Freie Universität Berlin

Conformational entropy from molecular simulations: statistical mechanics using the tools of information theory

Dissertation zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von

Jorge Numata

aus Mexiko-Stadt

Berlin, Januar 2012

Die vorliegende Arbeit wurde unter Anleitung von Prof. Dr. E. W. Knapp im Zeitraum 09.2005-12.2011 am Institut für Chemie / Physikalische und Theoretische Chemie der Freien Universität Berlin im Fachbereich Biologie, Chemie und Pharmazie durchgeführt.

1. Gutachter: Prof. Dr. Ernst-Walter Knapp,
   Freie Universität Berlin
2. Gutachter: Priv.-Doz. Dr. Martin Falcke,
   Max Delbrück Center for Molecular Medicine

Tag der Disputation: 10.04.2012

Dedicatoria

Esta tesis doctoral está dedicada a mi padre, Jorge Armando Numata. Le agradezco mucho haberme dado el aliento y los recursos para seguir mi camino intelectual, sin ser él mismo un científico. Tres ejemplos son: haberme conseguido una Commodore 64, haberme enseñado el sistema binario de niño, y haberme brindado la oportunidad de estudiar Ingeniería Química en el ITESM.

Esta tesis también está dedicada a Raquel Bronsoler, mi maestra de ciencias y matemáticas de niño y adolescente. Le agradezco haber nutrido mi curiosidad de forma tan agradable.

This thesis summarizes my doctoral research work. It is mainly based on two peer-reviewed papers:

Numata, J.; Knapp, E. W., *Balanced and bias-corrected computation of conformational entropy differences for molecular trajectories*. J. Chem. Theory Comput. **2012**, 8(4), 1235-1245. Submitted on 20-dec-2011. Accepted on 14-mar-2012. http://dx.doi.org/doi:10.1021/ct200910z
Presented in detail on Chapter 2 of this thesis.

Numata, J.; Juneja, A.; Diestler, D. J.; Knapp, E. W., *Influence of Spacer-Receptor Interactions on the Stability of Bivalent Ligand-Receptor Complexes*. J. Phys. Chem. B **2012**, *116*(8), 2595-2604. Submitted on 23-nov-2011. Accepted on 25-jan-2012. http://dx.doi.org/10.1021/jp211383s
Introduced in Chapter 3 of this thesis.

During the same period I participated in related research projects that led to the following publications:

Numata, J.; Wan, M.; Knapp, E. W., *Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation*. Genome Informatics **2007**, *18*, 192-205.
Cited* 11 times[1-11].

Numata, J.; Ebenhöh, O.; Knapp, E.-W., *Measuring correlations in metabolomic networks with mutual information*. Genome Informatics **2008**, *20*, 112-122.
Cited* 4 times[12-15].

Numata, J., *Entropy and thermodynamics in biomolecular simulation*. In *Handbook of Research on Systems Biology Applications in Medicine*; **2008**. ISBN: 978-1605660769

Salwiczek, M.; Samsonov, S.; Vagt, T.; Nyakatura, E.; Fleige, E.; Numata, J.; Cölfen, H.; Pisabarro, M. T.; Koksch, B., *Position dependent effects of fluorinated amino acids on hydrophobic core formation of a coiled coil heterodimer*. Chem. Eur. J. **2009**, *15*, 7628-7636.
Cited 10 times*.

Juneja, A.; Numata, J.; Nilsson, L.; Knapp, E.-W., *Merging Implicit with Explicit Solvent Simulations: Polyethylene Glycol*. J. Chem. Theory Comput. **2010**, *6*, 1871-1883.
Cited 1 time*.

*citations on ISI Web of Knowledge and Google Scholar, excluding self-citations, until Jan. 2012.

Acknowledgements

Ernst-Walter Knapp for supervising my PhD work, allowing me the liberty of choosing my own project topic and providing input from his wealth of knowledge on molecular simulation and statistical mechanics.

Support from the collaborative research center SFB 765 “Multivalency as chemical organization and action principle” Project C1 and the IRTG “Genomics and Systems Biology of Molecular Networks”, both from the Deutsche Forschungsgemeinschaft (DFG), are gratefully acknowledged.

All the members of the Knapp group, who create a cooperative and friendly working environment.

Arieh Ben-Naim for stimulating discussions; and for sharing the passion for understanding and communicating the meaning of thermodynamic entropy as a particular case of Shannon’s missing information for molecular microstates.

Lev Levitin (Boston University) for giving me a formal foundation on information theory, and for his warm-hearted and intelligent presence.

Dennis J. Diestler for being so generous with his knowledge and being always available to shine some light on otherwise opaque equations during his stays in Berlin.

José Luis López (JoLuLo) for waking my passion for simulation and numerical methods as a student of Chemical Engineering. Burkhard Schmidt and Giulia Morra for guiding my first steps in molecular dynamics.

Vladimir Hnizdo and Alexander Kraskov for helpful e-mail correspondence. Riccardo Baron for writing a clear and useful PhD Thesis. Michael Gilson for releasing his code as open source.

Arturo Robertazzi, Kay Wishöth and Alexander Berzin for proofreading the manuscript. Isabel Arnaud for the artistic rendering of Fig. 1.1.

Statutory Declaration

I hereby testify that this thesis is the result of my own work and research, except for any explicitly referenced material, whose source can be found in the bibliography. This work contains material that is the copyright property of others which cannot be reproduced without the permission of the copyright owner.

Jorge Numata

List of Abbreviations

MD, Molecular dynamics

MC, Monte Carlo

MI expansion, mutual information expansion

MIEn, n^{th} order MI expansion

BAT, bond-angle-torsion

NMR, nuclear magnetic resonance

PCA, principal component analysis

QHA, quasi-harmonic approximation

RW, random walk

cGMP, cyclic guanidine mono-phosphate

ER, estrogen receptor

PEG, polyethylene glycol;

RET, nucleotide-gated ion channel in bovine rod photoreceptor cells;

SASA, solvent accessible surface area

Papers referencing my publications

(1) Okimoto, N.; Futatsugi, N.; Fuji, H.; Suenaga, A.; Morimoto, G.; Yanai, R.; Ohno, Y.; Narumi, T.; Taiji, M., *High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations.* PLoS Comput. Biol. **2009**, *5*, e1000528. http://dx.doi.org/doi:10.1371/journal.pcbi.1000528

(2) Baron, R.; Hünenberger, P. H.; McCammon, J. A., *Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties.* J. Chem. Theory Comput. **2009**, *5*, 3150-3160. http://dx.doi.org/doi:10.1021/ct900373z

(3) Sawada, T.; Fedorov, D. G.; Kitaura, K., *Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations.* J. Am. Chem. Soc. **2010**, *132*, 16862-16872. http://dx.doi.org/doi:10.1021/ja105051e

(4) Woolf, N. J.; Priel, A.; Tuszynski, J. A., *Novel Modes of Neural Computation: From Nanowires to Mind.* In *Nanoneuroscience* **2010**, p 227-273. doi:10.1007/978-3-642-03584-5_6

(5) Watanabe, H.; Tanaka, S.; Okimoto, N.; Hasegawa, A.; Taiji, M.; Tanida, Y.; Mitsui, T.; Katsuyama, M.; Fujitani, H., *Comparison of binding affinity evaluations for FKBP ligands with state-of-the-art computational methods: FMO, QM/MM, MM-PB/SA and MP-CAFEE approaches.* Chem-Bio Inf. J. **2010**, *10*, 32-45. http://dx.doi.org/doi:10.1273/cbij.10.32

(6) Suárez, E.; Díaz, N.; Suárez, D., *Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations.* J. Chem. Theory Comput. **2011**, *7*, 2638-2653. http://dx.doi.org/doi:10.1021/ct200216n

(7) Yanga, S.-Y.; Yanga, X.-L.; Yao, L.-F.; Wang, H.-B.; Sun, C.-K., *Effect of CpG methylation on DNA binding protein: Molecular dynamics simulations of the homeodomain PITX2 bound to the methylated DNA.* J. of Mol. Graph. Model. **2011**, *29*, 920-927. http://dx.doi.org/doi:10.1016/j.jmgm.2011.03.003

