lactosamine (LacNAc) residues at internal positions. The formation of different multivalent complexes w.r.t. to the distance between the LacNAc residues (Figure 5.2 II and III) can then be controlled by hybridization with selected DNA template strands – and thus with a minimum of required synthesis. The binding behavior of the different complexes is evaluated using the LacNAc-specific *Erythrina cristagalli* lectin (ECL) [177] (Figure 5.1).

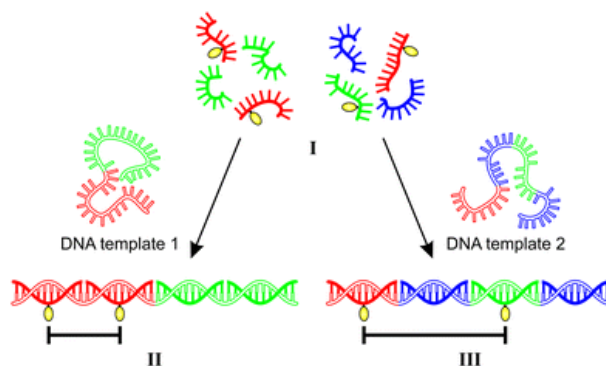


FIGURE 5.2: Modular assembly of multipartite PNA-DNA complexes to control the valency and spatial arrangement of a multivalent display of glycoligands (yellow). Figure taken from [161], reprinted with kind permission of Royal Society of Chemistry.

5.2.2 Design and evaluation of LacNAc-PNA-DNA complexes

PNA was chosen as the scaffold of choice for this application because PNA-DNA duplexes have a higher stability than DNA-DNA duplexes, and are not susceptible to enzymatic degradation by proteases [185], a feature that would facilitate prospective studies in biological environments. The construction of the multivalent LacNAc assemblies follows a modular approach that allows for the rapid and facile assembly of a wide

range of supramolecular structures with defined valency. The key element is a modified, cysteine-derived PNA monomer that permits the conjugation of LacNAc residues via thioether formation. PNA oligomers that can carry the modified residue are then hybridized to complementary DNA strands – a convenient means to “program” the presentation distance between the LacNAc ligands.⁴ From five different PNA-oligomers with three independent anticodon sequences (two anticodon sequences were used twice, once with LacNAc attached and once in an unmodified version) and DNA templates that provide three codon segments distributed over four positions, in principle a total of 324 different complexes can be formed. Only a fraction of these complexes was evaluated in the binding studies. The distances between the LacNAc residues were estimated based on the known solution NMR structure of PNA-DNA duplexes, i.e. a helical twist that comprises 13 base pairs with a pitch of 42 Å [186]. Accordingly, PNA oligomers of length 13 were synthesized. In order to correct for smaller discrepancies in the parallel presentation of the LacNAc-ligands arising from helical torsion, some flexibility is given by the maleimido-linker used for tethering thiol-modified PNA and ligands, as well as by the possibility to incorporate nick sites into the backbone (cp. Figure 5.2 III).

ECL was chosen as the model substrate for testing the binding affinities of the novel compounds by surface plasmon resonance (SPR) because its carbohydrate binding affinities and structure have been well characterized [177]. The lectin exists as a homodimer in which the binding sites for galactose-containing carbohydrates, including LacNAc, are located on opposite sides of the protein (cp. Figure 5.1). The distance between the binding sites in the crystal structure is approximately 65 Å. However, given that a linear scaffold that presents multiple ligand moieties for simultaneous binding to ECL would have to adapt a bent conformation in answer to the convex receptor surface, a larger presentation distance (in the range of approximately 100 Å) would be in order.

For the initial SPR experiments, the density of immobilized lectin on the gold surface was set to approximately 2700 RU (1 RU corresponds to ca. 1 pg of lectin per mm²) [187] and the influence of the number of LacNAc ligands on the binding behavior was investigated.⁵ As expected, no binding was found for the PNA-DNA duplex carrying no LacNAc ligands, and complex **17**, presenting only a single LacNAc residue, was revealed to have a low binding affinity ($K_D = 800 \mu\text{M}$). For complex **18**, in turn, involving two different PNA-LacNAc conjugates, the SPR sensorgram reveals a 33-fold enhancement of binding affinity per ligand ($K_D = 12 \mu\text{M}$), a trend that is followed by trivalent complex **19** ($K_D = 2.6 \mu\text{M}$) and tetravalent complex **20** ($K_D = 1.1 \mu\text{M}$), the latter showing a 182-fold enhancement in binding affinity compared to the monovalent complex **17** (for

⁴For details on the synthesis, as well as for the sequences of PNA oligomers and DNA templates, please refer to the original Ref. [161].

⁵For a full report on the SPR measurements, including kinetic data, please refer to Ref. [161], Table 1, Figure 2, and the associated electronic supplementary information (ESI), Table 2, Figures S1-S21.

structural representations of complexes **17**, **18**, **19** and **20**, cp. Ref. [161], Table 1). Interestingly, in the initial SPR measurements, little to no difference in the binding affinities of complexes **18** and **22–27** was observed (Figure 5.3, gray bars), despite the fact that the distance of LacNAc presentation in this series is varied from 42 Å (complex **22**) to 146 Å (complex **27**). It was soon suspected that the high density of ECL molecules immobilized on the sensor chip promotes the occurrence of cross-ECL binding between adjacent receptor molecules, and thus cancels out the differences w.r.t. to simultaneous recognition of two binding sites on a single ECL molecule. It was shown earlier that crosslinking of lectin molecules, also referred to as aggregation, can govern binding arrangements at high concentrations [188]. With the intention of reducing crosslinking, the density of immobilized ECL was reduced to 700 RU, and indeed, this step uncovered the differences in the binding behavior of the bivalent complexes **18** and **22–27** (Figure 5.3, black bars): The dissociation constants reach an optimum for LacNAc-LacNAc presentation distances between 104 and 127 Å, and deteriorate for presentation distances that are either significantly shorter or longer. The decrease in binding affinity is less striking for long presentation distances, which suggests that an increase in crosslinking probability might partially compensate for unfavorable single-ECL fit. In summary, an obvious relationship to the spatial arrangement of the binding sites of ECL (at a distance of ca. 100 Å, including compensation for convexity) can be established.

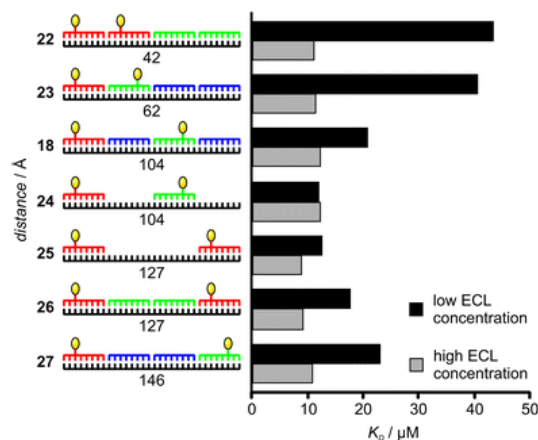


FIGURE 5.3: Influence of the LacNAc-LacNAc distance on the binding affinities of bivalent complexes **18** and **22–27**. The canonical distance of ligand presentation is given in Å below the structural representations of the complexes. Figure taken from [161], reprinted with kind permission of Royal Society of Chemistry.

In addition to the distance between the LacNAc ligands, the flexibility of the complexes also appears to be a factor that influences binding to ECL. Initially, it was envisioned that nick sites in the PNA backbone (i.e. at the end of each PNA 13mer) introduce local flexibility to the structure and thus facilitate bending of the PNA-DNA complex around the protein. In fact, complexes **24** and **25**, arguably the most flexible structures

in the comparison due to their unpaired regions of the DNA template (complex **24**: 2×13 bp unpaired strand, complex **25**: 1×26 bp unpaired strand) showed the smallest dissociation constants of all tested bivalent substrates. This indicates that unpaired single strand regions bend more readily than nick sites in duplex regions (complexes **18**, **22**, **23**, **26** and **27**) and thus allow complexes **24** and **25** to better adapt the curved conformation required for simultaneous binding to the ECL binding sites. The objective of the following computational study was to describe (and ideally quantify) the structural origin of this phenomenon.

5.2.3 Results and discussion (computational part)

The main obstacles for molecular simulation in this context was given by the lack of software tools for generating three-dimensional structures of PNA-DNA duplexes from sequence, and, more severely, the lack of force field parameters for PNA, in particular for the basic PNA residues needed for pairing with the standard DNA base set.

First, the PNA-DNA duplex solution structure (PDB ID 1PDT), obtained from the Protein Data Bank, was aligned to a generic (ideal) DNA double helix created with the software `nab` from AmberTools 1.2 [155]. By energy minimization with position restraints using the modeling and visualization software `Amira` [139], the (somewhat warped) solution NMR structure was rectified and turned into an idealized version of the PNA-DNA duplex. Overlapping substructures of the generic duplex structure were cut out and used as templates for force field parametrization of the PNA residues with the bases adenine, cytosine, guanine and thymine, as well as a number of capping residues for N and C-termini. All residues were parametrized for the Amber-99SB force field [149] using the software `Antechamber` [153, 154], with charges calculated by the AM1-BCC method [156, 157]. The novel PNA force field parameters were carefully adjusted to match both the existing peptide backbone parameters as well as the canonical parameters of the DNA bases. Finally, the modified PNA residue containing the maleimido-linker was modeled with `Amira` and parametrized according to the above protocol. The attached LacNAc ligand moiety was not explicitly modeled. The complete set of PNA parameters was included into the residue database of the Amber-99SB force field for the MD software `GROMACS`, therewith enabling its native tool `pdb2gmx` for the preparation of PNA-DNA complexes for simulation. In order to be able generate the relevant PNA-DNA complexes from sequence, a Python wrapper for `nab` was written⁶ that replaces one DNA backbone in a custom DNA double helix structure with a matching PNA backbone (unpaired regions are supported as well). The resulting structure/parameter

⁶courtesy of Ole Schütt

combination of the PNA-DNA duplex was evaluated in MD simulations and showed reasonable stability and agreement with the solution NMR structure.

Subsequently, MD simulations in explicit water were carried out in order to investigate the influence of nick sites and unpaired regions on the flexibility of the nucleic acid structure in more detail. For this purpose, the structures of complexes **18**, **24** and – a purely hypothetical – complex **28** were generated and prepared for simulation. These complexes possess an equal number of bases between the LacNAc ligands and thus qualify for a direct distance comparison. Complex **28** shares the same sequence as complex **18**, but has a continuous PNA backbone without nick sites. For further reference, complex **22**, featuring a short LacNAc-LacNAc distance, was prepared for simulation as well. The MD simulations were started from the linear complex conformation. Due to the fact that the DNA templates have a length of 52 bp, the use of rather large solvent boxes of about 20 nm side length was required. After an initial 10 ns for equilibration, the simulations were extended to a length of 30 ns in total. Although this simulation length is probably not sufficient to sample the complete conformational space of such large complexes (in particular w.r.t. to the single strand segments), it can be assumed that differences in flexibility are revealed.

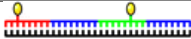

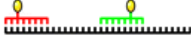

Complex		Distance μ / σ [Å]	Torsion μ / σ [°]
28		102.89 ± 3.86	129 ± 3.38
18		98.77 ± 5.46	112 ± 4.80
24		109.67 ± 5.54	59 ± 7.60
22		43.58 ± 3.37	35 ± 2.37

TABLE 5.1: LacNAc-LacNAc presentation distances and angles as measured during 20 ns of MD simulation (mean μ and standard deviation σ).

The most rigid complex (**28**) shows a mean LacNAc-LacNAc distance that hardly differs from the estimated, canonical distance of 104 Å, which confirms that the force field parameters are in order (Table 5.1). For the flexible complexes **18** and **24**, a deviation of about 5 Å from the predicted value can be found. While the probability for adapting bent conformations – and thus for displaying shorter LacNAc-LacNAc distances – should increase for both complexes, only complex **18** shows an actual decrease in LacNAc-LacNAc distance: Due to an increase in base to base distance and a decrease in helix diameter, the single strand regions of complex **24** expand to a larger degree than double strand segments, and thereby (over)compensate the distance decrease that is connected with the increase in bending propensity – at least in the limited time frame that is accessible to the simulation. Notably, even the most rigid complex (**28**) shows a certain amount of oscillations w.r.t. to the ligand presentation distance, leading to a

standard deviation of 3.86 Å. This corresponds to a contribution of 0.12 Å of “uncertainty” per residue in a non-interrupted backbone (the distance between the two ligands in this complex is 31 bp). The standard deviation of the ligand presentation distance for complexes **18** and **24** is in the same order (≈ 5.5 Å), a fact that underlines the effectiveness of nick sites.

The MD simulations also suggest differences in the dispersion of the torsion angle between the LacNAc presentation sites. While for double stranded complexes **18** and **28** the angle is in the same range, a clearly different value is found for complex **24**. This and the increased standard deviation of the torsion angle obtained for complex **24** shows that the single strand segment indeed confers a higher flexibility than double strands. The torsion angle standard deviation of complex **18** (two nick sites between the LacNAc moieties) is about double the value of complex **22** (one nick site between the LacNAc moieties), which suggests that the torsional freedom of the strand per nick site might be additive.

In a second step, it was evaluated to which extent nick sites and unpaired regions promote the formation of bent conformations as required for bivalent binding to ECL. As the initial 30 ns of MD only led to moderate bending of the strands, MD simulations were set up where a constant force acts on the strand so that the required LacNAc-LacNAc distance can be adjusted. After obtaining bent, distance-adjusted conformations for complexes **18**, **24** and **28**, short MD simulations were performed for measuring the force that is necessary to remain in these potential binding (i.e. correct LacNAc-LacNAc distance) conformations (Figure 5.4).

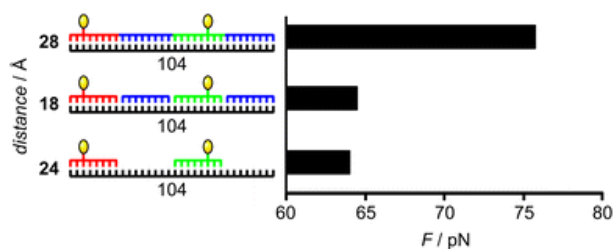


FIGURE 5.4: Mean force acting on the strand in the bent conformation with adjusted LacNAc-LacNAc distance, measured at the ligand attachment sites over 400 ps. Figure taken from [161], reprinted with kind permission of Royal Society of Chemistry.

These values give a measure of strand flexibility, as they can also be interpreted as the force that has to act on the strand to adopt the bent conformation. As expected, the highest force is obtained for the duplex with a continuous PNA backbone (complex **28**). The introduction of nick sites (complex **18**) or unpaired single strand regions (complex **24**) increases torsional freedom and the ability to stretch as the force needed to keep the complexes in the bent conformation decreases (Figure 5.5).

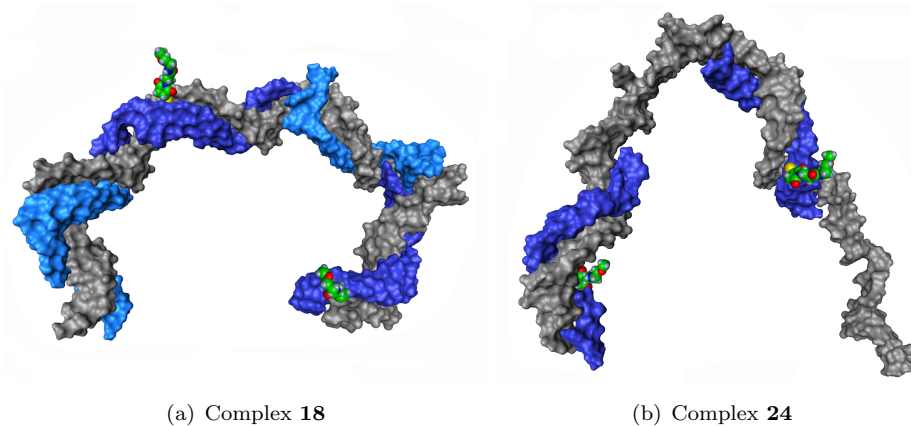


FIGURE 5.5: Surface representations of the complex conformations obtained by MD simulations in which a constant force acted on the strand to bend the complexes into a conformation suitable for bivalent binding to ECL. (a) In complex **18**, bending occurs mainly at the nick sites while the duplex segments remain in a linear conformation. (b) In complex **24**, the bent is formed by a nick in the flexible unpaired region (dark gray = DNA, blue = PNA strands, green = linker). Figure taken from [161], reprinted with kind permission of Royal Society of Chemistry.

The measured effect is slightly higher for the complex involving single strand regions (**24**). Hence, conformations suitable for bivalent binding to ECL (and arguably any spherical target protein) are more likely to be observed for these complexes. Given the relatively distinct difference in binding affinity w.r.t. to complexes **18** and **24** (cp. Figure 5.3), one could have expected a somewhat larger difference in mean force as well. Due to the fact that the bent conformations were generated by applying a rather artificial pulling force, the resulting complexes may not in detail represent realistic bent conformations as they appear in solvent (or possibly upon encountering of a target protein) without the application of artificial strain. Therefore, a “true” bent conformation of complex **24** may require even less force. Here, the large number of degrees of freedom in the unpaired regions of the DNA template is pushing the limits of molecular simulation.

For complex **22**, a bent conformation with the correct LacNAc-LacNAc distance could not be obtained – with an adjusted LacNAc-LacNAc distance of 65 Å, the strand is rather linear than bent – which may explain the comparatively poor binding affinity (Figure 5.6). In general, the observations from the MD simulations are in agreement with the results from the SPR experiments: The binding affinity data obtained from experiment coincides with the structural adaptability to the target structure as revealed by simulation in terms of strand bending force and strand rotational freedom.

The following second study on rigid and semi-rigid nucleic acid scaffolds will now explicitly consider the biological receptor molecule as well.

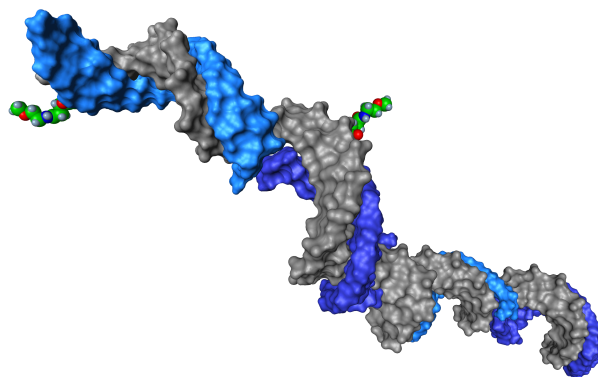


FIGURE 5.6: Surface representations of the conformation of complex **22** obtained by MD simulations in which a constant force acted on the strand to bend the complex into a conformation suitable for bivalent binding to ECL. In this case, the short distance between the ligand attachment sites leads to a linear rather than a bent conformation (dark gray = DNA, blue = PNA strands, green = linker).

5.3 Study R2: DNA-controlled bivalent presentation of ligands for the estrogen receptor

This section is based on (and contains content from) the following publication:

- F. Abendroth, A. Bujotzek, M. Shan, R. Haag, M. Weber, O. Seitz: DNA-controlled bivalent presentation of ligands for the estrogen receptor. *Angew. Chem. Int. Ed.*, 50(37):8592–8596, 2011. URL <http://onlinelibrary.wiley.com/doi/10.1002/anie.201101655/abstract>.

Study R1 has proven that the structural properties of nucleic acids are advantageous for the controlled presentation of ligands in multivalent arrays. The self-organization of complementary strands allows for generating a high number of different (and yet controlled and monodisperse) structural permutations with a minimum of synthetic effort. The rigid character of the double helix permits to position the attached ligand moieties very precisely, and thus turns nucleic acids into molecular rulers for screening the distance of binding sites of yet unknown protein targets. One of the main advantages of nucleic acid scaffolds is the possibility to introduce structural flexibility in small, controlled “doses” wherever needed, either by adding backbone nick sites or unpaired single strand segments of varying length. Again, this can be implemented by changes in the template strand, and thus rather conveniently. The effects of these flexible substructures on the parameters of ligand presentation have been the focus of the computational part of study R1.

The second study on rigid and semi-rigid spacers and scaffolds (and the final part of this thesis dealing with spacer and scaffold design) revisits the model system of the previous

chapter, estrogen receptor (ER). In this study, the favorable properties of nucleic acid scaffolds – in this case DNA – are exploited in order to create potent bivalent ligands for ER, outperforming not only a series of bivalent ligands using flexible spacers (cp. Chapter 4.3), but also the monovalent control. In this respect, the study is unique as it shows that multivalent enhancement can also work for individually potent ligand moieties with nanomolar affinities presented at low valencies – and for a relevant biological target. The computational part of this study focuses on the modeling of receptor binding modes of different DNA-SERM complexes, with the aim to facilitate the interpretation of the relative binding affinity (RBA) measurements. This once more involves the elusive hydrophobic subsite on the surface of ER α (cp. Chapter 4.1.2)

5.3.1 Rigid and semi-rigid DNA scaffolds for the bivalent presentation of estrogen receptor ligands

The precise positioning of functional groups conjugated to DNA by mutual recognition of DNA sequences has been demonstrated in a number of different scenarios, e.g. for chromophores [189, 190], metals [191, 192], catalytic units [193, 194], nanoparticles [195, 196], fluorophores [197, 198], and even proteins [199–201]: The ability to arrange functional units at well-defined distances is required for many applications, and DNA hybridization (without neglecting the use of other nucleic acids) is a convenient means to do so. While until recently the focus of these applications has been on material science as well as on the immobilization of biomolecules, using self-organizing DNA conjugates for addressing biological problems promises to be a fruitful venture [202–204]. In the previous section, Study R1 has presented the rapid spatial screening of lectin binding sites by rigid and semi-rigid multivalent PNA-DNA-LacNAc complexes with minor synthetic effort. Other examples include the interrogation of a tandem SH2 domain with DNA-peptide conjugates [18], or of a dimeric death receptor with PNA-peptide macrocycle conjugates [205]. The common motif is the covalent attachment of a ligand for a biological receptor to an oligonucleotide. Bi- or even multivalent nucleic acid-ligand conjugates can be generated in a very straightforward manner by hybridization of more than one modified oligonucleotide to an (unmodified) template strand, which, by variations in sequence and length, can be used to adjust the distances at which the ligands are presented to the receptor. The concept has been discussed at length in Sections 5.2.1 and 5.2.2. The results presented so far have already indicated that nucleic acid scaffolds offer notably more control w.r.t. to the parameters of ligand presentation than can be implemented by these use of flexible spacers. The study presented here marks the first example where DNA scaffolds are used for the spatial screening of a protein receptor using potent small molecule ligands. The model system is again the dimeric ER α , and the

study incorporates the two well-known SERMs 4-hydroxytamoxifen (Tam) and raloxifene (Ral), but only the latter ligand is incorporated into the computational part of the study. A series of bivalent ligands based on Ral (including the monovalent control) from Study F2 is integrated into Study R2 for comparison.