(8) Mukherjee, A., *Entropy Balance in the Intercalation Process of an Anti-Cancer Drug Daunomycin.* J. Phys. Chem. Lett. **2011**, *2*, 3021-3026. http://dx.doi.org/doi:10.1021/jz2013566

(9) Liu, F.-F.; Dong, X.; Dong, X.-Y.; He, L.; Middelberg, A. P. J.; Sun, Y., *Molecular Insight into Conformational Transition of Amyloid beta-Peptide 42 Inhibited by (-)-Epigallocatechin-3-gallate Probed by Molecular Simulations.* J. Phys. Chem. B **2011**, *116*, 11879-11887. http://dx.doi.org/10.1021/jp202640b

(10) Polyansky, A. A.; Zubac, R.; Zagrovic, B., *Estimation of Conformational Entropy in Protein–Ligand Interactions: A Computational Perspective.* Meth. Mol. Biol. **2012**, *819*, 327-353. http://dx.doi.org/doi:10.1007/978-1-61779-465-0_21

(11) Homeyer, N.; Gohlke, H., *Free Energy Calculations by the Molecular Mechanics Poisson–Boltzmann Surface Area Method.* Mol. Inf. **2012**, *In Print.* http://dx.doi.org/doi:10.1002/minf.201100135

(12) Mrabet, Y.; Semmar, N., *Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts.* Curr. Drug Metab. **2011**, *11*, 315-341. http://dx.doi.org/doi:10.2174/138920010791514333

(13) Subramaniam, S.; Fahy, E.; Gupta, S.; Sud, M.; Byrnes, R. W.; Cotter, D.; Dinasarapu, A. R.; Maurya, M. R., *Bioinformatics and systems biology of the lipidome.* Chem. Rev. **2011**, *111*, 6452-6490. http://dx.doi.org/doi:10.1021/cr200295k

(14) Koo, I.; Zhang, X.; Kim, S., *Comparison of Spectral Similarity Measures for Compound Identification.* Proc. Bioinf. and Biomed. Eng. **2011**, 1-4. http://dx.doi.org/doi:10.1109/icbbe.2011.5780011

(15) Wexler, E. M.; Rosen, E.; Lu, D.; Osborn, G. E.; Martin, E.; Raybould, H.; Geschwind, D. H., *Genome-Wide Analysis of a Wnt1-Regulated Transcriptional Network Implicates Neurodegenerative Pathways.* Science Signaling **2011**, *4*, ra65. http://dx.doi.org/doi:10.1126/scisignal.2002282

Table of contents

This page was intentionally left blank

1 Introduction

1.1 Biological thermodynamics

We are far away from being able to understand, model and simulate biological processes from first principles at all scales of detail. Nevertheless, it is firmly established that general physical principles are applicable to living matter. Among them, thermodynamics seems to me the most useful and general principle. Thermodynamics determines for instance the directions of reactions inside a cell and the amount of energy that is stored and transferred to synthesize a given metabolite[1]. Thermodynamics predicts the direction of spontaneous processes, such as protein association events, and the extent of biochemical reactions. It quantifies equilibrium, phase changes and stability using unmeasurable quantities like energy and entropy. These are coupled to experimentally measurable ones, like temperature and pressure, through mathematical relationships. This way, thermodynamics creates a system of explanation for physicochemical transformations in micro- and macromolecular systems.

> "[Biology has] become the paramount science, exceeding other disciplines, including physics and chemistry at least, in the creative tumult of its disputations. […] I'll also be so bold at this point to suggest that we are now at the edge of establishing the two fundamental laws of biology:
>
> The first law is that all of the phenomena of biology, the entities and the processes, are ultimately obedient to the laws of physics and chemistry. Not immediately reducible to them, but ultimately consistent and in consilience with them, by a cause and effect explanation.
>
> The second law is that all biological phenomena, these entities and processes that define life itself, have arisen by evolution through natural selection."
>
> E.O. Wilson, speaking at the 50th anniversary of New Scientist magazine, 2006

The concept of free energy is the main criterion to predict if, and to what extent, a process will occur in a spontaneous way. It is a refinement of the qualitative idea of "chemical affinity", widespread until the 19th century. Free energy allows us to describe the equilibrium in chemical reactions and physicochemically driven processes such as non-covalent association. Important processes governed by non-covalent interactions are hormone binding to receptors, mRNA codon recognition by the ribosome[2] and protein-protein interactions. Free energy allows us to predict

the strength of such non-covalent interactions and the corresponding equilibrium constants. Estimation of free energy in protein folding and molecular recognition is one of the central tasks of theoretical chemistry concerning macromolecules and the subject of many reviews[3-8]. Free energy can be loosely described as the interplay and competition between energy and entropy.

1.2 An intuitive notion of energy and entropy

Energy quantifies the ability to do work. Entropy measures the quality of that energy; the lower its entropy, the more useful that energy is.

Fig. 1.1: The Sun as the source of lowered entropy for planet Earth. Figure courtesy of Isabel Arnaud.

At first sight, the Earth seems to be kept alive by the energy arriving from the Sun. This is a superficial understanding, because in the steady state, the amount of energy arriving from the Sun and the amount radiated back into space are equal. If the energy arriving from the sun remained, the Earth would become unbearably warmer every day. As noted by Schrödinger[9] in an article directed to a lay audience, life is maintained by a constant influx of low entropy. He coined the term *negentropy*, which in this context means that living organisms are constantly expelling high entropy and feed on nurturing low entropy to survive. Plants use the low entropy radiation through photosynthesis to lower their own entropy. Animals eat these plants, for the same

purpose. With each step of metabolism in single organisms, and each trophic level of an ecosystem, the total entropy increases in irreversible processes. In other words, the "quality" of the energy is lowered along the food chain.

The earth receives low-entropy electromagnetic radiation, which partly trickles down through the food chain and metabolic networks, and is ultimately emitted back as high-entropy radiation. High-frequency, visible light arrives to the earth. Infrared, low frequency radiation is emitted back into space (see Fig. 1.1). Consider the proportionality between frequency and energy, known as Planck's relation

$$E = h\nu\,, \tag{1.0}$$

where E is energy, h is Planck's constant and ν is the frequency. The arriving high frequency photons carry more energy per photon than those leaving. To keep the balance of energy in the steady state, more photons leave than those that arrive. A larger number of photons means more degrees of freedom, and thus higher entropy. For more on this, see Chap. 27 of ref [10].

Entropy involves energy dissipation and the irreversibility of processes. This can be illustrated with a waterfall analogy. A small amount of water at the top of a mountain falls into the ocean, which is flattened and all at the same level. In this process, its energy is not lost - it just becomes more dispersed. As it falls down, it may or may not be used to drive an industrial or biological process.

Entropy is in many relevant cases a measure of *disorder* and *uncertainty*. Understanding entropy as *disorder* should be taken with a grain of salt, as this analogy does not always apply[11]. A crystal is an example of a low entropy material because of its predictable regularity. This is not to say that particles in a crystal are static at finite temperature, but their displacements due to thermal energy are relatively small. If we take a crystal with particles vibrating around fixed lattice points and heat it, it will become a liquid and its entropy will rise. If we heat the system further, it may become a gas that fills the whole room. The entropy (*uncertainty*) is now much larger than in the crystal. It has become much harder to say where in the room each particle is, that is, the *missing information*

about the positions has grown. The crystal will not spontaneously reform, putting everything back into place the way it was before. This irreversibility results in an arrow of time that points just in one direction into the more *disordered* future[12].

Entropy is a measure of the missing information about the possible arrangements of a system. About one century after entropy was postulated in thermodynamics[13], entropy was formulated again in the context of telecommunications by Shannon[14] to provide a measure for channel transmission capacity. Shannon's ideas went on to become the foundation of information theory, a whole branch of applied mathematics closely related to statistics. Information theory turned out to be more general, with thermodynamic entropy being a particular case thereof[15]. In the present doctoral thesis, I use tools from information theory and apply them to statistical thermodynamics.

Finally, entropy is a measure of *multiplicity* and *variability* within a system. It comprises *counting states* on a logarithmic scale. An intuitive connection between the *quality of energy* understanding of entropy and the *multiplicity of states* view from information theory can be gained through the following example: Consider how we rub our hands together on a cold day. We use high quality energy gained from food to apply very directed work, which is a collective effort of many muscle cells applying a force in the same direction (low *multiplicity*). It gets transformed into low quality energy that we perceive as a rise in temperature. This transformed energy has a high *multiplicity* because it quickly becomes spread out in all directions and involves the random, undirected vibrations of many particles. It is of lower quality because it cannot be completely turned back into directed motion, as dictated by the Second Law of Thermodynamics.

If entropy is always increasing in real-world processes, it is legitimate to ask: why was it so low in the first place? Why was the entropy of the universe so low after the Big Bang? A possible partial answer has been postulated by Roger Penrose in Conformal Cyclic Cosmology[16], where the universe exists in cycles of time that reset the entropy through rescaling[17].

Today, the concept of entropy has found widespread application in science and engineering. The generality of thermodynamics has afforded it a place in engineering[18], astrophysics[17,19] and of

course the life sciences[20]. Entropy also lives a parallel life in statistics[21] and information theory[14,22], where it is applied to quantify information in processes as varied as communication channels and cell signaling dynamics[23,24]. However, entropy is still often misunderstood, ignored or pointed to as the cause of inexplicable results.

1.3 From steam engines to actin filaments: the laws of thermodynamics

Laws of Thermodynamics in Lay Terminology
1st Law: It is impossible to obtain something from nothing, but one may break even.
2nd Law: One may break even but only at the lowest possible temperature.
3rd Law: One cannot reach the lowest possible temperature.
Implication: It is impossible to obtain something from nothing, so one must optimize resources
- Annamalai & Puri

Advanced thermodynamics engineering; CRC Press, 2002.

Although thermodynamics was born in the realm of industrial plants[25], its wide applicability has conferred it a place in biology. An example is the exploration of the thermodynamics of actin filaments[26], which are self-assembling units that play key roles in muscle contraction and in the formation and reshaping of the cytoskeleton. Actin filaments achieve cell motility by exploiting entropic forces[27-30].

Thermodynamics consists of a set of tools to reason about energies and entropies. The basic building blocks are two laws and some multivariate calculus[31].

1st law (energy balance):

$$dU = \delta q + \delta w, \tag{1.1}$$

where U is the internal energy, q is heat and w is work. d indicates U is a path-independent state variable, while δ means that heat and work depend on the application path.

2nd law (total entropy never decreases):

$$dS \geq 0, \tag{1.2}$$

where S is entropy[13,32].

The First Law simply states that energy in all forms is conserved, and that it can be exchanged through heat and work. The Second Law can be seen as "half a conservation law", because entropy can be created but not destroyed[33].

By combining the First and Second Laws at constant number of particles (N), Volume (V) and Temperature (T), we may obtain an expression describing how a system can reach equilibrium with the inequality

Free energy differential: $$dF = dU - TdS \leq 0 \qquad (1.3)$$

1.4 Free energy in processes involving proteins

Protein folding and receptor-ligand binding occur in a spontaneous and specific way when the folded and bound states have a lower free energy than their denatured and unbound counterparts, respectively. The Helmholtz free energy change ΔF or the Gibbs free energy change $\Delta G = \Delta F + P\Delta V$ predict the equilibrium constant (K_{eq}) for folding and binding. For macromolecules solvated in incompressible fluids like water, the volume term $P\Delta V$ is negligible, so

$$\Delta G \approx \Delta F = \Delta U - T\Delta S = -k_B T \ln K_{eq} . \qquad (1.4)$$

For macromolecules and soft matter in general, the understanding of the driving forces that together result in a given free energy or binding constant requires consideration of flexibility and dynamics. The amino acid sequence of proteins encodes structure, flexibility[34], thermodynamics[35] and dynamics[36-39], which in turn code function.

The search for minimum free energy balances a rise in total entropy and a fall in total energy. A rise in entropy need not happen uniformly for all the components of a system. Maximizing entropy in a subset of the system may be the driving force for organization in another subset. This is indeed the case in protein folding, where the hydrophobic effect often maximizes the entropy the water by collapsing the protein chain into a more orderly, low conformational entropy state[40].

The hydrophobic effect at room temperature for small solutes is also primarily due to a maximization of water entropy[41]. Nevertheless, the hydrophobic effect can also have an enthalpic origin, for instance at high temperatures[42] or in the association of ligands to protein cavities[43-45].

Fig. 1.2: Proteins are molecular machines with motion networks[39] that catalyze reactions. They obey the laws of thermodynamics, as first laid out for steam engines. Photo of a sculpture taken by myself at Tacheles, Berlin in 2006.

1.4.1 Predicting ligand binding and protein-protein interactions

The difference in free energy between two states tells us if a process will occur spontaneously and to what extent. For a thermodynamic state function such as free energy to be meaningful, the start and end states should be clearly defined. For example, the stability of a protein against unfolding is given by $\Delta G_{fold} = G_{folded} - G_{denatured}$. If ΔG_{fold} is negative, thermodynamics will favor the folded state. Similarly, the binding free energy for a ligand-receptor complex is:

$$\Delta G_{binding} = G_{complex} - \left(G_{ligand} + G_{protein}\right) = -RT \ln K_a \qquad (1.5)$$

A factor-of-ten increase in the binding affinity constant K_a translates into a change of 1.3 kcal/mol in $\Delta G_{binding}$ at room temperature. This additional stability can come from either enthalpic or entropic contributions within the whole system (ligand, receptor protein, solvent, ions, etc)[46].

1.4.2 Conformational entropy

The net enthalpic (ΔH) and entropic ($T\Delta S$) contributions from all particles (solute and solvent) almost even out in natural folding or properly engineered proteins[47]. Stability of proteins against denaturing is typically[48] around ΔG_{fold} = 5 to 15 kcal/mol (K_{eq} = 10^{-4} to 10^{-11}). Upon folding, the protein becomes more rigid and loses conformational entropy. This unfavorable contribution is typically $T\Delta S_{conf}$ = 10 to 100 kcal/mol. Any estimation of free energy lacking this contribution will grossly overestimate the stability of proteins against unfolding.

Entropy is defined in terms of probabilities of to occupy specific microstates. A microstate is an individual conformation of a molecule. The conformational entropy is

$$S_{conf} = -k_B \sum p_i \ln p_i \,, \tag{1.6}$$

where k_B is the Boltzmann constant and p_i is the probability of occupancy of each microstate. The p_i represent the net probabilities of occupancy of given microstates, including all correlations and statistical dependencies connected with it. Energetic interactions between particles give rise to correlations. It is known from information theory that neglecting correlations will cause an overestimation of entropy[15], which is the explanation for the famous difference between the Boltzmann and Gibbs entropies[49]. Such statistical correlations may even manifest physically to produce work[50]. We are nowadays certain about the existence of particles and the need for statistics to count the multiple ways in which they arrange. However, entropy was originally defined as a macroscopic quantity[13], without any reference to particles or statistics.

1.5 The statistical in mechanics

> Max Planck joined the Physical Society of Berlin, of which he wrote: "In those days I was essentially the only theoretical physicist there, whence things were not so easy for me, because I started mentioning entropy, but this was not quite fashionable, since it was regarded as a mathematical spook". Max Planck
>
> Elektrotechnischer Verein Berlin, G. *ETZ: Elektrotechnische Zeitschrift (VDE-Verlag) 69 (A)*, 1948.