The synthesis of the SERM-oligonucleotides was realized by introducing an alkyne modified uridine building block during the automated DNA synthesis. The SERMs were equipped with azido functions to enable the covalent attachment to the oligonucleotides by Cu-catalyzed 1,3-dipolar cycloaddition. The resulting conjugates were obtained in 30–70 % yield. The affinities of the complexes and compounds were evaluated with the HitHunter enzyme fragment complementation (EFC) estrogen chemiluminescence assay (DiscoverRx) [141]. All binding affinities are expressed as RBA values (Figure 5.7), that is, relative to the IC_{50} values of the reference, E_2 (3.11 nM, RBA = 100%).⁷

		RBA / %			
hexestrol		300			
4-hydroxytamoxifene		145 ^[a]			
raloxifene		30			
	R	R			
5'-ACCAGGGCGCAGAXG		TXCTCCATGGTGGCC-3'			
3'-TGGTCCCGCTCTAC- Y_n -		AAGAGGTACCACCGG-5'			
Y_n	n ; distance in nt	R = Tam, Ral			
/	0; 3 nt	240	120		
C	1; 4 nt	110	9		
AC	2; 5 nt	180	< 1		
AAC	3; 6 nt	210	80		
GAAC	4; 7 nt	300	7		
TGGAAC	6; 9 nt	120	40		
GATGGAAC	8; 11 nt	200	20		
TGGATGGAAC	10; 13 nt	110	50		

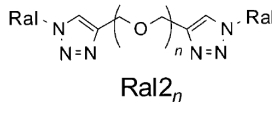
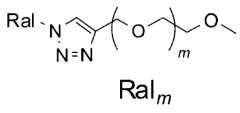
		RBA / %	
	$4R_n$		
		n	
		1	70
		7	10
		10	5
		13	5
		m	
		3	10
		5	3

FIGURE 5.7: Relative binding affinity (RBA) for the relevant complexes and compounds (RBA for E_2 = 100 %, $IC_{50}(E_2)$ = 3.11 mM for $ER\alpha$). [a] Calculated for 1:1 mixture of the *cis/trans* isomers. nt = nucleotide. Note that the flexible bivalent ligands **Ral2_n** appear in study F1 as RAL2EG_n. The same holds for the monovalent control: **Ral₃** = RALEG4Me, **Ral₅** = RALEG6Me (cp. Table 4.4).⁸ Figure taken from [17], reprinted with kind permission of John Wiley and Sons.

Tam and Ral maintained high $ER\alpha$ -binding affinity after conjugation to DNA. Interestingly, Ral, when conjugated to a ssDNA (single strand DNA) 15mer in **2Ral** (not shown) even surpassed the RBA of its small molecule counterpart, Ral (130 vs. 30 % RBA). Given that ER is a transcription factor, it is plausible that the oligonucleotide scaffold contributes to the affinity of the conjugate to ER. A variation of the nucleic acid sequence in **3Ral** (not shown), in turn led to a distinct drop in binding affinity (4 % RBA). Despite the undeniable contribution of the DNA scaffold to the binding affinity of the conjugates, control experiments with unmodified oligonucleotides showed – in the limits of the binding assay – no binding to $ER\alpha$ (cp. [17] Figure S32 and Table S4 in

⁷Details regarding synthesis and RBA measurements are omitted in this short summary. Please refer to the original Ref. [17] and the associated Supporting Information.

the Supporting Information).

The actual spatial screening of the ER α binding pockets was performed using the self-assembled complexes in which two different Tam-oligonucleotide conjugates or two different Ral-oligonucleotide conjugates were annealed to a template strand. The distance between the ligands was varied by changing the number of unpaired template nucleotides Y_n in the resulting bivalent, ternary complexes **4R_n** (cp. Figure 5.7). The stability of the ternary complexes was evaluated by melting experiments, and proved to be virtually unaffected by the conjugation of the ER ligands (cp. [17] Figure S18–S20 in the Supporting Information). The experiments with the bivalent complexes revealed a remarkably high level of up to 300 % RBA for ER α , a noteworthy result given that studies involving flexibly linked SERM dimers have shown relatively low binding affinities (cp. Ref. [135] and Study F2 in Chapter 4.3).

In order to evaluate the effect of bivalent ligand presentation, the complexes **4_n** were compared with a monovalent control, i.e. complexes **5_n** (not shown), compromising the same DNA architecture, but featuring only one ligand moiety. The bivalent complexes **4Tam₀** and **4Tam₄** showed five to seven times higher affinity to ER α than the monovalent complexes **5Tam₀** and **5Tam₄**, which clearly underlines the advantage of bivalent ligand presentation.

The highest binding affinities were obtained for complexes in which the Tam moieties were separated by three (**4Tam₀**) or seven (**4Tam₄**) nucleotides. The complexes involving Ral show an even more pronounced distance dependency: Again, two maxima of the binding affinity were observed at a separation of three (**4Ral₀**) and six nucleotides (**4Ral₃**). Analogously to the Tam conjugates, these complexes exceed the binding affinity of monovalent Ral. The two distinct peaks in binding affinity of the bivalent complexes involving both Tam and Ral when presented at three and six, or seven, respectively, nucleotides apart, are noteworthy. In order to estimate the corresponding ligand presentation distances, it was assumed that the complexes adopt the structure of B-DNA, the most common form of DNA in living systems (20 Å helix diameter, 3.4 Å base-to-base distance, 10.2 nucleotides per turn), an assumption that is justified because it was shown that fully base-paired ternary complexes (such as **4Tam₀** and **4Ral₀**) maintain the structural characteristics of B-DNA [206]. Furthermore, in accompanying FRET studies (FRET = fluorescence resonance energy transfer) it was ascertained that a ternary DNA complex consisting of two double-helical segments separated by three unpaired nucleotides retains the length of a canonical B-duplex [18].

Based on this assumption, and taking into account the length of the linker (Figure 5.8) as well as helical torsion, it was concluded that a spacer of three nucleotides length will arrange the triazole units at a distance of less than 23 Å, while a spacer of six or seven nucleotides length will position the triazole units 38–40 Å apart. Furthermore, given that in the latter cases a segment of unpaired nucleotides is involved, a larger degree

of flexibility can be expected. For comparison, the nitrogen atoms of two raloxifene molecules in a co-crystal with the ER α dimer are positioned 34.7 Å apart [128].

In order to investigate the pronounced distance dependency revealed by the RBA measurements in more detail, the structural fit of different bivalent complexes w.r.t. to the ER α dimer was evaluated in terms of a computational study consisting of two parts: The first part aimed at pointing out a possible structural basis for the binding behavior by suggesting a number of complex-receptor binding modes by means of modeling, and subsequent evaluation of the suggested binding modes in MD simulations. The second (and more general) part aimed at characterizing the influence of unpaired single strand segments in ternary DNA complexes on structural properties that may play a role for the presentation of ligands. The results are presented in the following.

5.3.2 Results and discussion (computational part)

Modeling of bivalent steroid binding site modes

The complexes **Ral4**₀, **Ral4**₂, **Ral4**₃ and **Ral4**₄ were generated based on generic DNA double helix structures generated with the software **nab** from AmberTools 1.4 [155] and complemented by unpaired single strand segments as well as the nucleotide-Ral conjugates at the appropriate positions in the sequence with the modeling and visualization software **Amira** [139]. Given the strong evidence for ternary DNA complexes adopting the shape of B-type DNA (cp. 5.3.1), the DNA helices were modeled accordingly. Force field parameters for DNA were readily available in the Amber-99SB force field [149], so that only the novel nucleotide-Ral conjugate (Figure 5.8) had to be parametrized from scratch.

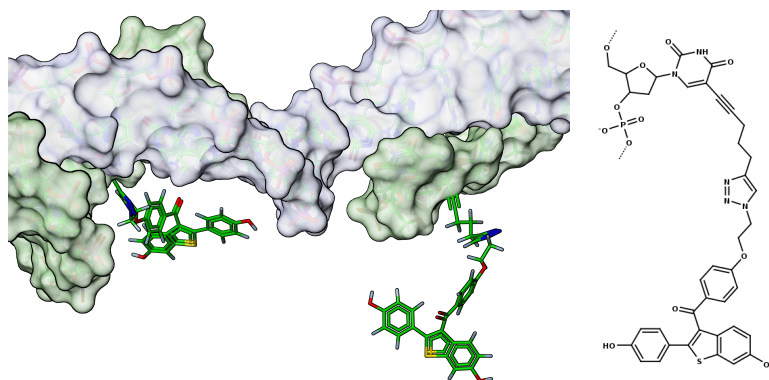
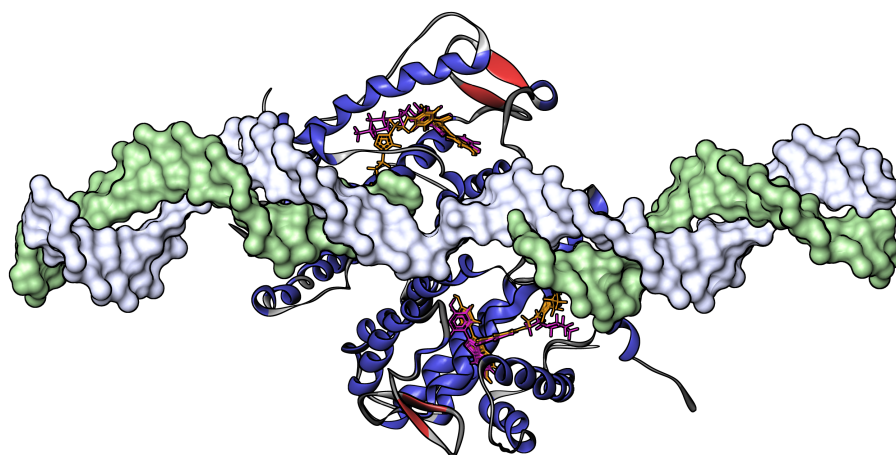


FIGURE 5.8: The nucleotide-Ral conjugate incorporates a semi-flexible linker that ascertains an adequate distance between Ral moieties and DNA duplex (**4Ral**₃).

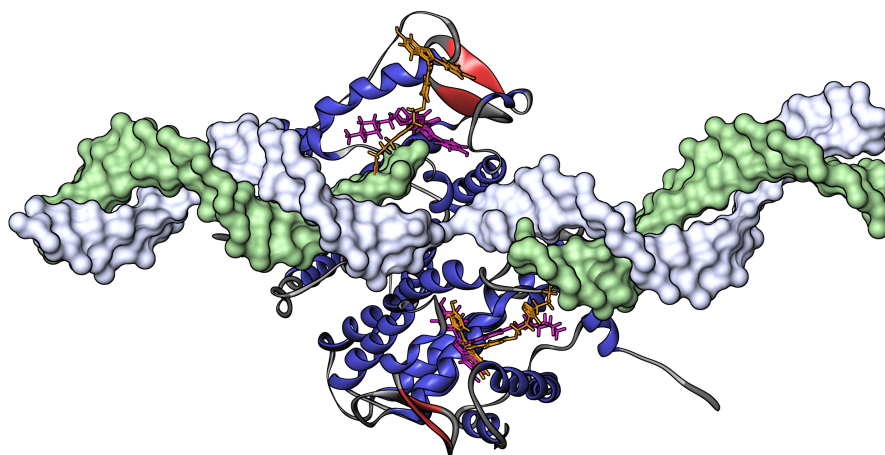
The first objective was to investigate the large difference in RBA between complex **Ral4**₃, with Ral moieties separated by six nucleotides, and complexes **Ral4**₂ and **Ral4**₄, having either one nucleotide less or one nucleotide more separation. With an estimated distance of 38–40 Å between the triazole units (**Ral4**₃), bivalent binding should be in range for all three complexes. In order to obtain ER α binding conformations of the complexes **Ral4**_{*n*} (*n* = 2–4), first an alignment of a single Ral moiety of complex **Ral4**_{*n*} with a single SERM (LY156681) as bound in complex structure with PDB ID 2R6W [136] was performed. By modeling with Amira, the remaining part of the **Ral4**_{*n*} complex was then arranged so as not to cause steric overlap with the protein structure. At the connection between duplex parts and unpaired single strand segments, the connection was first severed in order to allow for arranging the rigid duplex parts more easily, to be reconnected again afterwards. The remaining Ral moiety was arranged within the second binding pocket or close by, if possible. After an initial geometry with no or only moderate receptor overlap was found, a first energy minimization (steepest descent algorithm) was performed, using only the **Ral4**_{*n*} complex in explicit solvent, with position-restrained Ral moieties. The optimized – mainly w.r.t. to the conformation of the flexible segments – **Ral4**_{*n*} complex was then reintroduced to the ER α structure, put into a new solvent box, and a second energy minimization was performed, using the complete system consisting of protein, **Ral4**_{*n*} complex and explicit solvent, without position restraints, followed by a 200 ps position-restrained MD simulation in order to settle the solvent molecules. These simulations provided the starting geometries for unrestrained MD simulations of up to 10 ns lengths which were used for data collection and in order to evaluate the stability of the complexes. The resulting binding modes are shown in Figure 5.9.

The results show that complex **Ral4**₃ is able to achieve bivalent binding to the steroid binding sites of the ER α dimer without notable structural strain (cp. Figure 5.9 a). The flexible spacer of three nucleotides of unpaired single strand has a large contribution in this respect, as it can arrange within the narrow groove between the two ER α monomers more readily than a double-helical segment. The distance between the two triazole units is 28.6 Å (mean over 10 ns of MD), which allows for unhindered entry of the Ral moiety into the steroid binding sites. Again, similar to the binding mode of bivalent ligand **Ral2**₇/RAL2EG7 from Study F2 (cp. Chapter 4.3.3), the triazole units adopt a different positioning than the 4-methylpiperidin in the side chain of ligand LY156681, a fact that can probably be attributed to the attached linkers.

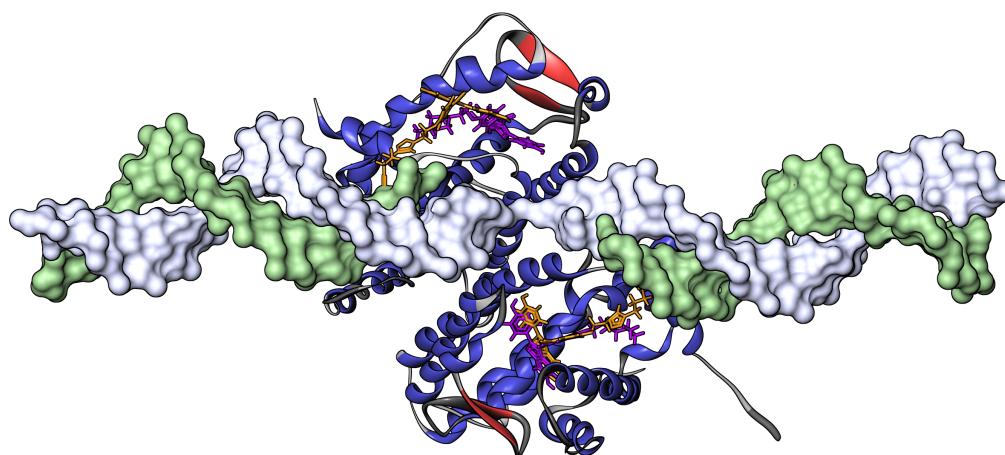
By contrast, the predicted binding modes of complexes **Ral4**₂ and **Ral4**₄ do not achieve simultaneous bivalent binding to the steroid binding sites of the ER α dimer without putting up with significant structural strain. A separation of five nucleotides between the Ral moieties (**Ral4**₂) leads to an only marginally smaller distance w.r.t. to the linear distance between the modified nucleobases in the strand; however, the angle of ligand



(a) **4Ral₃**-ER α complex (6 nt separation, 3 nt unpaired)



(b) **4Ral₂**-ER α complex (5 nt separation, 2 nt unpaired)



(c) **4Ral₄**-ER α complex (7 nt separation, 4 nt unpaired)

FIGURE 5.9: Proposed binding modes of bivalent complexes (a) **Ral₄₃**, (b) **Ral₄₂**, and (c) **Ral₄₄** to ER α (crystal structure with PDB ID 2R6W [136]): Ral moieties = orange, LY156681 ligands (SERM from crystal structure) = purple, template strand = light gray, conjugate strands = green. The pictures show snapshots from MD simulations in explicit water.

presentation is altered such that, given that one Ral moiety is already bound to the first steroid binding site, the second Ral moiety is directed towards the solvent (cp. Figure 5.9 b). This effects seems to be amplified by the rigid linear spacer segment that connects the nucleobase with the triazole unit of the Ral moiety. Consequently, despite the shorter separation in terms of nucleotides, the mean distance between the two triazole units increases notably to a value of 39.46 Å. The structural mismatch appears to be less severe for complex **Ral4₄** (cp. Figure 5.9 c), where the second Ral moiety only falls a little short of the steroid binding site (in fact, contact to the binding site entry can still be established). Again, the rigid linear spacer segment may be a double-edged sword, as it successfully ascertains the unobstructed presentation of the Ral moiety, however at the cost of not allowing for a more flexible compensation of structural mismatch in the scaffold.⁹ In this case, the mean distance between the two triazole units adds up to a value of 35.33 Å, which – considering the co-crystal of ER α with raloxifene – is technically very close to an optimal distance for bivalent binding. This might explain why complex **Tam4₄** (not included in the computational study), analogously to his counterpart having a separation of seven nucleotides between the ligand moieties, was found to have a very high RBA value of 300 %. Here, a small change in the ligand moiety might indeed make a difference.

In summary, the modeling results seem to represent a plausible structural basis for the RBA measurements presented above (cp. Figure 5.7). What remains is the question how **Ral4₀**, the complex with the highest RBA of all bivalent complexes using Ral, might bind to ER α . Following the above protocol, a bivalent binding mode of complex **Ral4₀** involving simultaneous binding to (or near to) the steroid binding sites of the ER α dimer could not be realized. In this case, the ligand moiety separation distance of only three nucleotides and the lack of a flexible unpaired segment render bivalent binding to the steroid binding sites – from the modeling perspective – impossible. Considering the high RBA of **Ral4₀**, the occurrence of a different bivalent binding mode, involving a hydrophobic subsite associated with the coactivator binding site of ER α , might be the explanation. In the spirit of Study F2 (cp. Chapter 4.3), this would presume the existence of an intermolecular bivalent binding mode (involving both steroid binding sites on the opposing sides of ER α dimer, cp. Figure 5.9) and an intramolecular bivalent binding mode (involving both steroid binding site and a hydrophobic subsite on a single ER α monomer). In the next section follows an evaluation of in how far **Ral4₀** might be able to adapt such an intramolecular bivalent binding mode.

⁹If the nucleic acid scaffold is to be used as a molecular linker for screening distances between binding sites, an excess of structural compensation by flexible linkers may be counterproductive.

Modeling of bivalent subsite binding modes

In the computational part of study F2, the ER α binding modes of a series of bivalent Ral ligands tethered by flexible oligo(ethylene glycol) (OEG) spacers, covering spacer variants from EG1 to EG7, were investigated by MD simulations, starting from a configuration where one Ral moiety of the bivalent ligand is bound to the first steroid binding site, while the remaining Ral moiety is dangling in the solvent (cp. Chapter 4.3.3). The result was not only that bivalent binding to the steroid binding sites in all likelihood would require a spacer of at least seven EG units (intermolecular bivalent binding was shown for ligand **Ral2**₇/RAL2EG7), but also that bivalent ligands tethered by shorter spacers adapt several other binding modes, some of which involving contact of the “dangling” Ral moiety to the surface of ER α . This phenomenon might also be the source of the high RBA value of complex **Ral4**₀.

In order to assess this in more detail, it was evaluated where the “dangling” Ral moiety of compound **Ral2**₁/RAL2EG1, the highest affinity bivalent ligand tethered by a short flexible spacer, can find favorable interaction sites on the ER α surface. For this purpose, the free Ral moiety was “docked” at several accessible locations in the vicinity of the steroid binding site, and interaction energies and complex stability were evaluated by conducting short MD simulations. A stable binding mode for the auxiliary Ral moiety that is accessible with the very short spacer of **Ral2**₁/RAL2EG1, given that the first Ral moiety is bound to the steroid binding site, could be identified in terms of a hydrophobic patch (or groove) that is formed by residues Leu354, Leu 355 and Ile358 “below” the steroid binding site entry, and residues Leu539 and Met543 on Helix 12 (Figure 5.10).

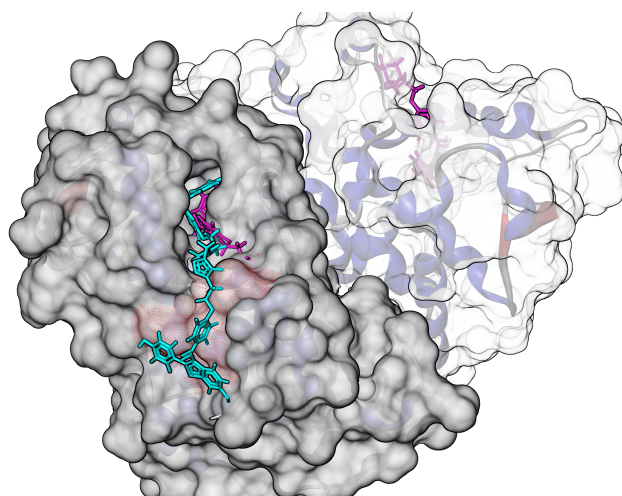


FIGURE 5.10: **Ral2**₁/RAL2EG1 (cyan) adapting an intramolecular binding mode involving both steroid binding site and hydrophobic subsite (red) on the same ER α monomer (gray). For comparison, the LY156681 ligands from the crystal structure (PDB ID 2R6W [136]) are shown in purple. The picture shows a snapshot from an MD simulation in explicit water.

This region belongs to the coactivator recognition interface of ER α (cp. Chapter 4.1.1, in particular Figure 4.1) and has already been shown to bind hydrophobic peptides [207] and – in the case of its homologue ER β – small hydrophobic molecules (such as SERMs) as well (cp. Chapter 4.1.2, in particular Figure 4.2). Thus, considering the high RBA value of **Ral2**₁/RAL2EG1 and the accumulated evidence (including modeling results) it is likely that the suggested intramolecular binding mode as depicted in Figure 5.10 exists. Of course, the availability of the hydrophobic subsite is dependent on the positioning of Helix 12, which in turn is dependent on the ligand that is bound to the steroid binding site. The crystal structure used for modeling the binding mode (PDB ID 2R6W) is a co-crystal with SERM (and raloxifene derivative) LY156681, so that the positioning of Helix 12 should also match complexes involving **Ral2**_{*n*}/RAL2EG_{*n*} and **Ral4**_{*n*}. A very rough estimate of the interactions energies from the MD simulation indicates that binding of a single Ral moiety to the subsite yields approximately 40 % of the enthalpy of the steroid binding site.

Finally, it was evaluated if complex **Ral4**₀, having the shortest ligand presentation distance as well as the highest RBA of all bivalent complexes 4**Ral**_{*n*} can adapt a similar intramolecular binding mode as the flexible compound **Ral2**₁/RAL2EG1. For this purpose, the Ral moieties of **Ral4**₀ were aligned to the Ral moieties in the predicted intramolecular binding mode of **Ral2**₁/RAL2EG1, and the scaffold (a double-helical structure with a very limited number of degrees of freedom) had to arrange accordingly. Again, the complex was optimized over two cycles of minimization in explicit water according to the protocol described above, to be subsequently studied in an MD simulation. The resulting binding mode is shown in Figure 5.11.

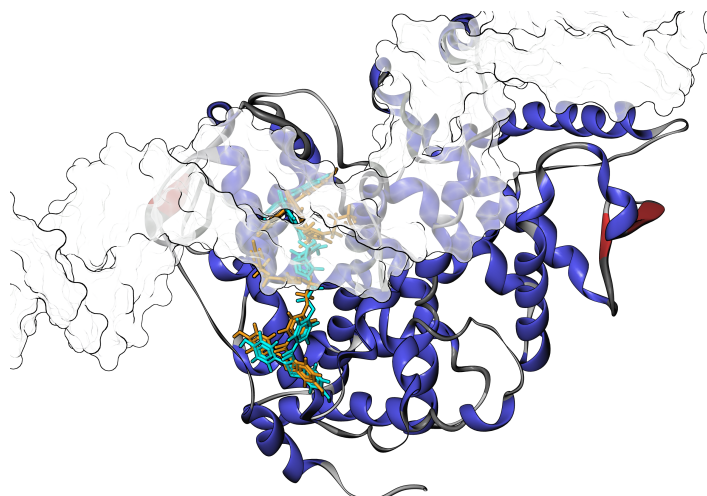


FIGURE 5.11: Overlay of **4Ral**₀-ER α and **Ral2**₁/RAL2EG1-ER α complex adapting an intramolecular bivalent binding mode involving both steroid binding site and hydrophobic subsite: DNA scaffold = transparent white, Ral moieties of **4Ral**₀ = orange, **Ral2**₁/RAL2EG1 = cyan.