Statistics deals with uncertainty and probabilities. This conjures for many the defeat of determinism. For this reason, statistics and the microscopic understanding of entropy had a difficult entry into science. Ludwig Boltzmann lived in the late 1800's and formulated a statistical approach[32,51] which took into account the stochastic nature of microscopic processes in which sharply defined macroscopic physical values become distributions[52]. Sadly, Boltzmann committed suicide before seeing the success of his theory. Albert Einstein's published PhD dissertation[53] deals with deterministic equations. This seems to have been a compromise, as his advisor Alfred Kleiner would not accept his molecular kinetic treatment of fluids, allegedly because of its statistical nature[54]. The hypothesis that Einstein originally intended to write a dissertation on statistical mechanics is supported by the fact that during the previous years, he had been publishing papers about entropy and thermodynamics with a strong statistical component[55-57].

For years, Planck also had upheld a macroscopic view of entropy and matter as a continuum. But with his solution of the blackbody radiation problem, Planck was not only introducing the *Wirkungsquantum*, but at the same time recognizing the need for a statistical treatment[58]. "I was, however, at that time still too far oriented towards the phenomenological aspect to come to closer quarters with the connection between entropy and probability […] I busied myself… with the task of elucidating a true physical character for the [entropy] formula, and this problem led me automatically to a consideration of the connection between entropy and probability, that is, Boltzmann's trend of ideas," said Planck in his Nobel prize lecture[59].

In the words of Jaynes, "It is possible to make a sharp distinction in statistical mechanics: the physical and the statistical. We formulate our partial knowledge into a physical model. This

model should deliver a correct enumeration of the states of a system and their properties. The statistical part is a straightforward example of inference."[60] The field of statistical physics[61,62] explains how the world we see around us arises from the interactions of uncountable numbers of microscopic particles. The observed fact that entropy always increases (2nd law, eq (1.2)) is in fact just a consequence of probability. A system that starts in a low-entropy state has many ways to move to a state of higher entropy, but only a few ways to move to a state with the same or lower entropy. Thus, you are more likely to see a system move from low to high entropy than vice versa, and when we consider macroscopic objects involving ~10^{23} particles, the probability of seeing entropy spontaneously decrease "quickly moves into monkeys-writing-Shakespeare territory"[63].

Arieh Ben-Naim argues that entropy can be reduced to plain common sense[64], as:
"1. The Second Law is basically a law of probability [as Boltzmann established].
2. The laws of probability are basically the laws of common sense [as Laplace said].
3. It follows from (1) and (2) that the Second Law is basically a law of common sense - nothing more."

1.5.1 Information over matter and energy

Recently, an experiment demonstrated the realization of a Szilárd-type[65-67] machine that transforms information into free energy. A non-equilibrium feedback manipulation of a Brownian particle on the basis of information about its position (configurational entropy) achieves this conversion[68]. Theoretical demonstrations have also shown how to convert statistical correlations into work[50]. Even more generally, both relativity theory and quantum mechanics can be understood in terms of information[69-71]. Some things can travel faster than light. For instance, quantum entangled states experimentally show *spooky action* at a distance[72,73]. But information cannot travel faster than light[71] (if Einstein's theory holds). This is confirmed by the fact that it is not possible to transmit information instantly using entangled states[71]. In my opinion, this shows that information has a higher hierarchy than matter and energy themselves.

1.6 Entropy is a logarithmic counting of microstates

A microstate is an individual conformation of a molecule that cannot be resolved within the framework of a conventional experimental setup. Entropy measures the multiplicity of such microstates on a logarithmic scale. In contrast, a macrostate[61,74] is a collection of microstates with given macroscopic properties (values of a small number of observables in a given[49] experiment).

According to the Boltzmann distribution, a microstate of lower energy will be more populated than one of higher energy in the canonical ensemble. Expressing the Boltzmann distribution in terms of the probability of individual microstates j and k yields:

$$\frac{p_j}{p_k} = \exp\left(\frac{-(E_j - E_k)}{k_B T}\right) \qquad \text{for T>0K .} \tag{1.7}$$

E_j is a microscopic energy, which is a function of the conformation *(j)*. *E* is a function of the number of moles and volume (*N*, *V*) but not of temperature or entropy (*T*, *S*). Higher temperatures mean that thermal energy $k_B T$ will allow significant occupation of correspondingly higher energy levels. The macrostate of a system acquires an internal energy *U* as a function of which microstates are actually populated. The macroscopic (average) internal energy *U* is the average over all microstates

$$U = \langle E \rangle = \sum_{i=1}^{t} p_i E_i \,, \tag{1.8}$$

where t is the total number of microstates.

The toy model in Fig. 1.2 captures many interesting properties of energy and entropy. This molecule has positively and negatively charged ends, which attract each other and may interact with a favorable energy -ε. We further assume that we have an experimental technique that can resolve whether the molecule is in the open or the compact conformation. Since the compact conformation has the lowest energy, one could reach the conclusion that it is the most populated. But what if the microstates of the open type are more numerous? This multiplicity (*W*) of states

of equal energy is quantified by the entropy, eq (1.6). For our model:

$S_o = -k_B \sum_{i=1}^{4} p_{o,i} \ln p_{o,i} = k_B \ln(4)$ and $S_c = k_B \ln(1)$. Since all microstates for each macrostate have the same energy, they are equiprobable, and we can simplify the entropy to $S = k_B \ln(W)$. The entropy change between macrostates is $\Delta S_{oc} = k_B \ln(4/1)$.

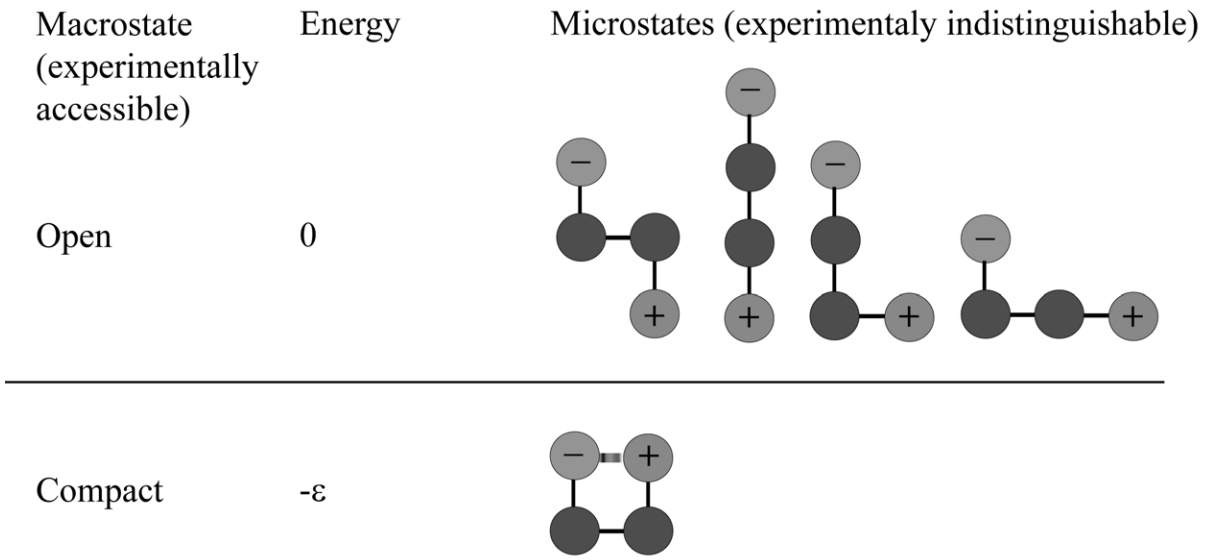

Fig. 1.2: A 4-bead toy model of a folding molecule. The compact state is unique and contains one attractive interaction with a favorable energy of –ε. The four open microstates have no long-range interactions. Based on Fig 10.1 of ref [31].