The modeling results indicate that **4Ral**₀ can adapt the suggested intramolecular binding mode analogously to its flexible counterpart, **Ral**₂₁/RAL2EG1, without putting structural strain on its (rather rigid) double-helical DNA scaffold. In contrast to the intermolecular binding modes for complexes **4Ral**₂, **4Ral**₃ and **4Ral**₄ (cp. Figure 5.9), the DNA (double) strand does not arrange in the central groove of the ER α dimer, but rather on the side of a single monomer.

Flexibility of ternary DNA complexes containing unpaired nucleotides

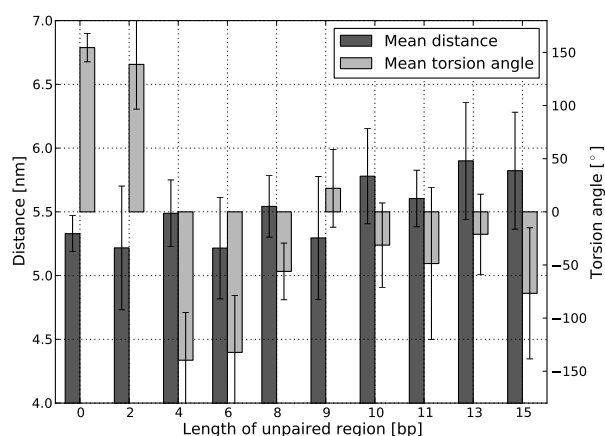


FIGURE 5.12: Mean distance and mean torsion angle between two nucleobases divided by a distance of 15 bp as a function of the number of unpaired nucleotides in between. Figure taken from the Supporting Information of [17], reprinted with kind permission of John Wiley and Sons.

The final part of the computational study changed the focus from modeling ER α binding modes to a more general view on conformations of ternary DNA complexes in water. In order to assess the impact of unpaired regions on the parameters of a DNA double helix structure, multiple 32 ns MD simulations of ternary DNA complexes of 42 bp length, and supplemented by variable segments of unpaired strand (0-15 bp), were conducted. The mean distance as well as the mean torsion angle w.r.t. the longitudinal axis of the strand were measured at two exemplary nucleobases divided by 15 bp distance, with the unpaired segment in between (Figure 5.12). The mean distance (dark gray) tends to increase proportional to the length of the unpaired region, as the DNA single strand helix diameter is smaller than the DNA double helix diameter, i.e. the unpaired segment extends in length. However, due to the increase in bending propensity in the unpaired region, the net mean distance increase is relatively mild (≈ 5 Å). The error bars (standard deviation over 32 ns simulation time) show that even very short unpaired segments introduce notable structural flexibility compared to the continuous double strand (length of unpaired region = 0). The same holds for the torsional freedom

of the strands (light gray, plotted from -180 to 180°): While the full double strand complex allows for only a very modest torsional freedom torsion of approximately 13° , single strand segments enable the strand to rotate up to more than 70° around the longitudinal axis. Although MD simulations of 32 ns length are not able to sample the complete conformational space of long DNA single strands, certain trends are clearly visible and may be helpful when using DNA as a variable scaffold for the presentation of ligands. The results are also in line with the findings that were presented for semi-rigid PNA-DNA complexes (cp. 5.2.3).

5.4 Conclusion

In summary, the results obtained from the molecular simulations were able to contribute to the interpretation of the experimental results, and thus help to draw certain conclusions regarding the design of spacers and scaffolds for the multivalent presentation of ligands to biological receptors.

First of all, nucleic acids scaffolds exceed flexible spacers in terms of controlled ligand presentation. Double-helical segments reliably retain a rod-like (linear) rigid structure that is successful in bridging distances of several nanometers while allowing for positioning ligand moieties with an error of only a few Ångström with a very manageable synthetic effort. Given that for a series of ternary complexes only two oligonucleotide-ligand conjugates have to be synthesized (while the remaining “work” is taken over by self-assembly), nucleic acids are near perfect molecular rulers for screening the distances between receptor binding sites. More complex architectures, such as cross- or grid-like scaffolds based on the Holliday junction [208], can be implemented as well. The DNA origami approach has impressively demonstrated how even functional three-dimensional structures, such as the DNA lockbox [209], can be constructed.

However, addressing multiple binding sites on a single receptor of low valency is a challenge of its own. In order to achieve high binding affinities, the preorganization of the ligand moieties has to go further than adjusting the presentation distance on a linear spacer. Study R1, presenting a distance screening approach for (low-affinity) binding sites on a lectin, has demonstrated how the introduction of flexible elements to nucleic acid scaffolds can improve the affinity for the receptor by adapting the scaffold to its convex shape. In this context, another advantage of nucleic acid scaffolds is that structural flexibility can be introduced in controlled quantities by the use of either nick sites in the back bone, or unpaired single strand segments of variable length. The structural impact of these flexible elements has been investigated at length in the course of this chapter. Finally, Study R2 has provided evidence that well-preorganized bivalent ligands based on DNA scaffolds are able to outperform both potent monovalent ligands – as well as

bivalent ligands using flexible spacers – in terms of binding affinity for a biological receptor. Of course, one has to tread carefully when making such claims, as, for instance, the DNA scaffold itself can contribute to an increase in affinity, not by preorganization, but in terms of enthalpy, especially when the target receptor is a transcription factor such as ER.

An important factor to consider when using nucleic acid scaffolds is that the distance of ligand presentation is correlated to the angle of ligand presentation, regardless of whether the ligand is attached to the backbone (e.g. a PNA backbone, as in Study R1) or to a modified nucleobase (as in Study R2). The binding mode comparisons in Section 5.3.2 reveal a very distinct difference in the angle of ligand presentation that is brought about by the distance of a only a single nucleotide. Thus, ligands can only be presented in the same angle when a helix turn is completed, or the rigidity of the structure is partly abandoned (at the cost of a higher variance in the ligand presentation parameters). Consequently, if one aims at a high degree of preorganization, considering a three-dimensional structure of the scaffold during the design process is advisable.

One of the main disadvantages of using nucleic acid scaffolds for (intracellular) biological applications is that they, being relatively large and negatively charged compounds, are not well internalized by cells, and, when within the cell, may trigger a response by the immune system, and/or quickly become subject of enzymatic degradation.¹⁰ Thus, scaffolds based on nucleic acids are facing similar adversities as synthetic siRNA [210]. Using more robust, synthetic constructs such as PNA is a possible way to circumvent these problems in the future.

By contrast, bivalent ligands tethered by flexible OEG spacers are readily internalized by cells and were also shown to possess biological activity (cp. Chapter 4.3.2). This advantage is lessened by the problems associated with flexible spacers (discussed at length in the previous chapter), namely folding, wrapping, as well as hydrophobic stacking interactions of ligand moieties in water, leading to a hydrophobic collapse of the structure and thus rendering bivalent binding unfavorable. All of this can be considered side effects of the comparatively low degree of preorganization that is associated with the use of very flexible structures. In this respect, the rigid and semi-rigid scaffolds observed in this chapter have been proven to be more successful: Short-range interactions between two ligand moieties are virtually impossible, and interactions between ligand moiety and spacer/scaffold could not be observed in the course of the simulation. However, the latter case may still occur if rather flexible ligands (e.g. peptide chains), possibly carrying positive charges, are attached to a scaffold that contains a negatively charged DNA backbone. The ligand-scaffold combination presented in Study R2, consisting of a

¹⁰Of course, there is a number of biological applications that involve interactions on the cell surface, or the glycocalyx, that are calling for the multivalent display of ligands, too, but would not be affected by these problems.

negatively charged DNA scaffold that presents hydrophobic ligands attached via short and partially rigid linkers, is not affected by such effects and thus ascertains an unhindered presentation of the ligand moieties.

The conformational entropy loss upon binding has not been studied explicitly for the nucleic acid scaffolds presented in this chapter. For double-helical segments, the entropy loss upon binding should be negligible, while for longer single strand segments, it is very hard to obtain a sampling that would be sufficient for conducting a meaningful conformational entropy estimate. In this respect, the use of experimental methods such as NMR spectroscopy that are able to take into account longer timescales is probably more fruitful (cp. Chapter ??). Still, it appears that, again, conformational entropy loss is not the limiting factor w.r.t to binding affinity, but rather enthalpy, in terms of sterical fit to the target. Complexes **18**, **24** and **25** (the latter complex was not modeled) from Study R1 provide clear evidence that partially flexible scaffolds are able to outperform their more rigid counterparts in terms of binding affinity, despite the fact that they have to suffer a larger loss in conformational entropy. This is in accord with the observations from Study F2, where the conformational entropy loss of flexible OEG chains upon binding of the bivalent ligand to ER α constitutes a comparatively small fraction of the estimated change in free energy (cp. Chapter 4.3.3). Consequently, conformational entropy loss appears to be the limiting factor only in the (hypothetical) limit of very long spacers or very flexible scaffolds for addressing more than a single receptor molecule. In summary, one could conclude that sterical considerations and correct pre-organization of the ligand moieties should be the main priority of spacer and scaffold design, while reduction of conformational entropy loss can be considered a means for further optimization of established binders.

5.5 Experimental setup

5.5.1 Study R1

Structures of the complexes were modeled using the visualization software *Amira* [139], based on the solution structure of a PNA-DNA duplex obtained from the Protein Data Bank (PDB ID 1PDT) [186] and a generic B-DNA double helix structure generated with the software *nab* from AmberTools 1.2 [155]. The modified PNA residue was modeled including the complete linker, but without the LacNAc ligand. The standard base set of PNA required for PNA-DNA duplexes and the modified PNA residue were parametrized for the Amber-99SB force field [149] using the software *Antechamber* [153, 154], with charges calculated by the AM1-BCC method [156, 157]. The protonation state of the PNA-DNA complexes in solvent was the normal one for pH 7. All simulations were

performed using the software GROMACS, version 4.07 [76] and the according port of the Amber force fields, ffamber [148]. For the 30 ns simulations, the linear complex conformations were used as starting configurations. The complex structures were solvated with roughly 176,000 water molecules (TIP3P water model [152]) in equally sized rhombic dodecahedron boxes of 19.6 nm side length. To neutralize the overall charge, an adequate amount of potassium ions was added to the simulation boxes. The energy of the systems was minimized with the steepest descent algorithm, and 200 ps position restrained MD simulations were performed in order to settle the solvent molecules. These systems were subsequently used as starting conformations for the unrestrained long-time simulations. In order to obtain bent complex conformations, a constant distance restraining potential was used for bending the complex structures such that the distance between the linkers was adjusted to the approximate distance of the ECL binding sites. This was done using implicitly modeled solvent. Afterwards the bent complexes were transferred into explicit TIP3P solvent, following the protocol described above for the linear complexes, including energy minimization and position restrained MD in order to settle the solvent molecules. These were then used as starting configurations for the force measurements. The forces were measured by applying an harmonic potential restraining the distance between atoms of the two linker attachment sites, and then calculating a 400 ps MD trajectory, yielding the mean the force that has to act on the strand to remain in the bent conformation with the correct linker-linker distance. To maintain a constant temperature of 300 K and a pressure of 1 bar, velocity rescaling [86] and Berendsen weak coupling [90] were applied. A twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm [75] was used for Coulomb interactions, with a switching distance of 1.0 nm. For the 30 ns simulations, bond lengths were constrained using the LINCS algorithm [84], allowing for an integration step of 2 fs. For the force measurement simulations, an integration step of 1 fs and unconstrained bond lengths were used.

5.5.2 Study R2

Structures of ligands and ligand-DNA complexes were modeled using the visualization software Amira [139]. DNA strands used in the ligand-DNA complexes were built based on generic B-DNA double helix structures generated with the software nab from AmberTools 1.4 [155]. The structure of ER α was obtained from the Protein Data Bank (PDB ID 2R6W) [136]. All molecules were parametrized for the Amber-99SB force field [149]. Molecule residues not available in the standard force field database, including the nucleotide-raloxifene conjugate, were parametrized using the software Antechamber

[153, 154], with charges calculated by the AM1-BCC method [156, 157]. The protonation state of all molecules in solvent was the normal one for pH 7. All simulations were performed using the software GROMACS, version 4.07 [76] and the according port of the Amber force fields, *ffamber* [148].

Simulations were performed in rhombic dodecahedron boxes with at least 1 nm solute-box distance filled with explicit water (TIP3P water model [152]). To neutralize the overall charge, an adequate amount of potassium ions was added to the simulation boxes. To maintain a constant temperature of 300 K and a pressure of 1 bar, velocity rescaling [86] and Berendsen weak coupling [90] were applied. A twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm [75] was used for Coulomb interactions, with a switching distance of 1.0 nm. Bond lengths were constrained using the LINCS algorithm [84], allowing for an integration step of 2 fs.

Chapter 6

Multivalent binding processes

6.1 Introduction

The previous chapters have focused on outlining certain guidelines regarding the design of spacers and scaffolds for the multivalent presentation of ligands. For this purpose, the structural properties of a number of multivalent compounds were investigated by means of molecular simulation, both in the unbound state (in solvent), and in the bound state (in complex with a biological receptor). This comparative approach is able to reveal a number of issues that can be of relevance, and, in particular when related to and verified by “wet” experimental results, allows to draw certain conclusions. One main drawback of this approach is that the modeling of the bound state is based on a number of assumptions that have to be made based on previous knowledge, in most cases in the form of available crystal structures: Docking algorithms typically need the approximate position of the binding site in the sub-nanometer range in order to produce meaningful results. For the modeling of the bivalent binding modes in the previous chapters, it was presumed that the ligand moieties adapt roughly the same conformation and orientation in the binding pocket as the monovalent ligand. The latter is largely predetermined by the presence of the attached spacer or scaffold as it has to protrude from the binding pocket – a fact that facilitates modeling.¹ In this particularly fortunate scenario, a number of co-crystal structures of estrogen receptor (ER) α with raloxifene and its derivatives were available, so that a very good guess of the bound state could be obtained by adapting the ligand conformation accordingly and then aligning the ligand moieties of the novel bivalent compounds to the ligands bound in the crystal structure. The result were a number of stable complexes that, to the best knowledge of the author, should reflect the character of the *in vitro* binding modes. Still, as one has not witnessed

¹Computational approaches that, lacking previous knowledge, test for all possible orientations of a small molecule ligand in a binding pocket require notable computing power, cp., e.g., Ref. [211].

the dynamical process that leads from the unbound to the bound state, there is no definite answer to the question if the system is able to adapt the complex in the suggested form. In the case of ER, this question is particularly interesting as the binding sites are buried beneath the receptor “surface” and thus require a rather mobile ligand – or some sort of induced fit by the receptor – in order to achieve binding. In summary, the question arises, instead of modeling a hypothetical bound state, why not simulate the proper binding process?

The answer to this question is related to a number of difficulties that arise from the general limitations of classical molecular simulation already hinted at in Chapters 1.4 and 3.2: The dynamics of the molecular system, which on the atomistic level is dominated (and limited in velocity) by friction with the surrounding solvent molecules has to be traversed using integration steps of 1 or 2 femtoseconds, while the most interesting molecular processes are only beginning to take place in the upper nanosecond regime, and above. Due to the rough potential energy landscape inherent to most molecular systems, the dynamics tends to remain within the same almost invariant subset of conformational space (i.e. a local low energy basin of the potential energy landscape) for a long time – long in relation to the length of the integration step – while transitions between different almost invariant subsets (i.e. conformational changes associated with the crossing of a more or less significant energy barrier) are rare events. Consequently, a lot of computing resources are spent on generating highly redundant data. The study of molecular binding processes is affected by this problem more than other disciplines, as it is dependent on sufficient sampling primarily in the transient regions of conformational space, i.e. exactly those that are the least populated regions due to the fact that molecular systems avoid adopting energetically unfavorable states.

Nonetheless, due to the increasing performance of computing clusters, the simulation of binding processes has come within reach of at least some scientists. In 2011, Buch *et al.* reported the complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics (MD) simulations [212]. The authors performed 495 explicit solvent MD simulations of 100 ns each with the ligand, benzamidine, starting in the unbound state, driven only by diffusion, 187 of which produced binding events to the enzyme, trypsin. The cumulative data was used to generate a Markov State Model (MSM) [213] of the binding process that revealed two metastable intermediate steps, one of which could be identified as being the rate-limiting one. This study reveals the huge computational cost connected with the sampling of binding processes even for a small ligand-receptor system as benzamidine-trypsin: Due to the fact that a single simulation does not generate sufficient information in the transition regions, the experiment has to be repeated an ample number of times.

A more efficient alternative to free ligand binding simulations is given by steered MD (SMD), where typically, the ligand is “pulled” out of the receptor binding pocket (or

through a membrane channel) by an artificial force that acts along a predefined reaction coordinate (cp., e.g., Ref. [214]). This approach requires that, unlike in free ligand binding simulations, certain assumptions regarding the binding process have to be made in advance, such as the position of the binding pocket, and the binding mode of the ligand. Furthermore, in order to obtain unbiased statistics, the pulling simulation has to be accompanied by a subsequent series of umbrella sampling simulations, in which, similar to the ZIBgridfree sampling scheme (cp. Chapter 3.2), additional data is generated in several restrained basins along the reaction coordinate [93]. These are then joined together by a thermodynamic reweighting approach such as the weighted histogram analysis method (WHAM) [215].² The pulling/umbrella sampling approach can be used with great success whenever the ligand is rather small, and its movement is more or less unidirectional, such as in the simulation of transport processes through membranes [216]. Furthermore, it is possible to relate the results of pulling simulations with data obtained from atomic force microscopy (AFM) experiments [217].

Unfortunately, for studying multivalent binding processes, force pulling/SMD is a somewhat problematic choice, as it is not immediately clear on which group of atoms the force would have to act (given that there are multiple ligand moieties), nor in which way to define the reaction coordinate(s) – given that the mechanism of multivalent binding may be rather intricate as it contains additional diffusive elements for the yet unbound ligand moieties. Therefore, the following sections present an approach that combines free diffusion ligand binding simulations with the aforementioned ZIBgridfree sampling scheme. The latter has the advantage that, being an umbrella-like approach, it allows for thorough sampling in the transient regions of conformational space, while it remains very flexible w.r.t. to the choice of reaction coordinates (or rather internal coordinates, cp. Chapter 3.2.2). The “presampling” of conformational space needed by ZIBgridfree is obtained by a free diffusion binding MD trajectory, which ascertains that the complete model of the binding process reflects a physically meaningful transition from the unbound to the (doubly) bound state – within the limits of classical molecular simulation. To the best knowledge of the author, the systematic study of a multivalent binding process in explicit solvent by means of atomistic molecular simulation is a first. The study incorporates a comparison to an analogously designed monovalent system.

²A different thermodynamic reweighting strategy that pursues the same goal has been introduced in Chapter 3.2.5.

6.2 How to presample a bivalent host-guest binding process

The model system for the following computational binding study is a bivalent crown ether-ammonium host-guest system, consisting of a anthracene linked 18-crown-6 dimer host molecule (**DiC6**), and a bivalent ammonium ion guest molecule, incorporating a short flexible spacer (**G3+2H**) (Figure 6.1). The compounds (among a number of variations of the guest molecule, mainly in terms of spacer length) were synthesized and studied in ITC experiments by Larissa v. Krbek³ at Freie Universität Berlin in the course of a master's thesis [66]. The thermodynamics of the formation of complex (**G3+2H**)•**DiC6** in the presence of tosylate counter ions (OTs) could be characterized in a mixture of chloroform and methanol. Using the “double mutant cycle” approach [60, 61], the system was shown to exhibit positive chelate cooperativity (cp. Chapter 2.4). Therefore, the author of the study reasoned that the singly bound open 1:1 complex does not persist as a (meta)stable species in the solvent, but rather represents a transient intermediate of the formation of the doubly bound cyclic complex – a hypothesis that ought to be verifiable by means of molecular simulation.

The synthetic host-guest system **DiC6**-(**G3+2H**) is probably one of the smallest bivalent ligand-receptor systems imaginable, and thus offers a welcome opportunity for the theoretical study of multivalent binding processes in general, more so than more complex and less well-controlled biological ligand-receptor systems as have been discussed in the previous chapters. Furthermore, a structurally analogous monovalent control system **C6**-(**MonoG1+H**) is available and thus can be integrated into the study for a direct comparison between monovalent and bivalent scenario. Typically, this feature is more difficult to implement when working with biological receptors.

The first task after the structures had been modeled and parametrized for the Amber-99SB force field was to generate a sequence of states that describe a binding path from the unbound to the (doubly) bound state of the system. Furthermore, the idea was to pursue an efficient approach, so that, even when a whole series of compounds (e.g. with varying spacer lengths in the guest molecules) has to be analyzed, the computational cost remains manageable. Given that both complexes proved to be very stable – when started from a the (singly or doubly) bound position obtained by energy minimization, both bivalent (**G3+2H**)•**DiC6** as well as monovalent (**MonoG1+H**)•**C6** did not dissociate in the course of a 100 ns MD simulations at 298 K – it became obvious that simulating a series of dissociation and (re)association events within a single trajectory would not be

³supervised by Prof. Dr. Christoph Schalley (also concept) at Institut für Chemie und Biochemie, Freie Universität Berlin, Takustraße 3, 14195 Berlin (Germany), co-supervised by Dr. Constantin Czekelius (same affiliation)

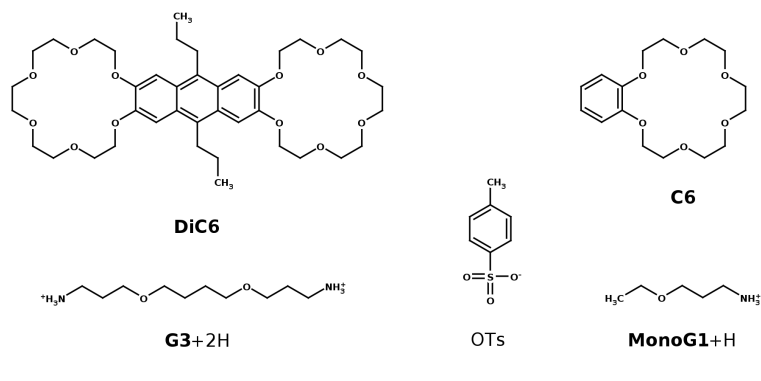


FIGURE 6.1: Structures of the bivalent host-guest system **DiC6-(G3+2H)** (left) and the monovalent control **C6-(MonoG1+H)** (right). The tosylate anion (OTs) was used as counter ion. Nomenclature of compounds according to Ref. [66].

realizable. Therefore, a different approach for the generation of an ensemble of binding paths was developed, which was coined the “vacuum shotgun method”.

The vacuum shotgun method works by repeatedly “shooting” the guest at the host molecule by performing short MD simulations with random starting impulse in the absence of solvent molecules. If the two binding partners do not by chance miss each other (and consequently drift away into empty space for the remaining simulation time), binding in vacuum is in general achieved very quickly, in most cases within the first 25 ps of the simulations.⁴ The advantage of this approach is that a very diverse set of binding paths and binding modes of the system (an ensemble of binding paths) can be generated within a very short time (Figure 6.2).

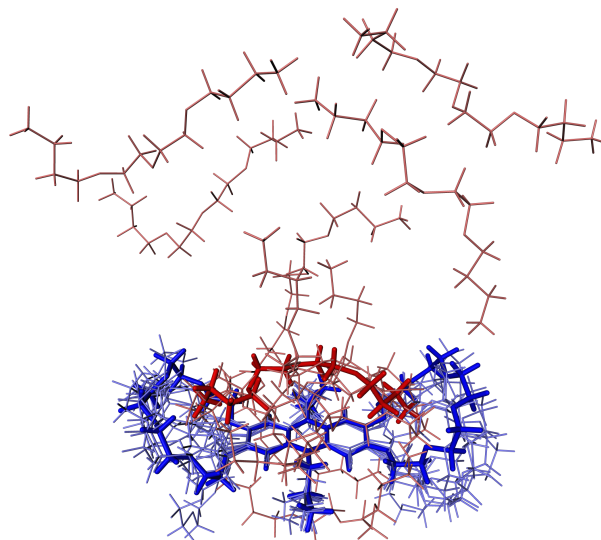


FIGURE 6.2: An exemplary set of states from an ensemble of binding paths for a bivalent host-guest system generated with the vacuum shotgun method.

⁴In this particular case, this corresponds to 4.655 s computing time on a single processor core, more precisely an Intel Core™ i7 870@2.93GHz.

After removal of redundant and continuously bound states, the ensemble of binding paths can be used as presampling input for the ZIBgridfree algorithm. The largest drawback of this approach is that it does not account for the proper positioning of solvent molecules. Therefore, the ZIBgridfree algorithm was extended such that it can generate energetically minimized solvent boxes for the states obtained by the vacuum shotgun presampling using the tools `zgf_solvate_nodes` and `zgf_genion`, which in turn are wrappers for the GROMACS applications `editconf`, `genbox` and `genion`. The retroactive solvation of the input states for ZIBgridfree generates additional computational cost, which, however, is negligibly small in comparison to working with explicit solvent from the beginning.

Unfortunately, when the vacuum shotgun method using retroactive solvation was devised, it was not yet foreseeable that (i) a rather large counter ion, namely OTs, would be involved in the binding process, and that (ii) a polar component in the solvent is needed in order to ascertain the proper solvation of OTs. This poses a number problem w.r.t. to the above approach, as it is not immediately clear if and in how far the counter ion intervenes with the complex formation (and thus where to place it), and how a mixture of solvents would array in the simulation box.

First equilibration simulations of solvent boxes including the final solvent mixture of 10:1 chloroform-methanol as well as the appropriate number of counter ions revealed that not only the OTs molecules in fact tend to array near the ammonium cation moieties of the guest (an indicator that stable tripartite complexes such as $\mathbf{C6} \bullet (\mathbf{MonoG1} + \mathbf{H}) \bullet \mathbf{OTs}$ exist), but also that the methanol component of the solvent mixture forms polar clusters that involve the positively charged guest molecules, the negatively charged counter ions, as well as, to a lesser degree, the host molecules' crown ethers (Figure 6.3).

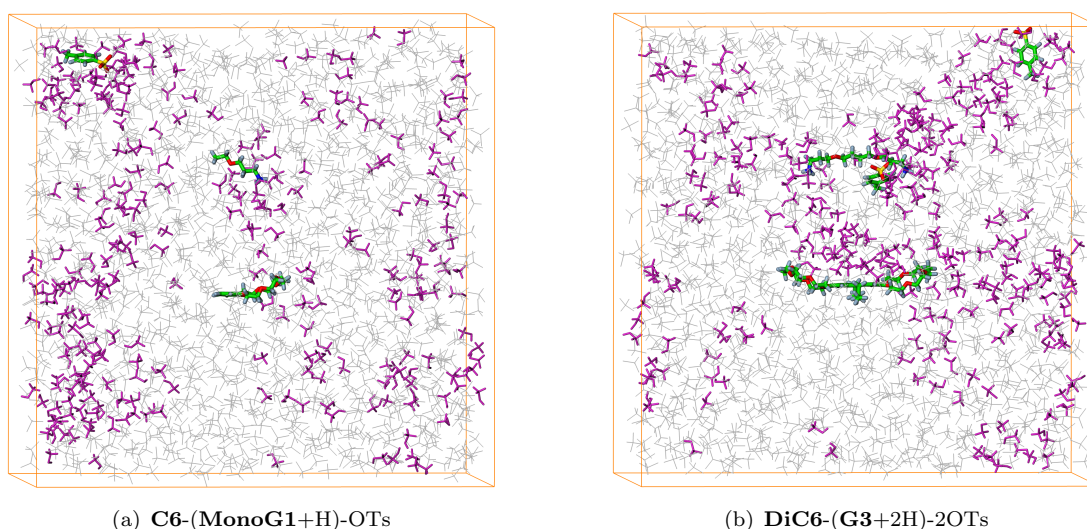


FIGURE 6.3: Host guest systems (a) $\mathbf{C6} \bullet (\mathbf{MonoG1} + \mathbf{H}) \bullet \mathbf{OTs}$ (monovalent) and (b) $\mathbf{DiC6} \bullet (\mathbf{G3} + 2\mathbf{H}) \bullet 2\mathbf{OTs}$ (bivalent) after a 2 ns equilibration of the solvent mixture (10:1 chloroform-methanol) in the position-restrained unbound state at 298 K (chloroform = gray, methanol = purple). Polar clusters of methanol molecules are clearly visible.

Considering the subtle effects connected with solvent mixture and counter ions, and given their expectable influence on the binding process, the vacuum shotgun/retroactive solvation approach was abandoned in favor of a presampling by conventional free diffusion simulations involving the complete explicit solvent and counter ion setup. For this purpose, for each of the two systems **DiC6-(G3+2H)-2OTs** (bivalent) and **C6-(MonoG1+H)-OTs** (monovalent control), five MD simulations of 10 ns length each were conducted, starting from the unbound state with about 1.5 nm separation between host and guest molecule (cp. Figure 6.3). For the monovalent system **C6-(MonoG1+H)-OTs**, one of the five simulations captured a binding event. In case of the bivalent system **DiC6-(G3+2H)-2OTs**, the five simulations produced one binding event leading to a singly bound open complex, and one binding event leading to a doubly bound cyclic complex. In summary, one can estimate that for the given pair of systems, approximately 50 ns of MD are required in order to observe a single binding event. This relatively low yield can be explained by the fact that (i) both host and guest molecule are rather small and mobile and therefore subject to rapid diffusion in the box and (ii) the complexation of host and guest is hindered by OTs counter ions associating with the ammonium moieties and thus obscuring the interaction site. Consequently, not every close contact between host and guest immediately induces complex formation.

As a side note, it shall be mentioned that bivalent guest molecule (**G3+2H**), despite featuring a flexible spacer for bridging the ammonium moieties, did not exhibit wrapped or folded conformations in the given solvent mixture: Quite contrary to the bivalent compounds equipped with hydrophobic ligand moieties presented in Chapter 4, guest (**G3+2H**) is driven towards the extended conformation by the positive charges on both ends as they prohibit closer interactions. Therefore, despite the flexible spacer structure, the preorganization of two interaction sites w.r.t. host **DiC6** is successful.

The obtained binding trajectories represent a physically meaningful path from the unbound to the (singly or doubly) bound state of the system that considers the influence of both solvent mixture and counter ions under the given conditions. These trajectories were used as the starting point for a full conformational analysis with ZIBgridfree that is described in the following.

6.3 Results and discussion

6.3.1 Discretization of conformational space

The (soft) conformational space discretization of both systems was created based on a set of internal coordinates consisting of three strongly correlated distances between ammonium moiety and 18-crown-6 ring per interaction site, leading to three internal

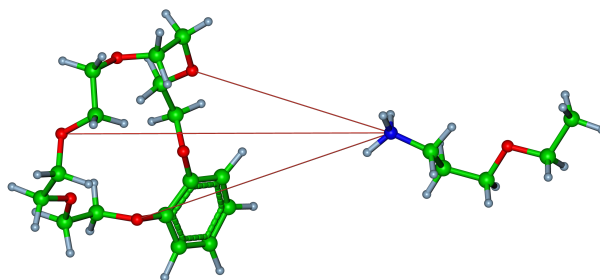


FIGURE 6.4: Three distances between ammonium moiety and binding site form the internal coordinates for system **C6-(MonoG1+H)**-OTs. Analogous internal coordinates were chosen for the bivalent system **DiC6-(G3+2H)**-2OTs.

coordinates for the monovalent system (Figure 6.4), and six internal coordinates for the bivalent system. Given the fact that all distances cover a similar range of values, the internal coordinates were weighted equally (cp. Chapter 3.2.2). In order to remove the abundance of unbound states not related to the binding process from the presampling data, states with distances of more than 1.8 nm distance between the interaction sites were discarded. In case of the bivalent system, this condition had to be fulfilled for both interaction sites in order to warrant the removal of a state from the presampling.

A number of 16 discretization nodes in case of **C6-(MonoG1+H)**-OTs and 31 discretization nodes in case of **DiC6-(G3+2H)**-2OTs – equidistantly placed in the conformational space revealed by the presampling – lead to an α value⁵ of approximately 15, a value that, during the validation of ZIBgridfree, had been proven to be a good compromise between restraining the sampling within a certain area while still allowing for a sufficient overlap between the basis functions (cp. Appendix A). The α value was set to exactly 15 for both discretizations in order to ascertain the comparability of the bivalent system and its monovalent control.

Finally, for each discretization node, 5×500 ps of MD in the NVT ensemble were simulated at a temperature of 298 K, with each 500 ps run starting at the initial position of the discretization node using a random starting impulse vector in order to allow for a good coverage of conformational space even in the transient regions of the binding process (for simulation details, see Section 6.5). This led to a joint sampling time of 40 ns for system **C6-(MonoG1+H)**-OTs, and 77.5 ns for system **DiC6-(G3+2H)**-2OTs.

6.3.2 Metastability analysis

The thermodynamic weights of the partial densities associated with the discretization nodes were calculated via the direct free energy reweighting approach and corrected

⁵The shape parameter α defines the degree of separation that is conveyed by the basis functions ϕ_i , $i = 1, \dots, s$, where s is the number of basis functions that form the discretization (cp. Chapter 3.2.1).

via the overlap integral matrix S as described in Chapter 3.2.5. The energy terms considered for thermodynamic reweighting were (i) the bonded energy terms of host and guest molecule (calculated by using the tool `zgf_rerun`) and (ii) the non-bonded energy terms between host and guest molecule (including self and long-range interactions) from the original simulation. Considering all energy terms from an explicit solvent simulation for thermodynamic reweighting is typically not an option, as the energetic “noise” of the solvent molecules will obscure the state of the solute, and thus leads to an uniform distribution. Therefore, meaningful energy groups for host (receptor), guest (ligand), and other relevant constituents of the system have to be defined prior to the simulation. The results of the thermodynamic reweighting are shown in Figures 6.5 (system **C6-(MonoG1+H)**-OTs) and 6.6 (system **DiC6-(G3+2H)**-2OTs).

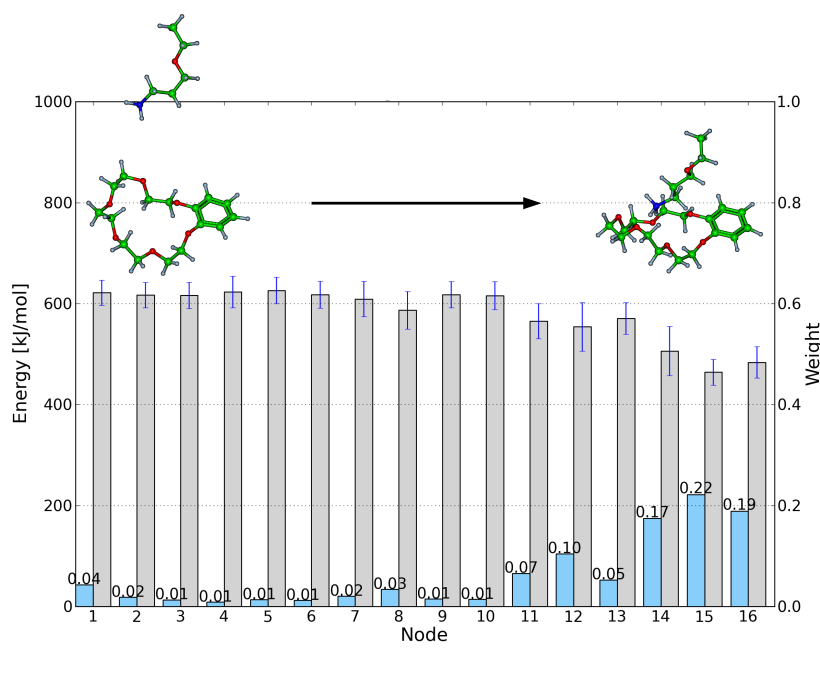


FIGURE 6.5: Mean potential energy (gray) and corrected discretization node weights (blue) for the 16 discretization nodes of system **C6-(MonoG1+H)**-OTs. The improvement in potential energy with decreasing distance between host and guest molecule is clearly visible. The decrease in potential energy goes along with an increase in the thermodynamic weights.

For the monovalent system (cp. Figure 6.5), the weighting documents a decrease in potential energy that is directly related to the distance of the host to the guest molecule (nodes 1–16 cover host-guest distances from approximately 1.7 nm to 2.5 Å). A notable increase in the interaction energy sets in with node 11 at an approximate host-guest distance of 7.5 Å, and culminates in the bound state (nodes 14, 15 and 16). While nodes 1–10 have similar (and low) thermodynamic weights, nodes 11–16, covering host-guest distances of 7.5 Å and nearer represent the largest share of the distribution. This finding also indicates that, due to shielding of the electrostatic interactions by solvent

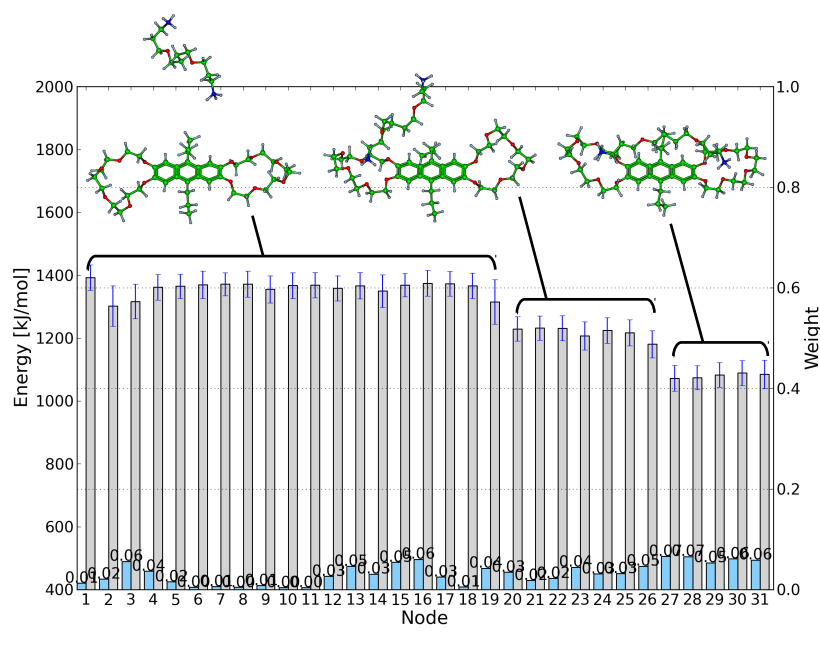


FIGURE 6.6: Mean potential energy (gray) and corrected discretization node weights (blue) for the 31 discretization nodes of system **DiC6-(G3+2H)-2OTs**. Three distinct levels of potential energy for non-complexed, singly bound open complex and doubly bound cyclic complex can be discriminated. The weights of the discretization nodes are distributed more evenly than in the monovalent case.

molecules, a relatively close vicinity between host and guest is required in order to induce complex formation – a completely different picture than for the *in vacuo* simulations reported in Section 6.2.

In contrast to the “continuous” picture that is found for the monovalent system, the bivalent system exhibits three distinct and evenly separated energy levels (cp. Figure 6.6): Nodes 1–19 represent the unbound state, nodes 20–26 correspond to the singly bound open complex, and nodes 27–31 form the doubly bound cyclic complex. In this case, the thermodynamic weights are distributed more evenly over all three energy levels, but nonetheless the largest accumulation of weight is found for nodes that are involved in the doubly bound state.

The overlap integral matrices S (or rather the corrected matrices $D^{-1}S$, cp. Chapter 3.2.5) for the two systems quantify, for each pair of basis functions involved in the discretization, the overlap of the associated partial densities, and thus can be interpreted as describing the transition behavior of the underlying system (Figure 6.7). Matrix rows with many off-diagonal entries indicate transient regions that involve notable overlap between the discretization nodes, while very stable or isolated discretization nodes are marked by large entries on the diagonal. Due to the fact that in this case the discretization is “presorted” by host-guest distances, a block-diagonal structure (an indicator for

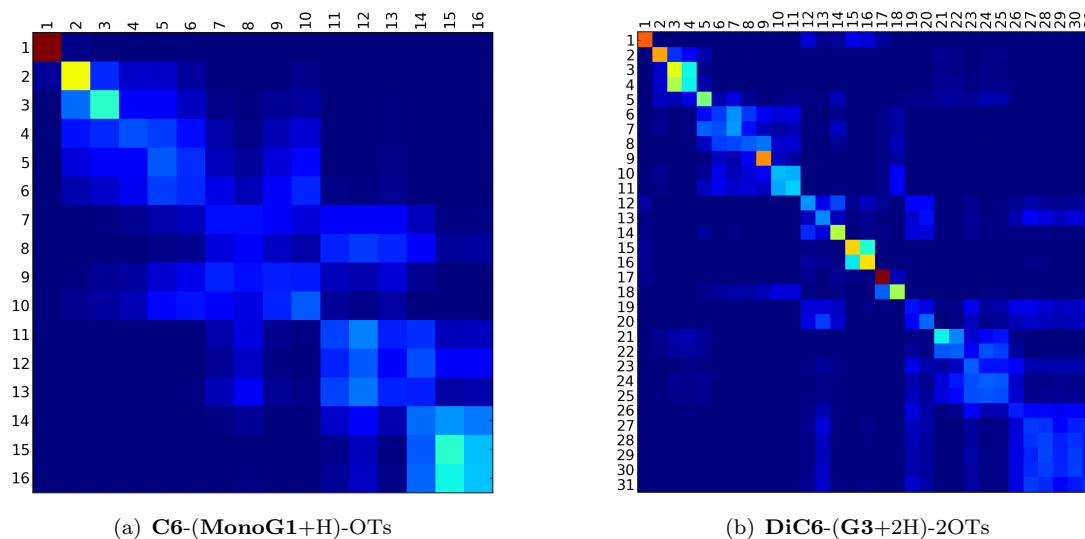


FIGURE 6.7: Overlap integral matrix S for (a) **C6-(MonoG1+H)**-OTs with 16 discretization nodes and (b) **DiC6-(G3+2H)**-2OTs with 31 discretization nodes. Large matrix entries (red, yellow) indicate no or few overlap with neighboring discretization nodes and represent isolated and/or stable regions. Discretization nodes with many off-diagonal entries (blueish) exhibit a significant overlap with their neighborhood and thus mark transient regions.

the presence of metastable states) can be found for both matrices even before permutation by the Robust Perron cluster analysis (PCCA+) algorithm. The S matrix for the monovalent system (cp. Figure 6.7 a) exhibits an isolated unbound state (node 1), an articulate “block” for the bound state (nodes 14, 15 and 16 in the lower right corner) and a large transition region in between. For the bivalent system, the picture is more diverse (cp. Figure 6.7 b): The diagonal contains approximately three small and isolated blocks of nodes, most likely representing different unbound states. The lower right corner (nodes 20–31) contains two relatively distinct blocks with significant mutual communication that should mark the singly bound and the doubly bound state.

By applying the PCCA+ algorithm on the S matrix (cp. Chapter 3.2.5), the number of metastable states n_C as well the degree of membership of the s basis functions (i.e. discretization nodes) with regard to each of the n_C conformations is revealed and formulated in terms of the matrix $\chi \in \mathbb{R}^{s \times n_C}$. Due to the fact that PCCA+ is a soft clustering approach, each discretization node can be assigned to multiple metastable states simultaneously, with varying degrees of membership. This form of eigenvector-based cluster analysis is based exclusively on the transition behavior that is embedded within the S matrix, and requires no predefinition of certain states or “core sets” (e.g. based on chemical previous knowledge). Once the χ matrix has been obtained, it is possible to compute the thermodynamic weights of the metastable states \tilde{w} as $\tilde{w} = \chi^\top w$, with w being the thermodynamic weights of the basis functions. The clustering results for the monovalent

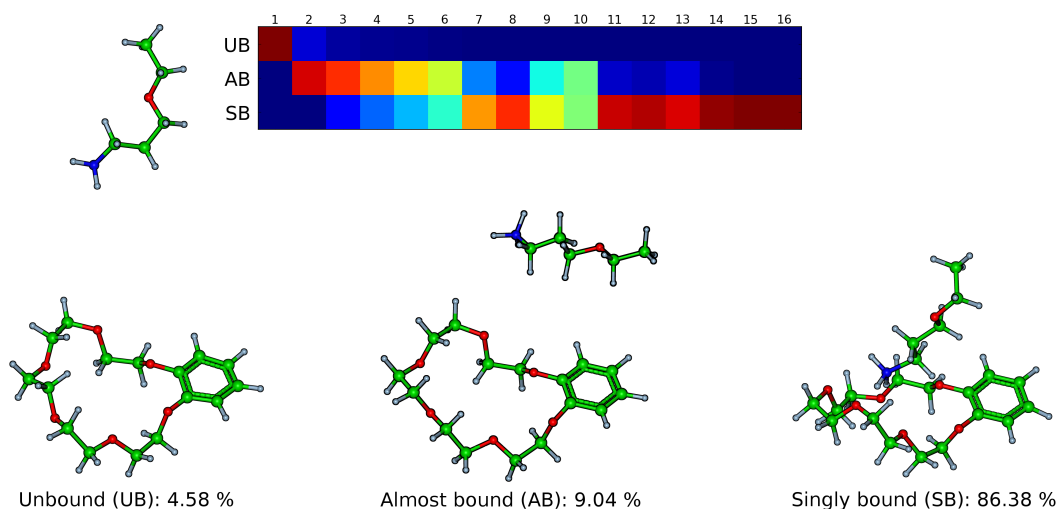


FIGURE 6.8: The χ^T matrix (top) of system **C6-(MonoG1+H)-OTs** groups the 16 discretization nodes into three metastable states: The unbound state (UB, left), the almost bound state (AB, center), and the singly bound state (SB, right). The colors in the χ^T matrix indicate the degree of membership of a discretization node to a given metastable state: dark red = highest degree of membership, dark blue = no membership. Nodes 6, 9 and 10 represent transition regions that belong almost evenly to the two metastable states AB and SB.

system, including a graphical representation of the χ matrix as well as representative molecular states, are shown in Figure 6.8: For system **C6-MonoG1+H)-OTs**, PCCA+ identifies three metastable states, namely the unbound state (UB) with a weight of 4.58 %, the almost bound state (AB) with a weight of 9.04 %, and the singly bound state (SB)⁶ with a weight of 86.38 %. State UB is detached from the rest of the system, save a small degree of communication involving nodes 2 and 3 that leads into state AB. State AB, in turn, exhibits a fluent transition into the SB state. Nodes 6, 9, and in particular 10 mark the transition “state” (or rather transition region in conformational space) between the two clusters AB and SB. Nodes 14, 15 and 16 have the highest membership w.r.t. to state SB, represent the proper bound state.

For the bivalent system, the interpretation of the clustering results is slightly more difficult (Figure 6.9): PCCA+ identifies five metastable states, three of which represent different unbound states, namely UB1 (weight 3.92 %), UB2 (weight 3.49 %), and UB3 (weight 14.46 %), leading to a total weight of 21.87 % for the unbound state. UB1, UB2 and UB3 differ in the relative positioning between host and guest molecule, but share the main characteristic, which is a relatively large distance between both interaction sites on host and guest (and thus beyond being able to induce the binding process). The remaining two metastable states are the almost bound state (AB) with a weight of 15.06

⁶Of course, in this case there is no other bound state than the singly bound one, as the system is monovalent. The nomenclature was chosen for better comparability with the bivalent system **DiC6-(G3+2H)-2OTs**.