The free energy uses both U and S to assess the stability of each macrostate. The relative stability in the canonical ensemble between the open and compact macrostates is measured by the free energy, eq (1.4). The opening ("unfolding") free energy for Fig. 1.2 is:

$$\Delta F_{oc} = (0 - (-\varepsilon)) - T(k_B \ln 4 - k_B \ln 1) = \varepsilon - k_B T \ln 4 \tag{1.9}$$

The equilibrium constant K_{eq} yields the proportion of compact vs. open conformations:

$$K_{eq} = \exp\left(\frac{-\Delta F_{oc}}{k_B T}\right) = \exp\left(\frac{-\Delta U_{oc}}{k_B T}\right)\exp\left(\frac{T\Delta S_{oc}}{k_B T}\right) \tag{1.10}$$

K_{eq} for the model in Fig. 1.2 is:

$$K_{eq} = \exp\left(\frac{-\varepsilon}{k_B T}\right)\exp\left(\frac{k_B T \ln(4)}{k_B T}\right) = 4 \cdot \exp\left(\frac{-\varepsilon}{k_B T}\right) \tag{1.11}$$

The entropy change cannot be interpreted straightforwardly when expressed as ΔS = 2.75 cal/(mol K). But its contribution to the equilibrium constant is a dimensionless number measuring multiplicity W = 4. This means simply 4 times more open conformations than compact conformations. In this light, entropy is more intuitive and easier to understand than energy itself!

Entropy can be interpreted in an analog way in protein folding. A conformational unfolding entropy change[75] of ΔS_{unfold} = 33.3 cal/(mol K) means there exist $W_{denatured} / W_{native}$ = $\exp(T\Delta S/(k_B T))$ = 1.9 × 10^7 times more denatured conformations than native ones. In this case, W refers to the weighted average multiplicity of states. The conformational entropy contribution is unfavorable and will oppose folding. Clearly, other driving forces such as the hydrophobic entropy gain and favorable enthalpic interactions have to compensate for this large conformational entropy loss for folding to occur.

1.6.1 Stabilization by conformational entropy

A simple example of a molecule whose dominant macrostate is stabilized by entropy is the peptide trialanine. The two conformers, compact α and extended β (shown in Fig. 1.3) can be distinguished experimentally[76,77] in solution. A conformer is a geometrically defined macrostate, or collection of microstates (conformations) with similar energies and geometries. Conformer α has a lower internal energy U_α due to having more favorable contacts than conformer β. As the energy difference $\Delta U_{\beta\alpha} = U_\beta - U_\alpha$ is smaller than thermal energy $k_B T$, significant interconversion occurs, and both conformers exist in equilibrium in the canonical ensemble.

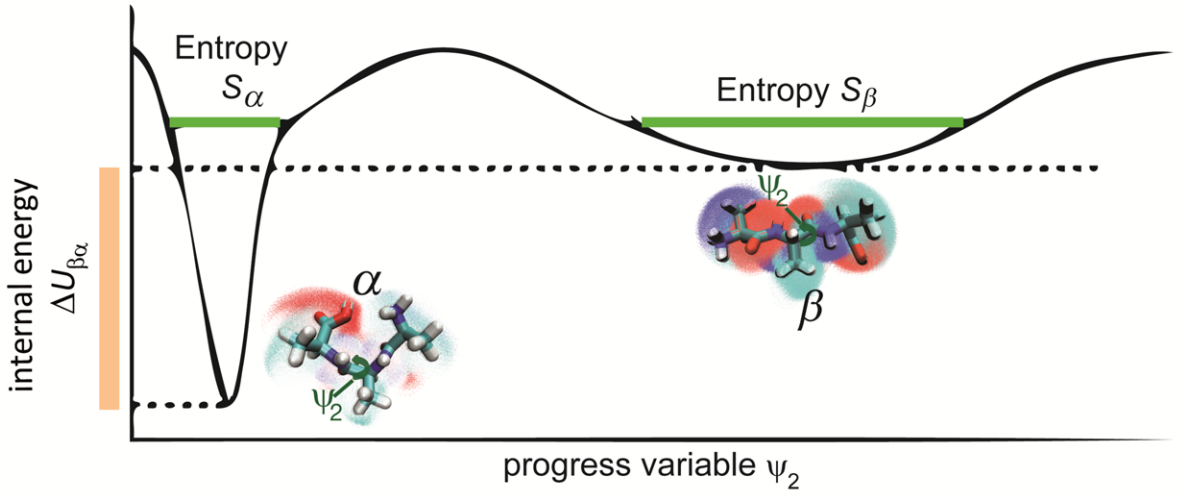

Fig. 1.3: Trialanine and its two conformers, α and β. The horizontal axis is a geometric "progress variable", in this case the central Ramachandran angle ψ_2. The conformation α is more compact and has lower energy (depicted as well height) than β. However, the extended β is favored by entropy (depicted as well width).

Although each individual high energy microstate in the basin of conformer β is less likely (see eq (1.7)), there are many more such states. The sheer multiplicity of states of conformer β allows it to be significantly populated, and indeed be the dominant conformer in solution[76,77]. Large multiplicity means high entropy, depicted here as the width of the well $S_\beta > S_\alpha$. In later sections, we use the concepts of microstate (individual conformation or frame in a molecular simulation) and macrostate (conformer) in the context of the thermodynamics of trialanine. We will use molecular dynamics simulations of the trialanine molecule to test the algorithms put forth in this thesis to estimate conformational entropy.

A biological example of conformational entropic stabilization, in this case of the folded state, occurs in hyperthermophilic organisms. Their genes code for proteins enriched in positively charged amino acids. But the positive charge is often achieved by the presence of lysines rather than arginines. This significant bias was recently explained[78] through the much higher number of rotamers available to lysine. Because the effective conformational freedom of lysine is greater, it constitutes a reservoir of conformational entropy, thereby stabilizing the protein's folded state.

A frequently unappreciated fact is that many proteins in the cell stably exist in a high entropy state. They populate a partially folded state sometimes called a "molten globule"[79], which has

significant secondary-structure but a fluid hydrophobic core[80]. Some of them acquire a more rigid structure only upon ligand binding, while others are intrinsically unstructured[81]. Such proteins are obviously underrepresented in the literature and the PDB database. In experiments, they are hard to discern from on-path folding intermediates. In molecular dynamics simulations, their high flexibility demands very long trajectories to acquire reliable statistics[82]. Nevertheless, there are recent theoretical studies in which molten globules were characterized with simulation[80].

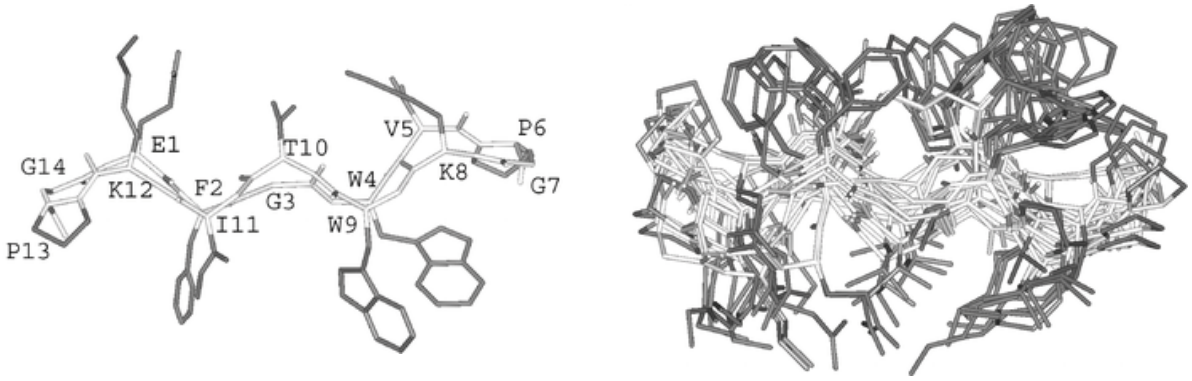

Fig. 1.4 Part of a polypeptide chain. Number of conformations in the native state (left) and the denatured state (right). The denatured state is clearly favored by conformational entropy, but is opposed by other thermodynamic forces that fold proteins.

1.6.2 Acceptable errors in theoretical estimations of protein thermodynamics

Enthalpy-entropy compensation is just enough to favor protein folding[83]. Proteins have evolved in such a way that changes in enthalpy from intramolecular interactions, electrostatic solvation free energy[84,85], a favorable hydrophobic effect[42,43,86,87] and an unfavorable conformational entropy loss almost cancel out. This tiny window for compensation makes simulation-based calculation of thermodynamic variables very challenging. Each one of the aforementioned factors can contribute 10 to 100 kcal/mol. The methods for estimating these enthalpic and entropic contributions need to be extremely precise in order not to miss the total free energy, which for a spontaneous process may be almost zero, but negative.