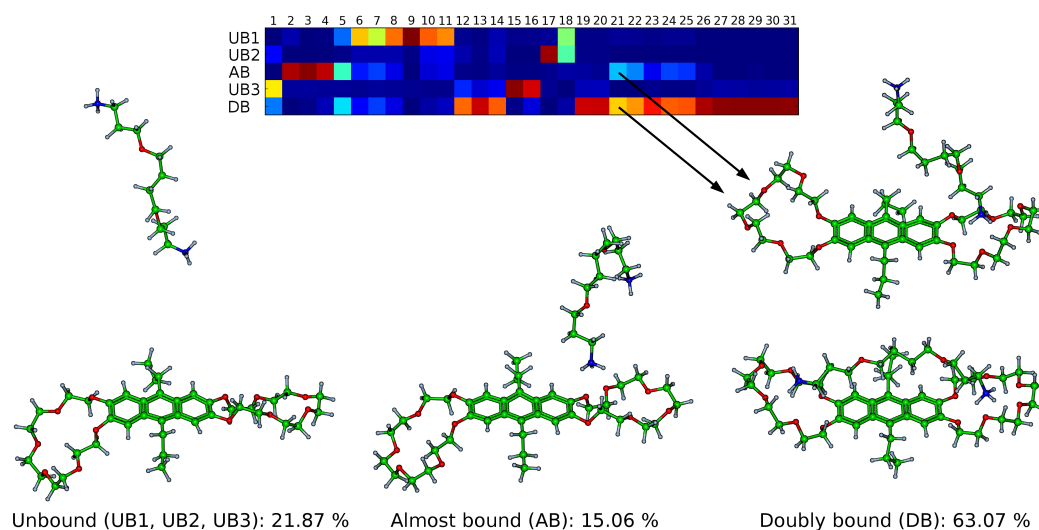


FIGURE 6.9: The χ^T matrix (top) of system **DiC6-(G3+2H)-2OTs** groups the 31 discretization nodes into five metastable states: Three distinct unbound states (UB1, UB2, and UB3, left), the almost bound state (AB, center), and the doubly bound state (SB, right). The colors in the χ^T matrix indicate the degree of membership of a discretization node to a given metastable state: dark red = highest degree of membership, dark blue = no membership. The singly bound state is not identified as a proper metastable entity, but rather assigned to the metastable states AB and DB.

%, and the doubly bound state (DB) with a weight of 63.07 %. Interestingly, based on the eigenvector-based clustering of the S matrix, the singly bound is not identified as a proper metastable state. Instead, it is assigned to state DB, and, to a much lesser degree, to state AB. Given that the separated energy levels for the singly bound and the doubly bound state (cp. Figure 6.6) indicated a seemingly stepwise process, this result is surprising: It means that the singly bound state does not exist as a proper kinetic entity, but rather as the transient component of the doubly bound state. This would indicate a distinct cooperative effect in the bivalent binding process.

In comparison, the bivalent system has a larger fraction of unbound state (21.87 vs. 4.58 %), and a smaller fraction of bound state (63.07 vs. 86.38 %), while the fraction of almost bound state is in the same range (15.06 vs. 9.04 %). This can possibly be attributed to the fact that the presampling of the bivalent system is imperfect in the regard that it covers the almost bound state (leading into the transient singly bound state) only for one of the two binding sites of **DiC6**, while a complete ensemble of binding paths would have to consider both cases. One can only speculate that, if a second almost bound state (“AB2”) had been sampled, the conformational weights would shift in a somewhat more favorable direction for the bivalent system. Finally, it should be mentioned that the weights calculated for the states of the kinetic models are valid only for the given simulation box. A direct comparison with experiment would require to include how often a ligand from the bulk solvent encounters a receptor within the distance that presumably

triggers a binding process.

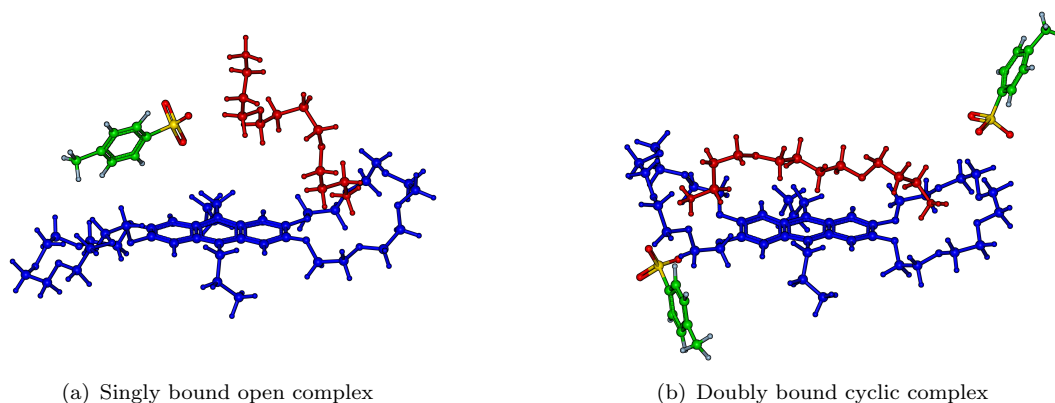


FIGURE 6.10: Snapshots from an MD simulation of system **DiC6-(G3+2H)-2OTs** that show the positioning of the OTs counter ion(s) (a) during the binding process, and (b) after formation of the doubly bound cyclic complex. If in the close vicinity of the guest molecule (red), OTs can act as a competitor for its ammonium interaction sites and thus delays the cyclization of the complex. The host molecule is shown in blue.

Finally, the role of the counter ion OTs throughout complex formation was investigated. It was found that (i) in the bivalent system, OTs counter ions are able to associate with the guest molecule's ammonium moieties and form $(\mathbf{G3+2H})\bullet\text{OTs}$ or $(\mathbf{G3+2H})\bullet\mathbf{2OTs}$ complexes that can persist for several nanoseconds, and (ii) both **C6-(MonoG1+H)-OTs** as well as **DiC6-(G3+2H)-2OTs** were found to form tripartite complexes, the latter both in the singly bound open complex, delaying cyclization (Figure 6.10 a), as well as in the “completed” doubly bound cyclic complex form of the bivalent system (Figure 6.10 b). This confirms the results of the quantum-mechanical DFT calculations presented in Ref. [66], and shows that, if in the vicinity of a guest molecule, OTs prefers the interaction with the ammonium moiety over a closed solvent shell formed by methanol molecules.

Transition probabilities

In order to look into the transition behavior on the level of the metastable states, additional unrestrained short-time MD simulations in the NVE ensemble were conducted. The unrestrained “transition nodes” (as opposed to the discretization nodes used for sampling the stationary distribution) were placed in regions of conformational space that mark interfaces between the different metastable states, and thus are prone to reveal the associated transition behavior more readily than simulations that are started exactly within the center of a metastable region.⁷ A thorough study and evaluation of

⁷The program code for this procedure, including the calculation of matrix $P_c(\tau)$, was devised by Adam Nielsen (same affiliation as the author of this thesis).

different strategies for placing transition nodes, along with a mathematical interpretation of the resulting transition probability matrices, can be found in Ref. [218].

For system **C6-(MonoG1+H)**-OTs, a total of 45 transition nodes was placed, each of which was started for ten runs of 100 ps each using a random starting impulse, leading to a total additional sampling time of 45 ns. Analogously, system **DiC6-(G3+2H)**-2OTs required 95 transition nodes in order to cover the larger number of metastable states. With the same number of simulation runs per node, this lead to an additional sampling time of 95 ns. The resulting transition probability matrices $P_c(\tau)$, with $\tau = 100$ ps, based on the cumulative transition sampling, are shown in the following.

$$P_c(\tau) = \begin{array}{c} \text{UB} \\ \text{AB} \\ \text{SB} \end{array} \begin{pmatrix} \text{UB} & \text{AB} & \text{SB} \\ 0.9868 & 0.0132 & 0. \\ 0.0489 & 0.7584 & 0.1928 \\ 0.0002 & 0.0029 & 0.9969 \end{pmatrix}, \text{ with } \tau = 100 \text{ ps} \quad (6.1)$$

Matrix 6.1 represents the transition probability matrix $P_c(\tau)$ for the monovalent case. Within the short time span of 100 ps, the system has a very high probability to remain in either the unbound state (UB, ≈ 99 %) or the bound state (SB, ≈ 100 %). In case where a transition starting in UB or SB takes place, it almost exclusively happens via the almost bound state (AB). Given the system is in state AB, it is more likely to make the transition into the bound state (≈ 19 %) than into the unbound state (≈ 5 %).

For the bivalent system, $P_c(\tau)$ shows a more diverse picture (Matrix 6.2): The three unbound states UB1, UB2 and UB3 have a somewhat more transient character, with probabilities to remain in the current state of approximately 74, 40, and 80 %, respectively. Surprisingly, all three unbound states are found to allow for a direct transition into the bound state, without having to take a “detour” via the almost bound state, most notably for UB2 and UB3 with a probability of approximately 12 and 15 %. This shows that the soft eigenvector-based clustering does not draw a clear line between unbound and almost bound states. The proper AB state, in turn, has about the same “stability” as in the monovalent case (≈ 75 %), but a slightly higher probability to make the transition into the bound state (≈ 23 % vs. 19 %). Finally, the probability to remain in state DB is about 96 %, and thus a little less than for the bound state in the monovalent system, which is surprising. However, due to the soft definition of states, a direct comparison between state SB in the monovalent system and state DB (encompassing both singly and doubly bound complexes) in the bivalent system may not be possible. Another reason may be that an OTs counter ion was directly involved in the binding process of system **DiC6-(G3+2H)**-2OTs (cp. Figure 6.10 a) – a factor that might increase the off rate – which was not the case for system **C6-MonoG1+H**-OTs, where the

OTs molecule associated with the complex only after its formation. Again, additional presampling covering more possible scenarios of counter ion placement throughout the binding process might be a remedy for this problem.

$$P_c(\tau) = \begin{matrix} & \text{UB1} & \text{UB2} & \text{AB} & \text{UB3} & \text{DB} \\ \text{UB1} & \left(\begin{matrix} 0.7354 & 0.0549 & 0.1181 & 0.0143 & 0.0773 \\ 0.2206 & 0.4077 & 0.0118 & 0.2394 & 0.1204 \\ 0.0215 & 0. & 0.7518 & 0. & 0.2267 \\ 0. & 0.0470 & 0. & 0.8021 & 0.1510 \\ 0.0022 & 0.0097 & 0.0211 & 0.0113 & 0.9557 \end{matrix} \right) & \text{with } \tau = 100 \text{ ps} \end{matrix} \quad (6.2)$$

Unfortunately, due to the fact that the singly bound state of system **DiC6-(G3+2H)-2OTs** is not identified as a metastable state of the system – and thus does not represent a proper kinetic entity in the model of the binding process – one cannot put a number on the transition between the singly bound state and the doubly bound state, and therewith on the probability of rebinding, once the complex has partially dissociated. Yet, one could also argue that it is precisely the lack of a proper singly bound state that indicates a significant impact of rebinding in this system, similar to a driving force that continuously pushes the system towards a stable doubly bound state. Compared to systems with gated binding pockets that require the crossing of an energy barrier, the system at hand has freely accessible binding sites that should promote the occurrence of rebinding events.

In their non-atomistic study of multivalent ligand-receptor systems, Weber *et al.* argue that the impact of rebinding in a given system, including scenarios where partially bound states do not form discrete kinetic entities, can be quantified by the degree with which the almost bound state of the system is attributed to the bound state [30]:

$$\tilde{w}_{\text{TB}} = w_{\text{TB}} + \underbrace{\chi_{\text{bound}}(\text{AB}) \cdot w_{\text{AB}}}_{\text{rebinding effect}}. \quad (6.3)$$

In Equation 6.3, the factual weight of the “totally bound” state TB (with all binding sites served), \tilde{w}_{TB} , is the sum of the thermodynamic weight of state TB, w_{TB} , and the degree of membership of the almost bound state AB to the bound state, expressed in terms of the membership function χ_{bound} . Consequently, if one interprets the transient singly bound state as being the almost bound state of the doubly bound state – the only stable bound state of the system – one can come to the conclusion that rebinding would indeed be an important factor for the stability of complex **(G3+2H)•DiC6**. The same holds true for system **C6-(MonoG1+H)-OTs**, which, despite being monovalent,

can benefit from rebinding events as well. In this respect, the binding model presented above provides more information w.r.t. to the occurrence of rebinding events than the long-time MD simulations conducted earlier, where a dissociation event (let alone a rebinding event) could not be recorded once. Of course, it would be possible to predefine the singly bound state as a kinetic entity, and then start additional short-time transition simulations from there to derive at least a propensity for rebinding. Finally, for this particular pair of systems, it has to be expected that the presence of counter ions may have a (possibly adverse) impact on the occurrence of rebinding events: The association of OTs to the unbound ammonium moiety of the guest might handicap or delay the rebinding process. This might promote the complete dissociation of the complex, in particular for the monovalent system.

6.4 Conclusion

In summary, the conformational analysis using ZIBgridfree delivered a characterization of the binding process in terms of a Markov State Model for both monovalent and bivalent system at a comparatively low cost, in particular w.r.t. to approaches that rely on large numbers of free diffusion trajectories [212]. The cost in terms of nanoseconds of MD is listed in Table 6.1. The simulations have given an insight into the mechanism of a bivalent binding process in explicitly modeled solvent, starting by free diffusion of host and guest molecule, and leading to the formation of a stable doubly bound complex.

	I. Presampling MD (<i>NVT</i>) [ns]	II. Restrained MD (<i>NVT</i>) [ns]	III. Unrestrained MD (<i>NVE</i>) [ns]	Σ [ns]
C6-(MonoG1+H)-OTs	50.0	16 \times 2.5	45 \times 1.0	135.0
DiC6-(G3+2H)-2OTs	50.0	31 \times 2.5	95 \times 1.0	222.5

TABLE 6.1: Computational cost of ZIBgridfree in terms of nanoseconds of MD for systems **C6-(MonoG1+H)-OTs** and **DiC6-(G3+2H)-2OTs**, for presampling, restrained sampling (stationary distribution), and unrestrained sampling (transition behavior).

The analysis revealed the differences between monovalent and bivalent binding process regarding the number and population of the associated metastable states. Based on the eigenvector-based cluster analysis by PCCA+, a strong indicator for the presence of a cooperative effect in the formation of the intramolecular binding in the bivalent system (i.e. the cyclization of complex **(G3+2H)•DiC6**) was found. Furthermore, the transition probabilities between the different metastable states of both systems were derived for a short time span of 100 ps. Although rebinding events have not been observed explicitly within the course of the simulations, the clustering results allow to draw the conclusion that both systems should benefit from rebinding, which of course is more likely to occur

in the bivalent system. Finally, it was evaluated in how far the counter ion OTs is involved in the binding process and the resulting host-guest complexes.

The main challenge in the study of binding processes by means of ZIBgridfree is the generation of a sufficient presampling. Due to the somewhat unpredictable counter ions and the heterogeneous solvent mixture, the vacuum shotgun method for obtaining a more complete ensemble of binding paths could not be employed (cp. Section 6.2). Consequently, if one is dependent on free diffusion binding trajectories such as in the case at hand, the approach loses some of its efficiency. For future studies, a system of similar size, but within homogeneous solvent (and preferably without bulky counter ions such as OTs) would be desirable. Finally, in order to study the impact of rebinding events more thoroughly, a ligand-receptor pair with low affinity – and thus being subject to frequent dissociation events, even on the nanosecond timescale – would be more promising. In this context, structurally different bivalent guest molecules ought to be investigated w.r.t. to differences in the (re)binding behavior, depending on the degree of preorganization that is conveyed by spacer length and rigidity.

6.5 Experimental setup

Structures of the host-guest complexes and the tosylate counter ion were modeled using the visualization software Amira [139], and parametrized for the Amber-99SB force field [149] using the software ACPYPE [219] and Antechamber [153, 154], with charges calculated by the AM1-BCC method [156, 157]. All simulations were performed using the double precision version of the software GROMACS, version 4.55 [76]. The initial complex configurations were obtained by energy minimization with the steepest descent algorithm *in vacuo*. The structures in complexed and non-complexed form (with a guest molecule displaced by 1.5 nm) were placed in cubic boxes of 6.5 nm side length and solvated in a 10:1 mixture of chloroform and methanol. The force field parameters for chloroform and methanol were obtained from the GROMACS Molecule & Liquid Database at URL <http://virtualchemistry.org/gmld.php> [220, 221]. To neutralize the overall charge, an adequate amount of counter ions was added to the simulation boxes. The energy of the systems was minimized with the steepest descent algorithm, and 2 ns position restrained MD simulations were performed in order to settle the solvent molecules. These systems were subsequently used as starting conformations for unrestrained simulations of five times 10 ns per system, i.e. monovalent **C6-(MonoG1+H)**-OTs and bivalent **DiC6-(G3+2H)**-2OTs, starting in the non-complexed form. The ten 10 ns simulations were complemented by two 100 ns long-time simulations of the two systems starting

in the complexed form meant for evaluating complex stability (and for recording dissociation events, which failed). To maintain a constant temperature of 298 K in the *NVT* simulations, velocity rescaling [86] was applied. A twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm [75] was used for Coulomb interactions, with a switching distance of 1.0 nm. The integration step was set to 1 fs for all simulations. The same simulation parameters were applied to the ZIBgridfree simulations, exempt from the transition node samplings, where no thermostat was applied in order to realize an *NVE* ensemble setup. The error threshold for the symmetrization of the *S* matrices was set to 10^{-3} .

Chapter 7

Conclusion and outlook

This thesis has demonstrated different approaches of studying multivalent chemical systems using the methodical toolkit and theoretical framework of molecular simulation, starting with the conformational analysis of isolated uncapped and capped spacer structures in solvent (Study F1, Chapter 4.2), and ending with the simulation of a complete bivalent host-guest binding process (Chapter 6). The structures under observation ranged from small synthetic compounds (Studies F1 and F2, Chapters 4.2 and 4.3) to large bioconjugate structures (Studies R1 and R2, Chapters 5.2 and 5.3), from biological structures such as estrogen receptor to completely synthetic structures such as the host-guest systems presented in Chapter 6.

Chapters 4 and 5 focused on aspects of the architecture of multivalent compounds, addressing topics such as the estimation of conformational entropy loss upon binding, the prediction of conformational changes in solvent, and the calculation of critical structure-dependent parameters of ligand presentation. The largest part of these studies was based on realistic multivalent systems provided by cooperating chemists. Consequently, wherever possible, efforts were made to relate the simulation results to data from “wet” experimental methods, namely NMR spectroscopy and binding affinity measurements, to at least obtain a qualitative validation of the computational predictions. The joint results of experimental and computational studies reveal a number of problems connected to using fully flexible spacer structures for bridging hydrophobic ligands in polar solvent, and confirm the advantages attested to more rigid scaffold structures constructed from nucleic acids regarding the defined and unobstructed multivalent presentation of the ligand moieties. The results also underline that, beyond nucleic acids, there is a demand for rigid, water-soluble as well as biocompatible compounds in order to extend the molecular construction kit that is available to the chemists who design multivalent spacers and scaffolds.

Furthermore, the results suggest that the conformational entropy loss upon binding in flexible spacer chains (at least in the scenario “hydrophobic ligands – polar solvent”) is comparatively low, mainly due to the fact that enthalpic effects such as π - π stacking interactions and spacer folding and wrapping dominate the dynamics and restrain the conformation of the flexible spacer even in the unbound state. Consequently, these factors, rather than conformational entropy loss, appear to be the limiting factor w.r.t. to binding affinity.

Chapter 6 changed focus from the rather static comparison between unbound and bound state to actually monitoring the dynamics of a bivalent binding process of a small synthetic host-guest system, including the comparison to the binding process of an analogous monovalent system. The sampling of binding processes in explicit solvent settings belongs to the challenges of molecular simulation, and was realized by applying the importance sampling scheme ZIBgridfree so that the sampling could be obtained at a relatively low computational cost in terms of overall simulation time. The analysis of the sampling data encompassed both the eigenvector-based cluster analysis of the overlap integral matrix – revealing the metastable states of the system and their statistical incidence – as well as the calculation of the transition probability matrix – describing the flux between the metastable states for a short time span. For both monovalent and bivalent system, a notable part of the metastable “almost bound” state is assigned to the bound state, which, according to the theory of Weber *et al.*, is an indicator for the propensity to rebind [30]. Given that both host-guest systems allow for energy-barrier free binding (not considering counter-ion related effects), rebinding should play an important role for the stability of the complexes – although it could not be observed in the preceding long-time simulations. For the bivalent system, the singly bound state was attributed in part to the almost bound state, and, in a much larger extent, to the doubly bound state. The fact that the singly bound state does not represent a metastable kinetic entity in the binding process is a strong indicator for the presence of a cooperative effect in the bivalent system, which would be in accord with the positive chelate cooperativity that was attributed to the complex by double mutant cycle analysis.

Despite this fact, the overall fraction of the unbound state was found to be larger for the bivalent system in the range of about 17 %. This might either be an artifact of insufficient sampling, or stem from subtle effects that are triggered by the association of the tosylate counter ion to the guest molecule. Unfortunately, the positioning of the counter ion could not be generalized for monovalent and bivalent case, so that a certain bias cannot be excluded. Furthermore, the transition probability matrix reveals that a notable part of the unbound state in the bivalent system allows for direct transitions into the doubly bound state, which indicates that the categorization into unbound and almost bound state (and consequently the weighting) in this case is somewhat diffuse.

Therefore, a direct comparison of the weights is flawed, as the soft definition of states by PCCA+ based on the eigenvectors of the overlap matrix leads to slightly different characterizations of what “unbound”, “almost bound” and “bound” means in the case at hand.

In summary, the author of this thesis hopes that the results can make a small contribution to the future design of multivalent compounds, as well as to the understanding of the mechanisms of multivalent binding. Of course, a lot of work remains to be done. In principle, the ZIBgridfree approach is efficient enough to be applied to larger series of compounds. The next step could be to compare binding processes of low valency (bi-, tri- and tetravalent) using different ligand moieties – possibly weaker ones with regard to observing rebinding events – and varying spacer/scaffold structures for the guest molecule. Given that a training set of “wet” experimental measurements will become available, a mid-term goal could be to setup a pipeline of computational tools that allow for the prediction of effective molarities (and hence chelate cooperativity), based on the analyses of binding processes obtained by molecular simulation, in a similar manner as it is done for predicting binding free energies of small molecules. In order to improve the comparability of results, the method should include the optional user-defined state definition in terms of “hard” Voronoi cells.

Another imminent task in the context of multivalency is to adapt the methods of molecular simulation such that they can meet the challenge of dealing with large-scale polyvalent systems, such as the interactions at the interface of cell surfaces and functionalized nanoparticles, e.g. for the determination of optimal particle shape and curvature, or ligand loading ratios. Although the atomistic simulation of such compounds will become more and more accessible with the continuing progress in parallel computing w.r.t. to both hardware as well as algorithms, the question remains if pursuing this approach is worthwhile. Structure preparation, data analysis and storage alone would demand a tremendous amount of resources that may not be justifiable – given that many questions associated with nanoscale compounds can be answered by electron microscopy and other non-computational methods in a more reliable manner.

This calls for approaches that maintain an efficient pipeline of model building, data generation and analysis that can keep up with the work that is done in lab in terms of time and resources spent. One means to achieve this is to resort to coarser levels of description (coarse-grained force fields, or even modeling by differential equations such as in systems biology and pharmacokinetics). Ideally, one would pursue an adaptive approach where critical substructures are simulated in atomistic simulations of manageable size and cost in order to obtain parameters that can then be fed into a coarse simulation model able to run on a laptop computer or web interface to enable direct comparison with experimental data.