Protein folding and binding involve non-covalent interactions. These have very narrow free energy ranges in order to remain reversible. For biological processes, the net difference between states is often[48] between 5 and 15 kcal/mol. This is about 10 times more than thermal energy at

physiological temperature $k_B T$ = 0.6 kcal/mol. Ken Dill set an error goal of 0.100 kcal/mol per amino acid for the estimation of free energy of proteins[88]. Theoretical entropy estimation is even more difficult to converge than free energy estimation[89]. The methods presented in this thesis to estimate entropy achieved an average precision in S of 0.3 J/(mol K) for trialanine, which translates into an average error in TS of 0.007 kcal/mol per amino acid at T = 300K. This is 14 times better than Dill's goal[88]. Nevertheless, questions remain about the transferability of the entropy estimation technique to larger systems, and of the quality of the force field used in the molecular dynamics simulation.

In this thesis, I concentrate on the calculation of solute conformational entropy, and not on water entropy. Nevertheless, by applying techniques such as Permutation Reduction[90], it is possible to extend the applicability of algorithms such as the ones developed in this work to estimate water entropy. In fact, others[91] have already applied my quasi-harmonic algorithm with nearest-neighbors correction[92] combined with Permutation Reduction[90] to estimate water entropic components in an anti-cancer drug binding to DNA[91].

1.7 The entropy of polymers

Folded proteins are relatively rigid, entangled biological polymers with a low (but difficult to estimate) residual conformational entropy. Synthetic flexible polymers usually have considerably more conformational freedom because of the lack of specific microscopic structural preferences.

In the macroscopic world of everyday experience, there is a clear difference between stretching a thin metal wire and a rubber band. This is connected to their microscopic structure. When a metal is stretched, the individual bond lengths between atoms are pulled away from equilibrium. The restoring force comes from the bonds pulling back and is enthalpically-driven. On the other hand, when an elastic polymer like rubber is stretched, the bond lengths remain basically unstrained. Instead, the torsional motion of the chain becomes restricted. This is coupled with a decrease in entropy. The chain's natural tendency towards disorder (enabled by the thermal energy) is experienced as a restoring entropic force towards the molecule's freer configuration[93]. These ideas go back to K.H. Meyer and Flory[94,95].

Cross-linked polymers such as vulcanized rubber and the elastin[96] in human skin owe their material properties to entropic forces. Cross-linking provides additional entropic restoring networks. The elastic fibers in human arteries, especially in the aortic, survive for more than 60 years, undergoing billions of cycles of stretching and relaxation[97]. Their resilience is due to the dominantly entropic nature of the elastic force. Entropic forces have also been shown to account for the activity of actin polymerization in cell motility[27-30].

Barring interference from enthalpic stabilization, the conformers of polymer chains tend to their state of highest entropy. In this thesis, we use the random walk polymer model to simulate the behavior of polyethylene glycol (PEG) chains, a biocompatible polymer. We calculate the free energy and entropy changes as the chain interacts with the surface of a protein. The application is bivalent binding, where a bivalent receptor binds two ligands, which are themselves tethered together with a PEG chain. The PEG chain interacts with the protein surface such that it experiences an entropy loss, which may be compensated by the hydrophobic effect and favorable energetic contacts between the polymer and protein.

1.7.1 Connection of thermodynamics to information theory

There exist many analogies useful to understand entropy, each of which carries a greater or lesser amount of truth: freedom, flexibility, chaos, disorder, accessibility, spreading of energy[98]. However, my experience in working on this thesis tells me that understanding entropy as the Shannon missing information about the molecular microstates[15,64,99,100], or equivalently as a logarithmic counting of such microstates, is by far more useful. Identifying thermodynamic entropy with the Shannon entropy of the molecular microstates probabilities will help not only our intuitive understanding, but also allow us to further statistical mechanics with the tools of information theory.

1.8 Medical applications of protein and drug thermodynamics

A complete understanding of the thermodynamics of drug compounds and their interactions with metabolic actors such as proteins has long been a holy grail of theoretical chemistry. Advances in

modeling of thermodynamics including conformational entropy will bring new developments to medicinal chemistry and biopharmaceuticals.

Most drugs currently on the market are agonists or antagonists that directly bind to the active site of a protein. But not all effective drugs bind to the active site. Allosteric modulators[101] bind outside the receptor binding site, but can nevertheless induce a change in binding affinity, and thus a change in activity. Allosteric modulators directly affect the protein motion networks. Conformational entropy changes are thus key in understanding and modulating allostery[102,103]. Ultimately, strategies that take biomolecular dynamics into account will yield new lead compounds and drugs. Recently it was proposed that apparent *mismatches* between the inhibitor compound topology and the crystal structure of the target protein are a sign that a drug is not "enthalpically optimized" but rather "entropically optimized" to fit the multiplicity of conformations in solution[104].

Most HIV-1 protease inhibitors[105,106] to date are antagonists of the active site. Design of drugs less susceptible to resistance may be accomplished by altering the thermodynamics of stability and folding of the protease (PR) dimer. Allosteric inhibitors bind to residues whose dynamics are coupled to the flap opening-closing collective motion network. They may either keep the flap open or closed shut, inhibiting its cleaving activity[107]. Another strategy is to inhibit folding of the PR dimer; this has been achieved with peptides that bind and reshape the free energy landscape to make inactive conformations thermodynamically stable[108].

Alzheimer's, Hungtington's[109] and Creutzfeld-Jakob (prion) diseases[110] all share protein misfolding and aggregation as a common feature. Experiments have lent credibility to the hypothesis that β-amyloid aggregates are causal in the pathogenesis, at least in Alzheimer's Disease[111]. The normal folded and the aggregated misfolded conformations represent two local minima in the free energy landscape[112,113]. The misfolded conformation is much lower in entropy, but is stabilized by enthalpy, mostly through tight van der Waals interactions in a so-called steric zipper[114]. The two free energy minima are separated by a kinetic barrier to oligomerization. The design of compounds that block aggregation will hopefully be assisted by a detailed understanding of the thermodynamics and kinetics of misfolding.

Furthermore, exploiting the multivalent effect is an interesting avenue in medicinal chemistry for the enhancement of binding affinity[115,116] and the reduction of toxicity. Multivalent applications often rely on polymer carriers to join together several drug molecules[117]. Therapeutic compounds attached to polyethylene glycol spacers or to dendritic polymers[118] create high local (effective) concentrations of drugs.

1.9 Aim of this work

In this thesis, I present improvements in the theoretical estimation of conformational entropy and in the modeling of the multivalent effect using flexible polymer spacers. It is hoped that these algorithms and models may contribute to further advance theoretical methods and help understand, model and simulate biological processes from first principles for the advancement of medicine.

1.10 References for Chapter 1

(1) Alberty, R. A. *Biochemical Thermodynamics -Applications of Mathematica*, **2006**. ISBN:978-0471757986

(2) Almlöf, M.; Andér, M.; Åqvist, J., *Energetics of Codon-Anticodon Recognition on the Small Ribosomal Subunit.* Biochemistry **2007**, *46*, 200-209.