Appendix A

Validation of a ZIBgridfree-GROMACS hybrid sampling framework

This section is based on (and contains content from) the following publication:

- A. Bujotzek, O. Schütt, A. Nielsen, K. Fackeldey, M. Weber: Efficient conformational analysis by partition-of-unity coupling. Accepted for publication in *J. Math. Chem.* 2013.

A.1 Introduction

ZIBgridfree as a sampling algorithm that implements the notion of conformation dynamics (cp. Chapter 3.2.1) has been introduced as early as 2005 [82, 96]. The original algorithm was tailored for the conformational analysis of small, drug-like molecules in vacuum. The corresponding implementation of the algorithm was centered around the hybrid Monte Carlo (HMC) algorithm [80, 81] for generating the sampling, using a customized (but merely sequential) version of the Merck molecular force field [222].

More recently, the demand to apply the algorithm to increasingly large systems, and a change of focus from single-molecule conformational analysis to the study of ligand-receptor binding processes [223] has created the need to modify the algorithm and its implementation accordingly, in particular with regard to allowing for a flexible choice of force fields, and, last but not least, the ability to conduct explicit solvent simulations. Therefore, it was decided to delegate the propagation of the dynamics (and most of all

force field evaluation) to a molecular dynamics (MD) software package.¹

In a first draft of the new implementation, it was attempted to retain the HMC scheme of the original ZIBgridfree and to use GROMACS only for the generation of the trial state \tilde{q} (cp. Chapter 3.1.3). This approach suffered significantly from two main issues: (i) For large systems (and certainly as soon as explicit solvent was involved) the acceptance probability of HMC deteriorated rapidly due to very pronounced energy fluctuations, and (ii) for short MD trajectories (as required in HMC), parallel computation was rendered inefficient due to the large computational overhead caused by initiating the communication between the processors via the message-passing interface (MPI). The latter point nullified the performance gain that can be achieved by parallel force field evaluation, and thus was the reason why this idea had to be abandoned.

A second draft of the new implementation was centered around integrating the ZIBgridfree potential modification (cp. Chapter 3.2.1) directly into the GROMACS source code (written in C), and extending the GROMACS executables such that they are able to read in all necessary information regarding the discretization so that an evaluation of the basis functions throughout the simulation is possible. This approach, however, involved extensive modifications in several parts of the GROMACS libraries, and would have required time-consuming maintenance work whenever a new version of the original software is released. Therefore, the integration into the MD code was abandoned as well, in favor of a hybrid framework that treats the MD software merely as a black box, which has the advantage that new versions of GROMACS can be plugged in as they are released. Furthermore, the user does not have to cope with compiling and maintaining an additional, non-standard version of GROMACS, which on large computing clusters can be a tedious task. The algorithmic modifications that are necessary to realize this approach have been discussed in Chapter 3.2.3.

The algorithmic framework of the current hybrid implementation is written in Python. Certain functions, such as for reading trajectories and resolving periodic boundary conditions, have been borrowed from the original GROMACS source. In general, the author of this thesis perceives Python as a good programming language for academic code, as it is accessible, enforces good readability by its syntax conventions, and has a comparably low maintenance. Extensive open source libraries for scientific computing are available, most notably `numpy` [224, 225] and `scipy` [226]. The successive steps of the algorithm have been organized into separate scripts that can easily be modified or extended in a modular fashion. Furthermore, focus was laid on making the use of ZIBgridfree comprehensible and transparent. This motivation led to development of the graphical user interface `zgf_browser`, which offers a number of visualization and analysis tools for

¹The choice finally fell on the software GROMACS [76], for reasons of availability, performance, and its extensive documentation and community support.

monitoring the discretization, sampling progress and clustering results. Finally, ZIBgridfree was adapted to the HLRN computing cluster, so that ZIBgridfree sampling jobs can be submitted directly to the according HLRN job queue. This allows to exploit the massively parallel computing environment by using both a parallel force field evaluation (implemented by GROMACS) and a parallel evaluation the discretized conformational space in terms of the basis functions (implemented by ZIBgridfree). The implementation of this new version of the algorithm was supported by the programming work of Ole Schütt, and mathematical advice provided by Konstantin Fackeldey, both from the working group of PD Dr. Marcus Weber.²

The extensions of the algorithm and the broadening of its scope towards explicit solvent simulations demanded a re-evaluation of its performance, which is presented in the following sections.

A.2 Results

A.2.1 Pentane *in vacuo*

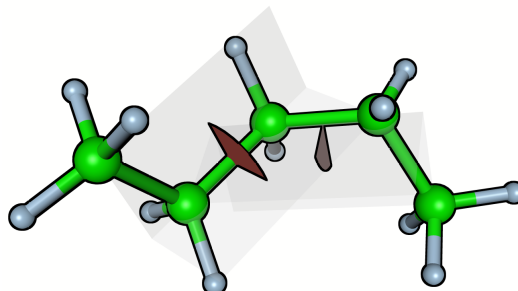


FIGURE A.1: Three-dimensional representation of *n*-pentane. The two backbone torsion angles chosen as internal coordinates are highlighted.

In order to evaluate basic properties of the algorithm, vacuum simulations of *n*-pentane, a small alkane with five carbon atoms (see Figure A.1), were conducted. The two backbone torsion angles of *n*-pentane were chosen as internal coordinates for the discretization. With regard to these internal coordinates, *n*-pentane has nine main conformations, separated by distinct energy barriers. The presampling of conformational space was obtained in terms of a 100 ns MD simulation at a very high (and physically unrealistic) temperature of 1000 K. Reference weights for the conformations of *n*-pentane were taken from the literature [227] (see Table A.1).

²same affiliation as the author, namely Arbeitsgruppe Mathematischer Moleküleentwurf, Abteilung Numerische Analysis und Modellierung, Zuse-Institut Berlin, Takustraße 7, 14195 Berlin (Germany)

c	tr/tr	g^-/tr	g^+/tr	tr/ g^-	tr/ g^+	g^+/g^+	g^-/g^-	g^+/g^-	g^-/g^+
\tilde{w}_c	0.473	0.120	0.132	0.117	0.132	0.013	0.012	< 0.005	< 0.005

TABLE A.1: Conformational weights of n -pentane at 300 K, derived from a hybrid Monte Carlo (HMC) simulation using the Merck molecular force field [222]. tr(ans): $\approx \pm 180^\circ$, $g(\text{auche})^+$: $\approx +60^\circ$, $g(\text{auche})^-$: $\approx -60^\circ$. Torsion angles are given on the scale $[-180, \dots, 180]$.

Stability regarding randomness of impulse and discretization

In order to monitor the impact of choosing a different discretization (placing of nodes in conformational space) on the sampling outcome, three experiments with ten runs of ZIBgridfree each were conducted: a) Equally placed nodes, but random MD starting impulse, b) randomly placed nodes, but equal MD starting impulse, and c) randomly placed nodes and random MD starting impulse. All runs were conducted with 20 discretization nodes and a minimum sampling time of 100 ps per node, leading to a mean overall sampling time per run of 2.8 ns. The results are shown in Figure A.2, left.

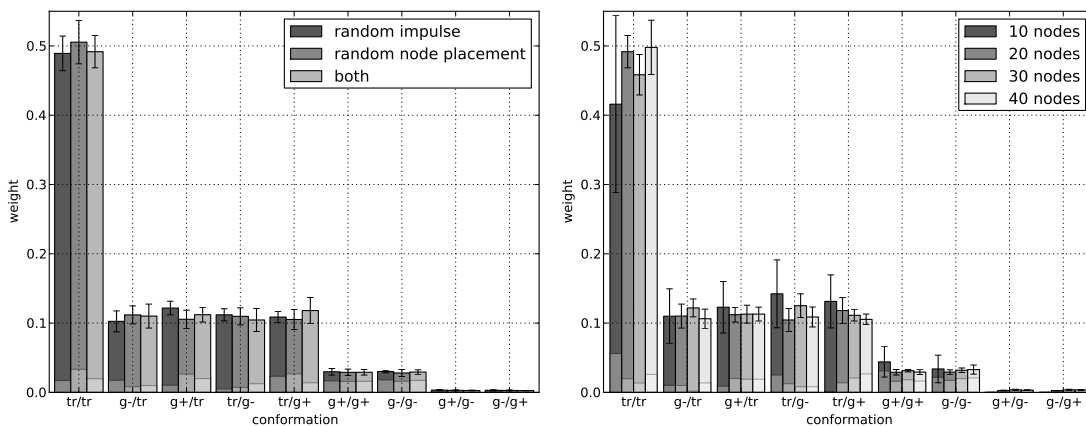


FIGURE A.2: Conformational weights of n -pentane. Error bars indicate the standard deviation w.r.t. 10 runs. Deviation from the literature values is indicated as intra-bar plot. Left: 20 nodes, 100 ps minimum sampling time per node, with equally placed nodes, random MD starting impulse (dark gray), randomly placed nodes, equal MD starting impulse (gray), and randomly placed nodes and random MD starting impulse (light gray). Right: 100 ps minimum sampling time per node, 10, 20, 30 and 40 nodes (dark gray to light gray), random MD starting impulse, random node placement.

Randomizing the MD starting impulse leads to a maximum standard deviation of 0.025 regarding the weight of the most dominant conformation, tr/tr. Randomizing the node placement by picking different initial seeds for the k -means algorithm leads to a maximum standard deviation of 0.031 for conformation tr/tr. When both MD starting impulse and node placement are randomized at the same time (mimicking a standard sampling setup), the maximum standard deviation is slightly smaller (0.23 for conformation tr/tr), which indicates that the uncertainty regarding both choices is not additive.

Stability regarding fineness of discretization

Similar simulations (random MD starting impulse, random node placement, 100 ps minimum sampling time per node) were performed with varying number of sampling nodes in order to evaluate the impact of the fineness of the discretization. For this experiment, automatic refinement of the discretization was switched off. The results are shown in Figure A.2, right. When only ten discretization nodes are used (only one more than the expected number of conformations), the error becomes very large (0.128 for conformation tr/tr), and, despite a relatively large mean overall sampling time of 3.2 ns per run, the rare conformations g+/g- and g-/g+ are not identified at all. For 20, 30 and 40 discretization nodes (mean overall sampling times 2.79, 4.45 and 5.5 ns per run), the results are comparable, but do not improve visibly with increasing fineness of the discretization.

Stability regarding sampling time

Finally, it was looked into how the sampling time per node determines the quality of the results. The outcome is shown in Figure A.3. A very short minimum sampling time of 10 ps per node produces a large error (0.099 for conformation tr/tr), but, given the mean overall sampling time of only 365 ps per run, the averaged conformational weights are acceptable. With increasing sampling time per node, the error can be significantly reduced. For a minimum sampling time of 1000 ps per node (mean overall sampling time 26.7 ns), the maximum standard deviation (conformation tr/tr) is reduced to 0.016, and below one percent for all other conformations. One can conclude that a rough estimate of the conformational weights can be obtained at a very low cost, whereas precise results have to be paid for with thorough sampling of the partial densities.

The results show a perceivable deviation w.r.t. to the conformational weights found in the literature (cp. Table A.1), which most likely can be attributed to the use of a different force field and (possibly) the different dynamics for propagating the system. For comparison, the conformational weights obtained from ZIBgridfree with 25 nodes and 1000 ps minimum sampling time per node, averaged over ten runs, are given in Table A.2.

c	tr/tr	g^-/tr	g^+/tr	tr/g^-	tr/g^+	g^+/g^+	g^-/g^-	g^+/g^-	g^-/g^+
\tilde{w}_c	0.486	0.113	0.113	0.116	0.110	0.027	0.029	0.003	0.004

TABLE A.2: Averaged conformational weights of *n*-pentane at 300 K, derived from ten runs of ZIBgridfree using the Amber-99SB force field.

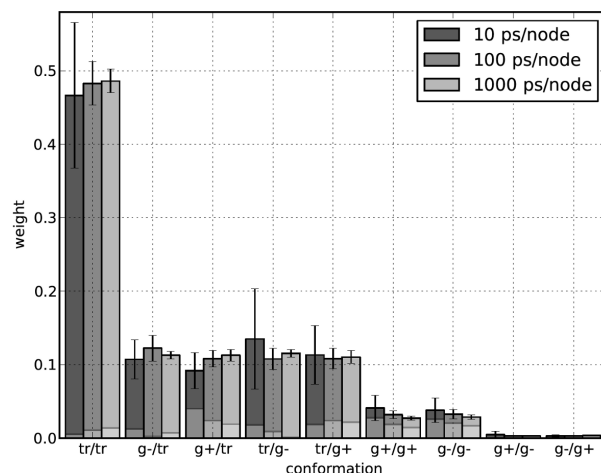


FIGURE A.3: Conformational weights of n -pentane. Error bars indicate the standard deviation w.r.t. 10 runs. Deviation from the literature values is indicated as sub-bar plot. 25 nodes, with 10, 100 and 1000 ps minimum sampling time per node (dark gray to light gray), random MD starting impulse, random node placement.

A.2.2 Alanine dipeptide in water

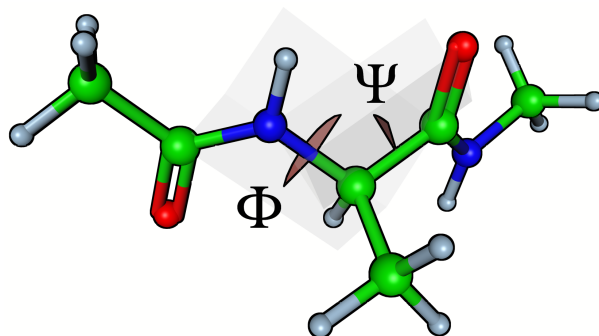


FIGURE A.4: Three-dimensional representation of alanine dipeptide (ACE-ALA-NME, i.e. terminally blocked alanine). The two backbone torsion angles Φ and Ψ chosen as internal coordinates are highlighted.

As a second example, the conformations of alanine dipeptide in explicit TIP4P-Ew water were studied. Alanine dipeptide is the most basic (or “minimal”) polypeptide and serves as a popular test case for evaluating biological force fields. The two backbone torsion angles Φ and Ψ span the relevant conformational space of alanine dipeptide, and were hence chosen as internal coordinates for the discretization. With regard to these internal coordinates, alanine dipeptide has six main conformations, which however are not as well-separated as in the previous example, n -pentane. Obtaining correct conformational weights from explicit solvent simulations is more difficult compared to vacuum or implicit solvent settings, as the dynamics of a solvated system is decelerated, while the computational cost of producing sufficient sampling data multiplies.

Reference weights for the conformations of alanine dipetide at 300 K in the NVT and in the NpT ensemble were obtained from two 200 ns MD simulations (see Table A.3).

	c	C_5	P_{II}	α_R	α_P	α_L	C_7^α
NVT	\tilde{w}_c	0.2696	0.4043	0.1745	0.1369	0.0136	0.0010
NpT	\tilde{w}_c	0.2794	0.4363	0.1563	0.1190	0.0070	0.0020

TABLE A.3: Conformational weights of alanine dipeptide at 300 K in the NVT and in the NpT ensemble, derived from two 200 ns MD simulations using the Amber-99SB force field. C_5 : $\approx 143^\circ/-158^\circ$, P_{II} : $\approx 70^\circ/-158^\circ$, α_R : $\approx 70^\circ/11^\circ$, α_P : $\approx 136^\circ/-11^\circ$, α_L : $\approx 55^\circ/-40^\circ$, and C_7^α : $\approx -60^\circ/\pm 180^\circ$. Torsion angles are given on the scale $[-180, \dots, 180]$. Conformation labels taken from Chodera *et al.* [228].

Explicitly modeled water also complicates the presampling of conformational space: High (or elevated) temperature presampling is possibly only to a certain extent, and requires a re-equilibration of the simulation boxes before the partial densities can be sampled at the target temperature. In principle, discretization nodes can also be picked from a vacuum or implicit solvent trajectory of the molecule of interest, to be put in explicit solvent only before the sampling of partial densities with ZIBgridfree is commenced (implemented in the tools `zgf_solvate_nodes` and `zgf_genion`). Again, another cycle of energy minimization and simulation box equilibration is needed before usable sampling data can be collected. For this example, the presampling consisted of a 100 ns MD trajectory at the target temperature of 300 K, which means that re-equilibration after node selection was not necessary.

Stability regarding sampling time

First, it was looked into how the sampling time per node determines the quality of the results using random MD starting impulse and random node placement in an NVT ensemble. The outcome is shown in Figure A.5. In comparison to the (vacuum) n -pentane example, a longer minimum sampling time per node is required in order to yield acceptable results. For a very short minimum sampling time of 10 ps per node, the results were not interpretable due to the large error (data not shown). A minimum sampling time of 100 ps per node (mean overall sampling time 2.4 ns) produces large errors of around 15 % in terms of standard deviation for the three largest conformations P_{II} , C_5 and α_R . When the minimum sampling time per node is increased to 500 ps (mean overall sampling time 7.7 ns), the error can be reduced below 6 % for all conformations (largest error is 0.0581 for conformation P_{II}). Finally, with a minimum sampling time of 1000 ps per node (mean overall sampling time 15 ns), the error is in the range of 5 %, and mainly below (largest error is 0.0533 for conformation P_{II}).

An auxiliary trial with a minimum sampling time of 1000 ps per node (mean overall sampling time 15.56 ns) using a double precision version of GROMACS did not lead to

a further decrease in standard deviation, contrary to what might have been expected from an increase in precision of coordinates and observables.

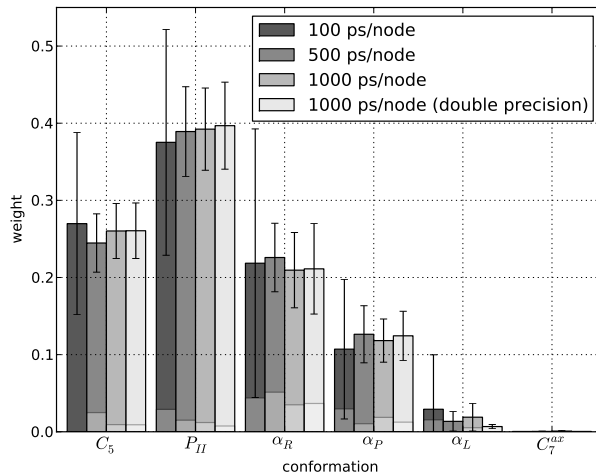


FIGURE A.5: Conformational weights of alanine dipeptide. Error bars indicate the standard deviation w.r.t. 10 runs. Deviation from the reference values is indicated as sub-bar plot. 15 nodes, with 100, 500 and 1000 ps minimum sampling time per node (dark gray to light gray), including an auxiliary 1000 ps double precision trial, random MD starting impulse, random node placement.

Stability regarding choice of dynamics

Second, similar simulations (random MD starting impulse, 15 randomly placed nodes, 1000 ps minimum sampling time per node) were performed while exchanging the common MD integrator with a stochastic dynamics (SD) integrator. Both integrators were compared in the context of an NVT and an NpT ensemble, the latter realized by using a Parrinello-Rahman barostat. All trial runs were conducted with a double precision version of GROMACS. The results are shown in Figure A.6. In both NVT and NpT ensemble, the SD integrator delivers better results with regard to the standard deviation over ten runs. In the NVT ensemble, the largest error obtained with the SD integrator is 3.618 % (conformation P_{II}), compared to 5.86 % when the MD integrator is used (conformation α_R). This gap becomes somewhat closer in the NpT ensemble, where the largest error obtained with the SD integrator is 5.35 %, compared to 6.3 % when the MD integrator is used (both w.r.t. conformation α_R).

The chosen dynamics also has an impact on the mean conformational weights. When the SD integrator is used, the largest conformation, P_{II} is sampled less dominant than with the MD integrator (NVT : 36.18 % compared to 39.68 %, and NpT : 39.02 % compared to 44.93 %). Instead, the conformational weight is distributed more equally over the minor conformations α_R , α_P and α_L .

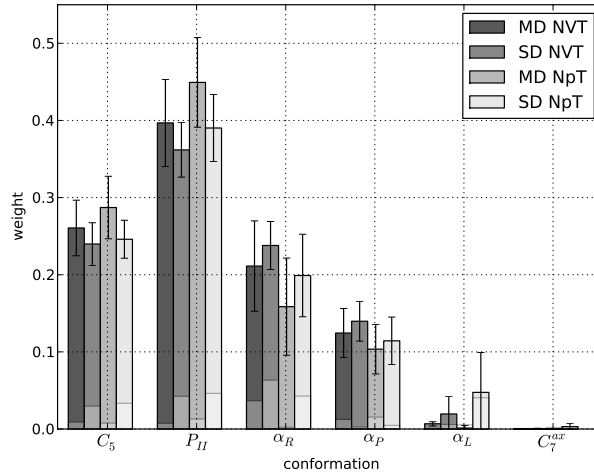


FIGURE A.6: Conformational weights of alanine dipeptide. Error bars indicate the standard deviation w.r.t. 10 runs. Deviation from the reference values is indicated as sub-bar plot. 15 nodes, 1000 ps minimum sampling time per node, random node placement, with MD integrator (NVT), SD integrator (NVT), MD integrator (NpT), and SD integrator (NpT), dark gray to light gray.

The results show an acceptable agreement with the reference weights that were extracted from the 200 ns MD trajectory for all runs using 500 ps or more minimum sampling time per node, at least for the runs conducted with the MD integrator (i.e. the same integrator that was used for the long-time trajectories used as reference). Long-time data from the SD integrator is not available, but it can be expected to deliver a slightly different distribution. In general, the largest deviation is found for the α_R conformation: ZIBgridfree tends to overweight α_R by about 4 %, a weight that is mostly drawn from the α_P , and partly from the α_L conformation. As the conformations of alanine dipeptide tend to have notable overlapping regions (as opposed to the well-separated conformations of n -pentane), the error might not only be due to insufficient sampling, but also to imperfect clustering of certain states in transient regions. For comparison, the conformational weights in the NVT and the NpT ensemble, obtained from ZIBgridfree with 15 nodes and 1000 ps minimum sampling time per node and averaged over ten runs, are given in Table A.4.

	c	C_5	P_{II}	α_R	α_P	α_L	C_7^α
NVT	\tilde{w}_c	0.2606	0.3968	0.2112	0.1244	0.0067	0.0003
NpT	\tilde{w}_c	0.2871	0.4493	0.1586	0.1035	0.0015	0.0001

TABLE A.4: Averaged conformational weights of alanine dipeptide at 300 K in the NVT and in the NpT ensemble, derived from ten runs of ZIBgridfree using the Amber-99SB force field (MD integrator, double precision GROMACS).

A.3 Conclusion

As far as the limited number of test cases allows, it was shown that algorithm and software perform reasonably well in determining the conformational weights of small molecular systems in both vacuum and solvent at a comparatively low computational cost. A direct performance comparison with long-time MD trajectories is difficult, as the applicability of ZIBgridfree is dependent on the availability of some kind of presampling of conformational space from which discretization nodes can be selected. Hence, the cost of obtaining the presampling would have to be added to the overall balance. The cost of generating an adequate presampling, in turn, is dependent on the system in question. For example, a series of docking results of a small molecule in a protein binding pocket would also serve as a valid starting point for using ZIBgridfree.

Given the efficiency of current MD code in generating even very long trajectories in a relatively short time, the need for a complex algorithm like ZIBgridfree can be questioned. However, depending on the size of the system under observation, the possibilities of force field parallelization can be limited. Here, another level of parallelization on the basis function level might be helpful. Furthermore, the analysis, handling and storage of very long trajectories can be tedious. ZIBgridfree uses short trajectories that are generated in a directed way, and the convergence of the sampling is monitored. Furthermore, the use of collective variables (i.e. internal coordinates) and the integrated clustering approach lead to a level of abstraction that significantly facilitates the analysis of the sampling data, the identification of relevant events and their biological or chemical interpretation.

The source code of ZIBgridfree is available free of charge at URL <https://github.com/CMD-at-ZIB/ZIBMolPy>.

A.4 Experimental setup

All molecular simulations were performed with GROMACS, versions 4.54 and 4.55 (single precision, unless stated differently). All molecules were parametrized for the Amber-99SB force field [149]. Residues not already included in the standard force field were prepared using the software ACPYPE [219] and Antechamber [153, 154] from AmberTools [155], with charges calculated by the AM1-BCC method [156, 157].

For the vacuum simulations (*n*-pentane), van der Waals and Coulomb interactions were computed without cut-off (all vs. all). For the explicit solvent simulations (alanine dipeptide), the TIP4P-Ew water model [150, 151] was used. The solute was placed in a rhombic dodecahedron simulation box of 4.0 nm side length, containing 1523 solvent molecules. A twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied

and the smooth particle mesh Ewald algorithm [75] was used for Coulomb interactions, with a switching distance of 1.0 nm.

In order to generate the NVT ensemble of states for the desired temperature of 300 K, either the velocity-rescaling thermostat [86] in combination with an MD leap-frog integrator, or a Langevin-type stochastic dynamics [229] integrator was used. For the explicit solvent NpT simulations (alanine dipeptide), the velocity-rescaling thermostat/stochastic dynamics integrator was supplemented by the Parrinello-Rahman barostat [91, 92], with a reference pressure of 1 bar. The integration step was set to 1 fs for all simulations. The error threshold for the symmetrization of the S matrices was set to 10^{-2} for pentane, and to 10^{-4} for alanine dipeptide.

Bibliography

- [1] C. Fasting, C.A. Schalley, M. Weber, O. Seitz, St. Hecht, B. Kokschi, J. Dervede, Ch. Graf, E.W. Knapp, and R. Haag. Multivalenz als chemisches Organisations- und Wirkprinzip. *Angew. Chem.*, 124(42):10567–10567, 2012.
- [2] M. Mammen, S.K. Choi, and G.M. Whitesides. Polyvalent interactions in biological systems: Implications for design and use of multivalent ligands and inhibitors. *Angew. Chem. Int. Ed.*, 37(20):2754–2794, 1998.
- [3] W.J. Lees, A. Spaltenstein, J.E. Kingery-Wood, and G.M. Whitesides. Polyacrylamides bearing pendant α -sialoside groups strongly inhibit agglutination of erythrocytes by influenza A virus: Multivalency and steric stabilization of particulate biological systems. *J. Med. Chem.*, 37(20):3419–3433, 1994.
- [4] M. Mammen, G. Dahmann, and G.M. Whitesides. Effective inhibitors of hemagglutination by influenza virus synthesized from polymers having active ester groups. Insight into mechanism of inhibition. *J. Med. Chem.*, 38(21):4179–4190, 1995.
- [5] P.I. Kitov, J.M. Sadowska, G. Mulvey, G.D. Armstrong, H. Ling, N.S. Pannu, R.J. Read, and D.R. Bundle. Shiga-like toxins are neutralized by tailored multivalent carbohydrate ligands. *Nature*, 403(6770):669–672, 2000.
- [6] D. Schwefel, C. Maierhofer, J.G. Beck, S. Seeberger, K. Diederichs, H.M. Moller, W. Welte, and V. Wittmann. Structural basis of multivalent binding to wheat germ agglutinin. *J. Am. Chem. Soc.*, 132(25):8704–8719, 2010.
- [7] Z. Lu, S. Deng, D. Huang, Y. He, M. Lei, L. Zhou, and P. Jin. Frontier of therapeutic antibody discovery: The challenges and how to face them. *World J. Biol. Chem.*, 3(12):187, 2012.
- [8] J.D. Badjic, A. Nelson, S.J. Cantrill, W.B. Turnbull, and J.F. Stoddart. Multivalency and cooperativity in supramolecular chemistry. *Acc. Chem. Res.*, 38(9):723–732, 2005.

- [9] A. Mulder, J. Huskens, and D.N. Reinhoudt. Multivalency in supramolecular chemistry and nanofabrication. *Org. Biomol. Chem.*, 2(23):3409–3424, 2004.
- [10] J. Huskens. Multivalent interactions at interfaces. *Curr. Opin. Chem. Biol.*, 10(6):537–543, 2006.
- [11] W. Jiang, K. Nowosinski, N.L. Löw, E.V. Dzyuba, F. Klautzsch, A. Schäfer, J. Huskonen, K. Rissanen, and C.A. Schalley. Chelate cooperativity and spacer length effects on the assembly thermodynamics and kinetics of divalent pseudorotaxanes. *J. Am. Chem. Soc.*, 134(3):1860–1868, 2012.
- [12] J.S. Román, A. Gallardo, and B. Levenfeld. Polymeric drug delivery systems. *Adv. Mater.*, 7(2):203–208, 1995.
- [13] S.K. Choi. *Synthetic Multivalent Molecules: Concepts and Biomedical Applications*. John Wiley & Sons, Inc., 2004.
- [14] R. Haag, A. Sunder, and J.F. Stumbé. An approach to glycerol dendrimers and pseudo-dendritic polyglycerols. *J. Am. Chem. Soc.*, 122(12):2954–2955, 2000.
- [15] I. Papp, J. Dervede, S. Enders, and R. Haag. Modular synthesis of multivalent glycoarchitectures and their unique selectin binding behavior. *Chem. Commun.*, 44:5851–5853, 2008.
- [16] M. Shan, A. Bujotzek, F. Abendroth, A. Wellner, R. Gust, O. Seitz, M. Weber, and R. Haag. Conformational analysis of bivalent estrogen receptor ligands: From intramolecular to intermolecular binding. *ChemBioChem*, 12(17):2587–2598, 2011.
- [17] F. Abendroth, A. Bujotzek, M. Shan, R. Haag, M. Weber, and O. Seitz. DNA-controlled bivalent presentation of ligands for the estrogen receptor. *Angew. Chem. Int. Ed.*, 50(37):8592–8596, 2011.
- [18] H. Eberhard, F. Diezmann, and O. Seitz. DNA as a molecular ruler: Interrogation of a tandem SH2 domain with self-assembled, bivalent DNA-peptide complexes. *Angew. Chem.*, 123(18):4232–4236, 2011.
- [19] F. Noé, S. Röblitz, B. Schmidt, Ch. Schütte, and M. Weber. Simulation von Biomolekülen. Lecture notes, Freie Universität Berlin, 2012. URL <http://page.mi.fu-berlin.de/bsch63/Lectures/SimBioMol/Material/script.pdf>.
- [20] P.W. Snyder, J. Mecinović, D.T. Moustakas, S.W. Thomas III, M. Harder, E.T. Mack, M.R. Lockett, A. Héroux, W. Sherman, and G.M. Whitesides. Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proc. Natl. Acad. Sci. USA*, 108(44):17889–17894, 2011.

- [21] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28: 235–242, 2000. URL <http://www.pdb.org>.
- [22] P.L. Freddolino, A.S. Arhipov, S.B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, 2006.
- [23] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.
- [24] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, and A.H. de Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111(27):7812–7824, 2007.
- [25] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23(1-3):3–25, 1997.
- [26] R.D. Taylor, P.J. Jewsbury, and J.W. Essex. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.*, 16(3):151–166, 2002.
- [27] E. Yuriev, M. Agostino, and P.A. Ramsland. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.*, 24(2):149–164, 2011.
- [28] J. Åqvist, C. Medina, and J.E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 7:385–391, 1994.
- [29] J. Numata, A. Juneja, D.J. Diestler, and E.W. Knapp. Influence of spacer–receptor interactions on the stability of bivalent ligand–receptor complexes. *J. Phys. Chem. B*, 116(8):2595–2604, 2012.
- [30] M. Weber, A. Bujotzek, and R. Haag. Quantifying the rebinding effect in multivalent chemical ligand-receptor systems. *J. Chem. Phys.*, 137(5):054111, 2012.
- [31] H. Lei and Y. Duan. Improved sampling methods for molecular simulation. *Curr. Opin. Struct. Biol.*, 17(2):187–191, 2007.
- [32] G. Ercolani, C. Piguet, M. Borkovec, and J. Hamacek. Symmetry numbers and statistical factors in self-assembly and multivalency. *J. Phys. Chem. B*, 111(42): 12195–12203, 2007.

- [33] G. Ercolani. Assessment of cooperativity in self-assembly. *J. Am. Chem. Soc.*, 125(51):16097–16103, 2003.
- [34] S.W. Benson. Statistical factors in the correlation of rate constants and equilibrium constants. *J. Am. Chem. Soc.*, 80(19):5151–5154, 1958.
- [35] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, 4th ed. Garland Science, 2002.
- [36] J.M. Woof and D.R. Burton. Human antibody–Fc receptor interactions illuminated by crystal structures. *Nat. Rev. Immunol.*, 4(2):89–99, 2004.
- [37] P.J. Delves, S.J. Martin, D.R. Burton, and I.M. Roitt. *Roitt's Essential Immunology*. Wiley-Blackwell, 2011.
- [38] A.D. McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology*, 2nd ed. (the “Gold Book”). Blackwell Scientific Publications, 1997. URL <http://dx.doi.org/10.1351/goldbook.C01012>. XML on-line corrected version created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins.
- [39] G. Schwarzenbach. Der Chelateffekt. *Helvetica Chim. Acta*, 35(7):2344–2359, 1952.
- [40] T. Christensen, D.M. Gooden, J.E. Kung, and E.J. Toone. Additivity and the physical basis of multivalency effects: A thermodynamic investigation of the calcium EDTA interaction. *J. Am. Chem. Soc.*, 125(24):7357–7366, 2003.
- [41] V. Vallet, U. Wahlgren, and I. Grenthe. Chelate effect and thermodynamics of metal complex formation in solution: A quantum chemical study. *J. Am. Chem. Soc.*, 125(48):14941–14950, 2003.
- [42] G.K. Ackers, M.L. Doyle, D. Myers, and M.A. Daugherty. Molecular code for cooperativity in hemoglobin. *Science*, 255(5040):54, 1992.
- [43] A. Schön and E. Freire. Thermodynamics of intersubunit interactions in cholera toxin upon binding to the oligosaccharide portion of its cell surface receptor, ganglioside GM1. *Biochemistry*, 28(12):5019–5024, 1989.
- [44] K.A. Dill and S. Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics and Nanoscience*, 2nd ed. Garland Science, 2011.
- [45] M. Mammen, E.I. Shakhnovich, J.M. Deutch, and G.M. Whitesides. Estimating the entropic cost of self-assembly of multiparticle hydrogen-bonded aggregates based on the cyanuric acid-melamine lattice. *J. Org. Chem.*, 63(12):3821–3830, 1998.

- [46] M. Mammen, E.I. Shakhnovich, and G.M. Whitesides. Using a convenient, quantitative model for torsional entropy to establish qualitative trends for molecular processes that restrict conformational freedom. *J. Org. Chem.*, 63(10):3168–3175, 1998.
- [47] R. Lumry and S. Rajender. Enthalpy–entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous property of water. *Biopolymers*, 9(10):1125–1227, 1970.
- [48] Y. Inoue, Y. Liu, L.H. Tong, B.J. Shen, and D.S. Jin. Calorimetric titration of inclusion complexation with modified. beta.-cyclodextrins. enthalpy-entropy compensation in host-guest complexation: from ionophore to cyclodextrin and cyclophane. *J. Am. Chem. Soc.*, 115(23):10637–10644, 1993.
- [49] D.H. Williams, E. Stephens, D.P. O’Brien, and M. Zhou. Understanding noncovalent interactions: Ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew. Chem. Int. Ed.*, 43(48):6596–6616, 2004.
- [50] V.M. Krishnamurthy, B.R. Bohall, V. Semetey, and G.M. Whitesides. The paradoxical thermodynamic basis for the interaction of ethylene glycol, glycine, and sarcosine chains with bovine carbonic anhydrase II: An unexpected manifestation of enthalpy/entropy compensation. *J. Am. Chem. Soc.*, 128(17):5802–5812, 2006.
- [51] N.J. de Mol, M.I. Catalina, F.J. Dekker, M.J.E. Fischer, A.J.R. Heck, and R.M.J. Liskamp. Protein flexibility and ligand rigidity: A thermodynamic and kinetic study of ITAM-based ligand binding to Syk tandem SH2. *ChemBioChem*, 6(12):2261–2270, 2005.
- [52] A. Cornish-Bowden. Enthalpy–entropy compensation: a phantom phenomenon. *J. Biosci.*, 27(2):121–126, 2002.
- [53] K. Fütterer, J. Wong, R.A. Grucza, A.C. Chan, and G. Waksman. Structural basis for Syk tyrosine kinase ubiquity in signal transduction pathways revealed by the crystal structure of its regulatory SH2 domains bound to a dually phosphorylated ITAM peptide1. *J. Mol. Biol.*, 281(3):523–537, 1998.
- [54] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [55] R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

- [56] R.H. Kramer and J.W. Karpen. Spanning binding sites on allosteric proteins with polymer-linked ligand dimers. *Nature*, 395(6703):710–713, 1998.
- [57] J. Rao, J. Lahiri, R.M. Weis, and G.M. Whitesides. Design, synthesis, and characterization of a high-affinity trivalent system derived from vancomycin and L-Lys-D-Ala-D-Ala. *J. Am. Chem. Soc.*, 122(12):2698–2710, 2000.
- [58] D.J. Diestler and E.W. Knapp. Statistical thermodynamics of the stability of multivalent ligand-receptor complexes. *Phys. Rev. Lett.*, 100(17):178101, 2008.
- [59] G. Ercolani and L. Schiaffino. Allosteric, chelate, and interannular cooperativity: A mise au point. *Angew. Chem. Int. Ed.*, 50(8):1762–1768, 2011.
- [60] S.L. Cockroft and C.A. Hunter. Chemical double-mutant cycles: Dissecting non-covalent interactions. *Chem. Soc. Rev.*, 36(2):172–188, 2007.
- [61] A. Camara-Campos, D. Musumeci, C.A. Hunter, and S. Turega. Chemical double mutant cycles for the quantification of cooperativity in H-bonded complexes. *J. Am. Chem. Soc.*, 131(51):18518–18524, 2009.
- [62] M.C. Misuraca, T. Grecu, Z. Freixa, V. Garavini, C.A. Hunter, P.W.N.M. van Leeuwen, M.D. Segarra-Maset, and S.M. Turega. Relationship between conformational flexibility and chelate cooperativity. *J. Org. Chem.*, 76(8):2723–2732, 2011.
- [63] H.J. Hogben, J.K. Sprafke, M. Hoffmann, M. Pawlicki, and H.L. Anderson. Step-wise effective molarities in porphyrin oligomer complexes: Preorganization results in exceptionally strong chelate cooperativity. *J. Am. Chem. Soc.*, 133(51):20962–20969, 2011.
- [64] M. Weber. *A subspace approach to molecular Markov state models via a new infinitesimal generator*. Habilitation thesis, Freie Universität Berlin, 2011. URL <http://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/1402>.
- [65] C.A. Hunter and H.L. Anderson. What is cooperativity? *Angew. Chem. Int. Ed.*, 48(41):7488–7499, 2009.
- [66] L. von Krbek. Multivalente Krone-Ammonium-Komplexe. Master’s thesis, Freie Universität Berlin, Fachbereich Chemie, Biologie, Pharmazie, 2012.
- [67] D. Mollenhauer. *Quantenchemische Untersuchungen zur Wechselwirkung von Pyridinderivaten mit Goldnanopartikeln – Vom Pyridin-Gold-Komplex zur Adsorption auf Goldoberflächen*. Doctoral thesis, Freie Universität Berlin, 2011.
- [68] F. Jensen. *Introduction to Computational Chemistry*. Wiley, 2007.

- [69] G.D. Purvis III and R.J. Bartlett. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.*, 76:1910, 1982.
- [70] M. Head-Gordon, J.A. Pople, and M.J. Frisch. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.*, 153(6):503–506, 1988.
- [71] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, 1976.
- [72] T.A. Halgren and W. Damm. Polarizable force fields. *Curr. Opin. Struct. Biol.*, 11(2):236–242, 2001.
- [73] E. Apol, R. Apostolov, H.J.C. Berendsen, A. van Buuren, P. Bjelkmar, R. van Drunen, A. Feenstra, G. Groenhof, B. Hess, P. Kasson, P. Larsson, E. Lindahl, P. Meulenhoff, T. Murtola, S. Páll, S. Pronk, R. Schulz, M. Shirts, D. Sijbers, A. van der Spoel, and P. Tieleman. *GROMACS User Manual*. 1991–2000: Department of Biophysical Chemistry, University of Groningen, The Netherlands. 2001–2010: The GROMACS development teams at the Royal Institute of Technology and Uppsala University, Sweden., version 4.5.4 edition, 1991–2010. Additional contributions by M. Abraham, Ch. Junghans, C. Kutzner, J.A. Lemkul, E. Marklund, M. Wolf.
- [74] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications, 2nd ed.*, volume 1 of *Computational Science Series*. Academic Press, 2002.
- [75] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577, 1995.
- [76] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.
- [77] A. Riemer. Accuracy, stability, convergence of rigorous thermodynamic sampling methods. Master’s thesis, Freie Universität Berlin, Fachbereich Bioinformatik, 2006.
- [78] A. Fischer. Die Hybride Monte-Carlo-Methode in der Molekülphysik. Master’s thesis, Freie Universität Berlin, Fachbereich Mathematik und Informatik, 1997.
- [79] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6): 1087–1092, 1953.

- [80] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987.
- [81] A. Brass, B.J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33(8):1307–1315, 1993.
- [82] H. Meyer. Die Implementierung und Analyse von HuMFree – einer gitterfreien Methode zur Konformationsanalyse von Wirkstoffmolekülen. Master’s thesis, Freie Universität Berlin, Fachbereich Bioinformatik, 2005.
- [83] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.
- [84] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, et al. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.
- [85] H.C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384, 1980.
- [86] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:014101, 2007.
- [87] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81:511, 1984.
- [88] W.G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695, 1985.
- [89] G.J. Martyna, M.L. Klein, and M. Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97:2635, 1992.
- [90] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.
- [91] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.
- [92] S. Nose and M.L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50(5):1055–1076, 1983.

- [93] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.
- [94] A. Amadei, A. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.*, 17(4):412–425, 1993.
- [95] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.
- [96] M. Weber and H. Meyer. ZIBgridfree – Adaptive conformation analysis with qualified support of transition states and thermodynamic weights. Technical Report 05-17, Konrad-Zuse-Zentrum für Informationstechnik, 2005.
- [97] P. Deuffhard. From molecular dynamics to conformation dynamics in drug design. In M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi, editors, *Trends in Nonlinear Analysis*, pages 269–288. Springer, 2003.
- [98] S. Kube and M. Weber. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.*, 126:024103, 2007.
- [99] M. Weber. *Meshless methods in conformation dynamics*. Doctoral thesis, Freie Universität Berlin, 2006. Department of Mathematics and Computer Science.
- [100] B.L. de Groot, D.M.F. van Aalten, R.M. Scheek, A. Amadei, G. Vriend, and H.J.C. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins: Struct. Funct. Genet.*, 29(2):240–251, 1997.
- [101] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.
- [102] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Sci.*, 7(4):457–472, 1992.
- [103] M. Klimm, A. Bujotzek, and M. Weber. Direct reweighting strategies in conformation dynamics. *MATCH Commun. Math. Comput. Chem.*, 65(2):333–346, 2011.
- [104] M. Weber, S. Kube, L. Walter, and P. Deuffhard. Stable computation of probability densities for metastable dynamical systems. *Multiscale Model. Simul.*, 6(2):396–416, 2007.
- [105] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annal. Math. Statistics*, 35(2):876–879, 1964.

- [106] P. Deuffhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.*, 398:161–184, 2005.
- [107] M. Weber and S. Kube. Robust Perron cluster analysis for various applications in computational life science. *Comput. Life Sci.*, pages 57–66, 2005.
- [108] M. Weber and K. Andrae. A simple method for the estimation of entropy differences. *MATCH Commun. Math. Comput. Chem.*, 63(2):319–332, 2010.
- [109] W. Blokzijl and J.B.F.N. Engberts. Hydrophobic effects. Opinions and facts. *Angew. Chem. Int. Ed.*, 32(11):1545–1579, 2003.
- [110] J. Numata. *Conformational entropy from molecular simulations: Statistical mechanics using the tools of information theory*. Doctoral thesis, Freie Universität Berlin, 2012. Department of Biology, Chemistry and Pharmacy.
- [111] J.B. Thompson, H.G. Hansma, P.K. Hansma, and K.W. Plaxco. The backbone conformational entropy of protein folding: Experimental measures from atomic force microscopy. *J. Mol. Biol.*, 322(3):645–652, 2002.
- [112] K.K. Frederick, M.S. Marlow, K.G. Valentine, and A.J. Wand. Conformational entropy in molecular recognition by proteins. *Nature*, 448(7151):325–329, 2007.
- [113] M.S. Marlow, J. Dogan, K.K. Frederick, K.G. Valentine, and A.J. Wand. The role of conformational entropy in molecular recognition by calmodulin. *Nat. Chem. Biol.*, 6(5):352–358, 2010.
- [114] C. Diehl, O. Engström, T. Delaine, M. Håkansson, S. Genheden, K. Modig, H. Lefler, U. Ryde, U.J. Nilsson, and M. Akke. Protein flexibility and conformational entropy in ligand design targeting the carbohydrate recognition domain of galectin-3. *J. Am. Chem. Soc.*, 132(41):14577, 2010.
- [115] O. Stern. Über eine Methode zur Berechnung der Entropie von Systemen elastisch gekoppelter Massenpunkte. *Annal. Phys.*, 356(19):237–260, 1916.
- [116] J. Schlitter. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.*, 215(6):617–621, 1993.
- [117] H. Schäfer, A.E. Mark, and W.F. van Gunsteren. Absolute entropies from molecular dynamics simulation trajectories. *J. Chem. Phys.*, 113:7809, 2000.
- [118] B.J. Killian, J.Y. Kravitz, and M.K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.*, 127(2):024107, 2007.

- [119] A. Bujotzek, M. Shan, R. Haag, and M. Weber. Towards a rational spacer design for bivalent inhibition of estrogen receptor. *J. Comput. Aided Mol. Des.*, 25(3): 253–262, 2011.
- [120] E. Ottow and H. Weinmann. *Nuclear Receptors as Drug Targets*. WILEY-VCH Verlag GmbH & Co. KGaA, 2008.
- [121] C. Avendano and J. C. Menendez. *Medicinal Chemistry of Anticancer Drugs*. Elsevier B.V., 2008.
- [122] V.C. Jordan. Antiestrogens and selective estrogen receptor modulators as multifunctional medicines. 1. Receptor interactions. *J. Med. Chem.*, 46:883–903, 2003.
- [123] V.C. Jordan. Antiestrogens and selective estrogen receptor modulators as multifunctional medicines. 2. Clinical considerations and new agents. *J. Med. Chem.*, 46:1081–1111, 2003.
- [124] A.K. Shiau, D. Barstad, P.M. Loria, L. Cheng, P.J. Kushner, D.A. Agard, and G.L. Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95(7):927–937, 1998.
- [125] B.L. Riggs and L.C. Hartmann. Selective estrogen-receptor modulators – Mechanisms of action and application to clinical practice. *N. Engl. J. Med.*, 348(7): 618–629, 2003.
- [126] Y. Wang, N.Y. Chirgadze, S.L. Briggs, S. Khan, E.V. Jensen, and T.P. Burris. A second binding site for hydroxytamoxifen within the coactivator-binding groove of estrogen receptor β . *Proc. Natl. Acad. Sci. USA*, 103(26):9908–9911, 2006.
- [127] B. Fisher, J.P. Costantino, D.L. Wickerham, C.K. Redmond, M. Kavanah, W.M. Cronin, V. Vogel, A. Robidoux, N. Dimitrov, J. Atkins, et al. Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project P-1 study. *J. Natl. Cancer Inst.*, 90(18):1371–1388, 1998.
- [128] A.M. Brzozowski, A.C.W. Pike, Z. Dauter, R.E. Hubbard, T. Bonn, O. Engstroem, L. Oehman, G.L. Greene, J. Gustafsson, and M. Carlquist. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, 389:753–758, 1997.
- [129] A.C.W. Pike, A.M. Brzozowski, R.E. Hubbard, T. Bonn, A.G. Thorsell, O. Engström, J. Ljunggren, J.Å. Gustafsson, and M. Carlquist. Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *EMBO J.*, 18(17):4608–4618, 1999.