(3) Lazaridis, T.; Karplus, M., *Thermodynamics of protein folding: a microscopic view.* Biophys. Chem. **2003**, *100*, 367-395. dx.doi.org/doi:10.1016/S0301-4622(02)00293-4

(4) Gilson, M. K.; Zhou, H.-X., *Calculation of Protein-Ligand Binding Affinities.* Annu. Rev. Biophys. Biomol. Struct. **2007**, *36*, 21-24. dx.doi.org/doi:10.1146/annurev.biophys.36.040306.132550

(5) Shakhnovich, E., *Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry and Biology Meet.* Chem. Rev. **2006**, *106*. dx.doi.org/doi:10.1021/cr040425u

(6) van_Gunsteren, W. F.et al., *Biomolecular Modeling: Goals, Problems, Perspectives.* Angew. Chem. Int. Ed. **2006**, *45*, 4064-4092. dx.doi.org/doi:10.1002/anie.200502655

(7) Chipot, C.; Scott_Shell, M.; Pohorille, A.; Andricioaei, I.; Hummer, G.; Pande, V.; Mark, A.; Simonson, T. *Free energy calculations: Theory and applications in chemistry and biology*, **2007**. ISBN:978-3540384472

(8) Wereszczynski, J.; McCammon, J. A., *Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition.* Q. Rev. Biophys. **2011**, *FirstView Article*, 1-25. dx.doi.org/doi:10.1017/S0033583511000096

(9) Schrödinger, E., *What is life? The Physical Aspect of the Living Cell.* **1944**.

(10) Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*; Vintage Books USA, **2005**. ISBN:978-0679776314

(11) Ben-Naim, A., *On the So-Called Gibbs Paradox, and on the Real Paradox.* Entropy **2007**, *9*, 132-136. dx.doi.org/doi:10.3390/e9030133

(12) Bragg, M.; Atkins, P.; Grady, M.; Gribbin, J. *BBC Radio 4-In our time-The second law of thermodynamics*, **2004**. www.bbc.co.uk/radio4/history/inourtime/inourtime_20041216.shtml

(13) Clausius, R., *Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie.* Ann. der Physik **1865**, *201*, 353-400. dx.doi.org/doi:10.1002/andp.18652010702

(14) Shannon, C. E.; Weaver, W., *A mathematical theory of communication.* Bell Syst. Tech. J **1948**, *27*, 379-423 dx.doi.org/doi:10.1145/584091.584093

(15) Ben-Naim, A. *A Farewell To Entropy: Statistical thermodynamics based on information*; World Scientific Publishing Company: Singapore, **2008**. ISBN:978-9812707079

(16) Penrose, R., *Before the big bang: an outrageous new perspective and its implications for particle physics.* Proc. of EPAC **2006**.

(17) Penrose, R. *Cycles of Time: An Extraordinary New View of the Universe*; Bodley Head, **2010**. isbn:978-0224080361

(18) Bejan, A. *Advanced Engineering Thermodynamics*; 2nd edition ed.; Wiley Interscience: New York, **1997**. ISBN:978-0471148807

(19) Hawking, S., *Black holes and thermodynamics.* Phys. Rev. D **1976**, *13*, 191-197. dx.doi.org/doi:10.1103/PhysRevD.13.191

(20) Atkins, P.; Paula, J. d. *Physical chemistry for the life sciences*; WH Freeman and Co./Oxford University Press: New York, **2006**. ISBN:978-0716773290
(21) Jaynes, E. T.; Bretthorst, G. L. *Probability Theory: The Logic of Science*, **2003**. ISBN:978-0521592710
(22) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; 2nd. ed., **2006**. ISBN:978-0471241959
(23) Skupin, A.; Falcke, M., *Statistical analysis of calcium oscillations.* Eur. Phys. J. **2010**, *187*, 231-240. dx.doi.org/doi:10.1140/epjst/e2010-01288-9
(24) Thurley, K.; Skupin, A.; Thul, R.; Falcke, M., *Fundamental properties of Ca2+signals.* Biochim. Biophys. Acta **2011**, *In Press*. dx.doi.org/doi:10.1016/j.bbagen.2011.10.007
(25) Smith, J.; Van_Ness, H.; Abott, M. *Introduction to chemical engineering thermodynamics*; 5th ed.; McGraw Hill, **1996**. ISBN:978-0071147378
(26) Sept, D.; McCammon, J. A., *Thermodynamics and Kinetics of Actin Filament Nucleation.* Biophys. J. **2001**, *81*, 667-674.
(27) Enculescu, M.; Gholami, A.; Falcke, M., *Dynamic regimes and bifurcations in a model of actin-based motility.* Phys. Rev. E **2008**, *78*, 031915. dx.doi.org/doi:10.1103/PhysRevE.78.031915
(28) Gholami, A.; Falcke, M.; Frey, E., *Velocity oscillations in actin-based motility.* New J. Phys. **2008**, *10*, 033022. dx.doi.org/doi:10.1088/1367-2630/10/3/033022
(29) Enculescu, M.; Sabouri-Ghomi, M.; Danuser, G.; Falcke, M., *Modeling of Protrusion Phenotypes Driven by the Actin-Membrane Interaction.* Biophys. J. **2010**, *98*, 1571-1581. dx.doi.org/doi:10.1016/j.bpj.2009.12.4311
(30) Zimmermann, J.; Enculescu, M.; Falcke, M., *Leading-edge–gel coupling in lamellipodium motion.* Phys. Rev. E **2010**, *82*, 051925. dx.doi.org/doi:10.1103/PhysRevE.82.051925
(31) Dill, K. A.; Bromberg, S. *Molecular Driving Forces*; Garland Science, **2003**. ISBN:978-0815320517
(32) Boltzmann, L. *Vorlesungen ueber Gastheorie*; J.A. Barth: Leipzig, **1896**.
(33) Falk, G., *Entropy, a resurrection of caloric-a look at the history of thermodynamics.* Eur J Phys **1985**, *6*, 108-115. dx.doi.org/doi:10.1088/0143-0807/6/2/009
(34) Demetrius, L., *Role of Enzyme-Substrate Flexibility in Catalytic Activity: an Evolutionary Perspective.* J Theor. Biol. **1998**, *194*, 175-194. dx.doi.org/doi:10.1006/jtbi.1998.0748
(35) Bastolla, U.; Moya, A.; Viguera, E.; van_Ham, R. C. H. J., *Genomic Determinants of Protein Folding Thermodynamics in Prokaryotic Organisms.* J. Mol. Biol. **2004**, *343*. dx.doi.org/doi:10.1016/j.jmb.2004.08.086
(36) Billeter, S. R.; Webb, S. P.; Agarwal, P. K.; Iordanov, T.; Hammes-Schiffer, S., *Hydride Transfer in Liver Alcohol Dehydrogenase: Quantum Dynamics, Kinetic Isotope Effects, and Role of Enzyme Motion.* J. Am. Chem. Soc. **2001**, *123*, 11262-11272. dx.doi.org/doi:10.1021/ja011384b
(37) Hammes-Schiffer, S., *Enzyme Motions Inside and Out.* Science **2006**, *312*, 208-209. dx.doi.org/doi:10.1126/science.1127654
(38) Bahar, I.; Chennubhotla, C.; Tobi, D., *Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation.* Curr Opin Struct Biol. **2007**, *17*, 633-640.
(39) Hammes-Schiffer, S.; Benkovic, S. J., *Relating protein motion to catalysis.* Annual Review of Biochemistry **2006**, *75*, 519-541. dx.doi.org/doi:10.1146/annurev.biochem.75.103004.142800