- [130] M. Salomonsson, J. Häggblad, B.W. O'Malley, and G.M. Sitbon. The human estrogen receptor hormone binding domain dimerizes independently of ligand activation. *J. Steroid Biochem. Mol. Biol.*, 48:447–452, 1994.
- [131] K.E. Bergmann, C.H. Woogef, K.E. Carlson, B.S. Katzenellenbogen, and J.A. Katzenellenbogen. Bivalent ligands as probes of estrogen receptor action. *J. Steroid Biochem. Mol. Biol.*, 49:139–152, 1994.
- [132] S. Groleau, J. Nault, M. Lepage, M. Couturea, N. Dallaire, G. Berube, and R. C.-Gaudreault. Synthesis and preliminary *in vitro* cytotoxic activity of new triphenylethylene dimers. *Bioorg. Chem.*, 27(5):383–394, 1999.
- [133] G. Berube, D. Rabouin, V. Perron, B. N'Zemba, R. C. Gaudreault, S. Parent, and E. Asselin. Synthesis of unique 17β -estradiol homo-dimers, estrogen receptors binding affinity evaluation and cytotoxic activity on breast, intestinal and skin cancer cell lines. *Steroids*, 71:911–921, 2006.
- [134] A.L. LaFrate, K.E. Carlson, and J.A. Katzenellenbogen. Steroidal bivalent ligands for the estrogen receptor: Design, synthesis, characterization and binding affinities. *Bioorg. Med. Chem.*, 17:3528–3535, 2009.
- [135] A.E. Wendlandt, S.M. Yelton, D. Lou, D.S. Watt, and D.J. Noonan. Synthesis and functional analysis of novel bivalent estrogens. *Steroids*, 75:825–833, 2010.
- [136] S.Y. Dai, M.J. Chalmers, J. Bruning, K.S. Bramlett, H.E. Osborne, C. Montrose-Rafizadeh, R.J. Barr, Y. Wang, M. Wang, T.P. Burris, J.A. Dodge, and P.R. Griffin. Prediction of the tissue-specificity of selective estrogen receptor modulators by using a single biochemical method. *Proc. Natl. Acad. Sci. USA*, 105(20):7171–7176, 2008.
- [137] H.J. Kreuzer, R.L.C. Wang, and M. Grunze. Effect of stretching on the molecular conformation of oligo (ethylene oxide): A theoretical study. *New J. Phys.*, 1: 21.1–21.16, 1999.
- [138] J. H., Lee, H. B. Lee, and J. D. Andrade. Blood compatibility of polyethylene oxide surfaces. *Prog. Polym. Sci.*, 20:1043–1079, 1995.
- [139] D. Stalling, M. Westerhoff, and H.-Ch. Hege. Amira: A highly interactive system for visual data analysis. In Ch. D. Hansen and Ch. R. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.
- [140] V.M. Krishnamurthy, V. Semetey, P.J. Bracher, N. Shen, and G.M. Whitesides. Dependence of effective molarity on linker length for an intramolecular protein-ligand system. *J. Am. Chem. Soc.*, 129(5):1312–1320, 2007.

- [141] R.M. Eglén. Enzyme fragment complementation: A flexible high throughput screening assay technology. *Assay Drug Dev. Technol.*, 1(1):97–104, 2002.
- [142] V.V. Tyulmenkov and C.M. Klinge. Interaction of tetrahydrocrysene ketone with estrogen receptors α and β indicates conformational differences in the receptor subtypes. *Arch. Biochem. Biophys.*, 381(1):135–142, 2000.
- [143] W.P. van Hoorn. Identification of a second binding site in the estrogen receptor. *J. Med. Chem.*, 45(3):584–589, 2002.
- [144] W.B. Panko, C.S. Watson, and J.H. Clark. The presence of a second, specific estrogen binding site in human breast cancer. *J. Steroid Biochem.*, 14(12):1311–1316, 1981.
- [145] D.J. Kojetin, T.P. Burris, E.V. Jensen, and S.A. Khan. Implications of the binding of tamoxifen to the coactivator recognition site of the estrogen receptor. *Endocr. Relat. Cancer*, 15(4):851–870, 2008.
- [146] S.H. Kim and J.A. Katzenellenbogen. Hormone–PAMAM dendrimer conjugates: Polymer dynamics and tether structure affect ligand access to receptors. *Angew. Chem. Int. Ed.*, 45(43):7243–7248, 2006.
- [147] D. Han, F.H. Försterling, X. Li, J.R. Deschamps, H. Cao, and J.M. Cook. Determination of the stable conformation of GABA_A-benzodiazepine receptor bivalent ligands by low temperature NMR and X-ray analysis. *Bioorg. Med. Chem. Lett.*, 14(6):1465–1469, 2004.
- [148] E. J. Sorin and V. S. Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.*, 88(4):2472–2493, 2005.
- [149] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Funct. Bioinf.*, 65(3):712–725, 2006.
- [150] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.*, 120:9665–9678, 2004.
- [151] H.W. Horn, W.C. Swope, and J.W. Pitera. Characterization of the TIP4P-Ew water model: Vapor pressure and boiling point. *J. Chem. Phys.*, 123:194504, 2005.
- [152] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926, 1983.

- [153] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general Amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.
- [154] J. Wang, W. Wang, P.A. Kollman, and D.A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006.
- [155] D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, 2005.
- [156] A. Jakalian, B. L. Bush, D.B. Jack, and C.I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, 21:132–146, 2000.
- [157] A. Jakalian, D.B. Jack, and C.I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23:1623–1641, 2002.
- [158] J. Schmidt-Ehrenberg, D. Baum, and H.-Ch. Hege. Visualizing dynamic molecular conformations. In *Proceedings of IEEE Visualization 2002*, pages 235–242. IEEE Computer Society Press, 2002.
- [159] P.W.K. Rothmund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.
- [160] P.E. Nielsen, M. Egholm, R.H. Berg, O. Buchardt, et al. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*, 254(5037):1497, 1991.
- [161] C. Scheibe, A. Bujotzek, J. Dervede, M. Weber, and O. Seitz. DNA-programmed spatial screening of carbohydrate–lectin interactions. *Chem. Sci.*, 2(4):770–775, 2011.
- [162] Y.C. Lee and R.T. Lee. Carbohydrate-protein interactions: Basis of glycobiology. *Acc. Chem. Res.*, 28(8):321–327, 1995.
- [163] J.J. Lundquist and E.J. Toone. The cluster glycoside effect. *Chem. Rev.*, 102(2):555–78, 2002.
- [164] R.J. Pieters. Maximising multivalency effects in protein–carbohydrate interactions. *Org. Biomol. Chem.*, 7(10):2013–2025, 2009.

- [165] G.B. Sigal, M. Mammen, G. Dahmann, and G.M. Whitesides. Polyacrylamides bearing pendant α -sialoside groups strongly inhibit agglutination of erythrocytes by influenza virus: The strong inhibition reflects enhanced binding through cooperative polyvalent interactions. *J. Am. Chem. Soc.*, 118:3789–3800, 1996.
- [166] S.K. Choi, M. Mammen, and G.M. Whitesides. Generation and *in situ* evaluation of libraries of poly(acrylic acid) presenting sialosides as side chains as polyvalent inhibitors of influenza-mediated hemagglutination. *J. Am. Chem. Soc.*, 119(18):4103–4111, 1997.
- [167] D.D. Manning, X. Hu, P. Beck, and L.L. Kiessling. Synthesis of sulfated neoglycopolymers: Selective P-selectin inhibitors. *J. Am. Chem. Soc.*, 119(13):3161–3162, 1997.
- [168] K. Aoi, K. Itoh, and M. Okada. Globular carbohydrate macromolecules “sugar balls”. 1. Synthesis of novel sugar-persubstituted poly(amido amine) dendrimers. *Macromolecules*, 28(15):5391–5393, 1995.
- [169] D. Zanini and R. Roy. Synthesis of new α -thiosialodendrimers and their binding properties to the sialic acid specific lectin from *Limax flavus*. *J. Am. Chem. Soc.*, 119(9):2088–2095, 1997.
- [170] P.R. Ashton, E.F. Hounsell, N. Jayaraman, T.M. Nilsen, N. Spencer, J.F. Stoddart, and M. Young. Synthesis and biological evaluation of α -D-mannopyranoside-containing dendrimers. *J. Org. Chem.*, 63(10):3429–3437, 1998.
- [171] J.M. de la Fuente, A.G. Barrientos, T.C. Rojas, J. Rojo, A. Fernández, and S. Penadés. Gold glyconanoparticles as water-soluble polyvalent models to study carbohydrate interactions. *Angew. Chem.*, 113(12):2317–2321, 2001.
- [172] Y. Chen, T. Ji, and Z. Rosenzweig. Synthesis of glyconanospheres containing luminescent CdSe-ZnS quantum dots. *Nano Lett.*, 3(5):581–584, 2003.
- [173] J. Zhang, C.D. Geddes, and J.R. Lakowicz. Complexation of polysaccharide and monosaccharide with thiolate boronic acid capped on silver nanoparticle. *Anal. Biochem.*, 332(2):253–260, 2004.
- [174] K. Matsuura, M. Hibino, Y. Yamada, and K. Kobayashi. Construction of glycoclusters by self-organization of site-specifically glycosylated oligonucleotides and their cooperative amplification of lectin-recognition. *J. Am. Chem. Soc.*, 123(2):357, 2001.
- [175] K. Matsuura, M. Hibino, T. Ikeda, Y. Yamada, and K. Kobayashi. Self-organized glycoclusters along DNA: Effect of the spatial arrangement of galactoside residues on cooperative lectin recognition. *Chem. Eur. J.*, 10(2):352–359, 2004.

- [176] K. Gorska, K.T. Huang, O. Chaloin, and N. Winssinger. DNA-templated homo- and heterodimerization of peptide nucleic acid encoded oligosaccharides that mimic the carbohydrate epitope of HIV. *Angew. Chem.*, 121(41):7831–7836, 2009.
- [177] K. Turton, R. Natesh, N. Thiyagarajan, J.A. Chaddock, and K.R. Acharya. Crystal structures of *Erythrina cristagalli* lectin with bound *n*-linked oligosaccharide and lactose. *Glycobiology*, 14(10):923–929, 2004.
- [178] E.K. Fan, Z.S. Zhang, W.E. Minke, Z. Hou, C.L.M.J. Verlinde, and W.G.J. Hol. High-affinity pentavalent ligands of *Escherichia coli* heat-labile enterotoxin by modular structure-based design. *J. Am. Chem. Soc.*, 122(11):2663–2664, 2000.
- [179] P.I. Kitov, H. Shimizu, S.W. Homans, and D.R. Bundle. Optimization of tether length in nonglycosidically linked bivalent ligands that target sites 2 and 1 of a Shiga-like toxin. *J. Am. Chem. Soc.*, 125(11):3284–3294, 2003.
- [180] A.M. Riley, A.J. Laude, C.W. Taylor, and B.V.L. Potter. Dimers of D-myo-inositol 1,4,5-trisphosphate: Design, synthesis, and interaction with Ins(1,4,5)P₃ receptors. *Bioconjugate Chem.*, 15(2):278–289, 2004.
- [181] N. Schaschke, G. Matschiner, F. Zettl, U. Marquardt, A. Bergner, W. Bode, C.P. Sommerhoff, and L. Moroder. Bivalent inhibition of human β -tryptase. *Chem. Biol.*, 8(4):313–327, 2001.
- [182] V. Wittmann and S. Seeberger. Kombinatorische Festphasensynthese von multivalenten cyclischen Neoglycopeptiden. *Angew. Chem.*, 112(23):4508–4512, 2000.
- [183] V. Wittmann and S. Seeberger. Spatial screening of cyclic neoglycopeptides: Identification of polyvalent wheat-germ agglutinin ligands. *Angew. Chem. Int. Ed.*, 43(7):900–903, 2004.
- [184] Z. Zhang, J. Liu, C.L.M.J. Verlinde, GJ Wim, and E. Fan. Large cyclic peptides as cores of multivalent ligands: Application to inhibitors of receptor binding by cholera toxin. *J. Org. Chem.*, 69(22):7737–7740, 2004.
- [185] E. Uhlmann, A. Peyman, G. Breipohl, and D.W. Will. PNA: Synthetic polyamide nucleic acids with unusual binding properties. *Angew. Chem. Int. Ed.*, 37(20):2796–2823, 1998.
- [186] M. Eriksson and P.E. Nielsen. Solution structure of a peptide nucleic acid–DNA duplex. *Nat. Struct. Mol. Biol.*, 3(5):410–413, 1996.
- [187] X. Zeng, T. Murata, H. Kawagishi, T. Usui, and K. Kobayashi. Analysis of specific interactions of synthetic glycopolypeptides carrying *n*-acetyllactosamine and related compounds with lectins. *Carbohydr. Res.*, 312(4):209–217, 1998.

- [188] J.C. Sacchettini, L.G. Baum, and C.F. Brewer. Multivalent protein-carbohydrate interactions. A new paradigm for supermolecular assembly and signal transduction. *Biochemistry*, 40(10):3009–3015, 2001.
- [189] T.J. Bandy, A. Brewer, J.R. Burns, G. Marth, T.N. Nguyen, and E. Stulz. DNA as supramolecular scaffold for functional molecules: Progress in DNA nanotechnology. *Chem. Soc. Rev.*, 40(1):138–148, 2011.
- [190] R. Varghese and H.A. Wagenknecht. DNA as a supramolecular framework for the helical arrangements of chromophores: Towards photoactive DNA-based nanomaterials. *Chem. Commun.*, (19):2615–2624, 2009.
- [191] E. Braun, Y. Eichen, U. Sivan, G. Ben-Yoseph, et al. DNA-templated assembly and electrode attachment of a conducting silver wire. *Nature*, 391(6669):775–778, 1998.
- [192] K. Tanaka, G.H. Clever, Y. Takezawa, Y. Yamada, C. Kaul, M. Shionoya, and T. Carell. Programmable self-assembly of metal ions inside artificial DNA duplexes. *Nat. Nanotechnol.*, 1(3):190–194, 2006.
- [193] S.K. Silverman. DNA – eine vielseitige chemische Verbindung für die Katalyse, zur Kodierung und zur Stereokontrolle. *Angew. Chem.*, 122(40):7336–7359, 2010.
- [194] S.K. Silverman. DNA as a versatile chemical component for catalysis, encoding, and stereocontrol. *Angew. Chem. Int. Ed.*, 49(40):7180–7201, 2010.
- [195] A.P. Alivisatos, K.P. Johnsson, X. Peng, T.E. Wilson, C.J. Loweth, M.P. Bruchez, and P.G. Schultz. Organization of ‘nanocrystal molecules’ using DNA. *Nature*, 382:609–611, 1996.
- [196] F.A. Aldaye, A.L. Palmer, and H.F. Sleiman. Assembling materials with DNA as the guide. *Science*, 321(5897):1795–1799, 2008.
- [197] F. Stühmeier, A. Hillisch, R.M. Clegg, and S. Diekmann. Fluorescence energy transfer analysis of DNA structures containing several bulges and their interaction with CAP. *J. Mol. Biol.*, 302(5):1081–1100, 2000.
- [198] A. Hillisch, M. Lorenz, and S. Diekmann. Recent advances in FRET: Distance determination in protein–DNA complexes. *Curr. Opin. Struct. Biol.*, 11(2):201–207, 2001.
- [199] C.M. Niemeyer. The developments of semisynthetic DNA–protein conjugates. *Trends Biotechnol.*, 20(9):395–401, 2002.

- [200] U. Feldkamp and C.M. Niemeyer. Rationaler Entwurf von DNA-Nanoarchitekturen. *Angew. Chem.*, 118(12):1888–1910, 2006.
- [201] U. Feldkamp and C.M. Niemeyer. Rational design of DNA nanoarchitectures. *Angew. Chem. Int. Ed.*, 45(12):1856–1876, 2006.
- [202] F. Diezmann and O. Seitz. DNA-guided display of proteins and protein ligands for the interrogation of biology. *Chem. Soc. Rev.*, 40(12):5789–5801, 2011.
- [203] L. Röglin and O. Seitz. Controlling the activity of peptides and proteins with smart nucleic acid–protein hybrids. *Org. Biomol. Chem.*, 6(21):3881–3887, 2008.
- [204] Z.L. Pianowski and N. Winssinger. Nucleic acid encoding to program self-assembly in chemical biology. *Chem. Soc. Rev.*, 37(7):1330–1336, 2008.
- [205] K. Gorska, J. Beyrath, S. Fournel, and N. Winssinger. Ligand dimerization programmed by hybridization to study multimeric ligand–receptor interactions. *Chem. Commun.*, 46(41):7742–7744, 2010.
- [206] E. Protozanova, P. Yakovchuk, and M.D. Frank-Kamenetskii. Stacked–unstacked equilibrium at the nick site of DNA. *J. Mol. Biol.*, 342(3):775–785, 2004.
- [207] C. Chang, J.D. Norris, H. Grøn, L.A. Paige, P.T. Hamilton, D.J. Kenan, D. Fowlkes, and D.P. McDonnell. Dissection of the LXXLL nuclear receptor-coactivator interaction motif using combinatorial peptide libraries: Discovery of peptide antagonists of estrogen receptors α and β . *Mol. Cell. Biol.*, 19(12):8226–8239, 1999.
- [208] R. Holliday. A mechanism for gene conversion in fungi. *Genet. Res.*, 5(2):282–304, 1964.
- [209] E.S. Andersen, M. Dong, M.M. Nielsen, K. Jahn, R. Subramani, W. Mamdouh, M.M. Golas, B. Sander, H. Stark, C.L.P. Oliveira, et al. Self-assembly of a nanoscale DNA box with a controllable lid. *Nature*, 459(7243):73–76, 2009.
- [210] H. Xia, Q. Mao, H.L. Paulson, and B.L. Davidson. siRNA-mediated gene silencing *in vitro* and *in vivo*. *Nat. Biotechnol.*, 20(10):1006–1010, 2002.
- [211] V. Durmaz, M. Weber, and R. Becker. How to simulate affinities for host–guest systems lacking binding mode information: Application to the liquid chromatographic separation of hexabromocyclododecane stereoisomers. *J. Mol. Model.*, pages 1–10, 2012.
- [212] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 108(25):10184–10189, 2011.

- [213] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 18(2):154–162, 2008.
- [214] L. Celik, B. Schiøtt, and E. Tajkhorshid. Substrate binding and formation of an occluded state in the leucine transporter. *Biophys. J.*, 94(5):1600–1612, 2008.
- [215] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.
- [216] F. Khalili-Araghi, J. Gumbart, P.C. Wen, M. Sotomayor, E. Tajkhorshid, and K. Schulten. Molecular dynamics simulations of membrane channels and transporters. *Curr. Opin. Struct. Biol.*, 19(2):128–137, 2009.
- [217] H. Lu and K. Schulten. Steered molecular dynamics simulation of conformational changes of immunoglobulin domain I27 interpret atomic force microscopy observations. *Chem. Phys.*, 247(1):141–153, 1999.
- [218] A. Nielsen. Von Femtosekunden zu Minuten – ein verallgemeinerter Operatoransatz in der Molekülsimulation. Master’s thesis, Freie Universität Berlin, Fachbereich Mathematik und Informatik, 2012.
- [219] A.W.S. da Silva and W.F. Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes*, 5(1):367, 2012.
- [220] C. Caleman, P.J. van Maaren, M. Hong, J.S. Hub, L.T. Costa, and D. van der Spoel. Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *J. Chem. Theory Comput.*, 8(1):61–74, 2011.
- [221] D. van der Spoel, P.J. van Maaren, and C. Caleman. GROMACS molecule & liquid database. *Bioinformatics*, 28(5):752–753, 2012.
- [222] T.A. Halgren. Merck Molecular Force Field: I-V. *J. Comput. Chem.*, 17(5-6):490–641, 1996.
- [223] A. Bujotzek and M. Weber. Efficient simulation of ligand-receptor binding processes using the conformation dynamics approach. *J. Bioinf. Comput. Biol.*, 7(05):811–831, 2009.
- [224] P.F. Dubois, K. Hinsin, and J. Hugunin. Numerical Python. *Computers Phys.*, 10(3), 1996.
- [225] P.F. Dubois. Extending Python with Fortran. *Comput. Sci. Eng.*, 1(5):66–73, 1999.

-
- [226] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL <http://www.scipy.org/>.
- [227] Ch. Schütte. *Conformational dynamics: Modelling, theory, algorithm and application to biomolecules*. Habilitation thesis, Freie Universität Berlin, 1999. Department of Mathematics and Computer Science.
- [228] J.D. Chodera, W.C. Swope, J.W. Pitera, and K.A. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, 5(4):1214, 2006.
- [229] W.F. Van Gunsteren and H.J.C. Berendsen. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.*, 1(3):173–185, 1988.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Anwendung von Methoden der klassischen Molekülsimulation auf multivalente Ligand-Rezeptor Systeme. Die Fragestellungen des angewandten Teils der Arbeit drehen sich erstens um das rationale Design von molekularen Abstandhalter- und Gerüststrukturen, die sich für eine multivalente Präsentation von Liganden für einen gegebenen (typischerweise biologischen) Rezeptor eignen, und zweitens um die Mechanismen, die einen multivalenten von einem monovalenten Bindungsprozess abgrenzen, etwa im Hinblick auf die für multivalente Systeme realisierbare Chelatkooperativität. Anhand der Analyse der Simulationsdaten und das in Verhältnis setzen der theoretischen Ergebnisse mit von Kooperationspartnern bereitgestellten experimentellen Daten können gewisse Grundlagen für das Design von molekularen Abstandhalter- und Gerüststrukturen erarbeitet werden, etwa im Hinblick auf den Einsatz von flexiblen und rigiden Elementen im modularen Aufbau, sowie lösungsmittelabhängige Faltungseffekte. Die Analyse des Bindungsprozesses eines synthetischen bivalenten Wirt-Gast-Systems im Hinblick auf metastabile Zustände und Übergangswahrscheinlichkeiten ergibt des Weiteren deutliche Hinweise auf das Vorhandensein eines kooperativen Effekts im Zusammenhang mit der intramolekularen Bindung bei der Bildung des zyklischen Komplexes. Der angewandte Teil der Arbeit wird ergänzt durch einen theoretischen Teil, der sich erstens mit den chemischen Hintergründen der in multivalenten Systemen wirkenden Effekten auseinandersetzt, und zweitens die Grundlagen der klassischen Molekülsimulation zusammenfasst. Der theoretische Teil der Arbeit wird vervollständigt durch eine Charakterisierung des ZIBgridfree-Sampling-Algorithmus, der im Rahmen dieser Arbeit erweitert und neu implementiert wurde, um die Simulation des bivalenten Bindungsprozesses unter Berücksichtigung des Lösungsmittels zu ermöglichen. Eine Validierung der Methode anhand der Konformationsanalyse zweier kleiner Moleküle in Vakuum und in explizit modelliertem Wasser ist im Anhang der Arbeit zu finden.