(40) Rispens, T. *Cycloadditions in weakly and highly organized aqueous media*, Rijksuniversiteit Groningen, **2004**. dissertations.ub.rug.nl/FILES/faculties/science/2004/t.rispens/thesis.pdf
(41) Chandler, D., *Interfaces and the driving force of hydrophobic assembly.* Nature **2005**, *437*, 640-647. dx.doi.org/doi:10.1038/nature04162
(42) Southall, N. T.; Dill, K. A.; Haymet, A. D. J., *A View of the Hydrophobic Effect.* J. Phys. Chem. B **2002**, *106*, 521-533. dx.doi.org/doi:10.1021/jp015514e
(43) Setny, P.; Baron, R.; McCammon, J. A., *How Can Hydrophobic Association Be Enthalpy Driven?* J. Chem. Theory Comput. **2010**, *6*, 2866-2871. dx.doi.org/doi:10.1021/ct1003077
(44) Myslinski, J. M.; DeLorbe, J. E.; Clements, J. H.; Martin, S. F., *Protein-Ligand Interactions: Thermodynamic Effects Associated with Increasing Nonpolar Surface Area.* J. Am. Chem. Soc. **2011**, *133*, 18518-18521. dx.doi.org/doi:10.1021/ja2068752
(45) Baron, R.; Setny, P.; McCammon, J. A., *Water in Cavity-Ligand Recognition.* J. Am. Chem. Soc. **2010**, *132*, 12091-12097. dx.doi.org/doi:10.1021/ja1050082
(46) Mihailescu, M.; Gilson, M. K., *On the Theory of Noncovalent Binding.* Biophys. J. **2004**, *87*, 23-26. dx.doi.org/doi:10.1529/biophysj.103.031682
(47) Zoete, V.; Meuwly, M.; Karplus, M., *Study of the Insulin Dimerization: Binding Free Energy Calculations and Per-Residue Free Energy Decomposition.* Proteins: Struct., Funct., Bioinf. **2005**, *61*, 79-93. dx.doi.org/doi:10.1002/prot.20528
(48) Day, A. *The Source of Stability in Proteins (Doctoral presentation at Birkbeck University, London)*, **1996**. www.cryst.bbk.ac.uk/PPS2/projects/day/TDayDiss/
(49) Jaynes, E. T., *Gibbs vs. Boltzmann Entropies.* Am J Phys **1965**, *33*, 391-398. dx.doi.org/doi:10.1119/1.1971557
(50) Levitin, L. B.; Toffoli, T., *Heat-to-Work Conversion by Exploiting Full or Partial Correlations of Quantum Particles.* Int. J. Theor. Phys. **2011**, *50*, 3844-3851. dx.doi.org/doi:10.1007/s10773-011-0886-8
(51) Boltzmann, L., *Ueber die sogenannte H-Curve.* Mathematische Annalen **1898**, *50*, 325-332. dx.doi.org/doi:10.1007/BF01448073
(52) Lazar, T., *Book Reviews: Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. By K. A. Dill, S. Bromberg.* Macromolecular Chemistry and Physics **2003**, *204*, 1800. dx.doi.org/doi:10.1002/macp.200390113
(53) Einstein, A., *Eine neue Bestimmung der Moleküldimensionen.* Ann. der Physik **1906**, *19*.
(54) Uffink, J., *Insuperable difficulties: Einstein's statistical road to molecular physics.* Studies in History and Philosophy of Modern Physics **2006**.
(55) Einstein, A., *Kinetische Theorie des Wärmegleichgewichtes und des zweiten Hauptsatzes der Thermodynamik.* Ann. der Physik **1902**, *314*, 417-433. dx.doi.org/doi:10.1002/andp.19023141007
(56) Einstein, A., *Eine Theorie der Grundlagen der Thermodynamik.* Ann. der Physik **1903**, *14*, 135-153. dx.doi.org/doi:10.1002/andp.200590002
(57) Einstein, A., *Zur allgemeinen molekularen Theorie der Wärme.* Ann. der Physik **1904**, *319*, 354-362. dx.doi.org/doi:10.1002/andp.19043190707
(58) Kragh, H., *Max Planck: the reluctant revolutionary.* Physics World **2000**.
(59) Planck, M. *Nobel Lecture: The Genesis and Present State of Development of the Quantum Theory*, **1918**. nobelprize.org/nobel_prizes/physics/laureates/1918/planck-lecture.html

(60) Jaynes, E. T., *Information Theory and Statistical Mechanics (part 1).* Phys. Rev. **1957**, *106*, 620-630. dx.doi.org/doi:10.1103/PhysRev.106.620
(61) Landau, L. D.; Lifshitz, E. M. *Statistical Physics Part 1-Vol. 5*; 3rd ed. ed.; Oxford: Pergamon Press, **1980**.
(62) Reichl, L. E. *A Modern Course in Statistical Physics*; Wiley, **2009**. ISBN:978-3527407828
(63) Orzel, C. *The Advent Calendar of Physics: Entropy*, **2011**. scienceblogs.com/principles/2011/12/the_advent_calendar_of_physics_17.php
(64) Ben-Naim, A. *Entropy demystified: the second law reduced to plain common sense*; Entropy demystified: the second law reduced to plain common sense, **2007**. ISBN:9789812700551
(65) Szilard, L., *Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen.* Zeitschr. für Physik A **1929**, *53*, 840-856.
(66) Maddox, J., *Maxwell's demon: Slamming the door.* Nature **2002**, *417*. dx.doi.org/doi:10.1038/417903a
(67) Leff, H.; Rex, A. F. *Maxwell's Demon 2: Entropy, Classical and Quantum Information*, **2003**. ISBN:978-0750307598
(68) Toyabe, S.; Sagawa, T.; Ueda, M.; Muneyuki, E.; Sano, M., *Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality.* Nat. Phys. **2010**, *6*, 988–992. dx.doi.org/doi:10.1038/nphys1821
(69) Lanyon, B. P.; Whitfield, J. D.; Gillett, G. G.; Goggin, M. E.; Almeida, M. P.; Kassal, I.; Biamonte, J. D.; Mohseni, M.; Powell, B. J.; Barbieri, M.; Aspuru-Guzik, A.; White, A. G., *Towards quantum chemistry on a quantum computer.* Nat. Chem. **2010**, *2*, 106-111. dx.doi.org/doi:10.1038/nchem.483
(70) Nielsen, M. A.; Chuang, I. L. *Quantum Computation and Quantum Information*; Cambridge University Press, **2000**.
(71) Seife, C. *Decoding the Universe: How the New Science of Information Is Explaining Everything in the Cosmos, from Our Brains to Black Holes*; Penguin, **2007**. ISBN:978-0143038399
(72) Einstein, A.; Podolsky, B.; Rosen, N., *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?* Phys. Rev. **1935**, *47*, 777-780. dx.doi.org/doi:10.1103/PhysRev.47.777
(73) Salart, D.; Baas, A.; Branciard, C.; Gisin, N.; Zbinden, H., *Testing the speed of 'spooky action at a distance'.* Nature **2008**, *454*, 861-864. dx.doi.org/doi:10.1038/nature07121
(74) Shalizi, C. R.; Moore, C., *What Is a Macrostate? Subjective Observations and Objective Dynamics.* arXiv.org cond-mat **2003**.
(75) Makhatadze, G. I.; Privalov, P. L., *Hydration effects in protein unfolding.* Biophys. Chem. **1994**, *51*, 291-309. dx.doi.org/doi:10.1016/0301-4622(94)00050-6
(76) Schweitzer-Stenner, R.; Eker, F.; Huang, Q.; Griebenow, K., *Dihedral Angles of Trialanine in D2O Determined by Combining FTIR and Polarized Visible Raman Spectroscopy.* J. Am. Chem. Soc. **2001**, *123*, 9628-9633.
(77) Hamm, S. W. a. P., *Structure Determination of Trialanine in Water Using Polarization Sensitive Two-Dimensional Vibrational Spectroscopy.* J. Phys. Chem. B **2000**, *104*, 11316-11320. dx.doi.org/doi:10.1021/jp001546a
(78) Berezovsky, I. N.; Chen, W. W.; Choi, P. J.; Shakhnovich, E. I., *Entropic Stabilization of Proteins and Its Proteomic Consequences.* PLoS Computational Biology **2005**, *1*, e47. dx.doi.org/doi:10.1371/journal.pcbi.0010047.eor

(79) Ptitsyn, O. B., *Molten Globule and Protein Folding.* Adv. Prot. Chem. **1995**, *47*, 83-229. dx.doi.org/doi:10.1016/S0065-3233(08)60546-X

(80) Naganathan, A. N.; Orozco, M., *The Native Ensemble and Folding of a Protein Molten-Globule: Functional Consequence of Downhill Folding.* J. Am. Chem. Soc. **2011**, *133*, 12154-12161. dx.doi.org/doi:10.1021/ja204053n

(81) Wrighta, P. E.; Dyson, H. J., *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J. Mol. Biol. **1999**, *293*, 321-331. dx.doi.org/doi:10.1006/jmbi.1999.3110

(82) Gruebele, M.; Ervin, J.; Larios, E.; Osváth, S.; Schulten, K., *What Causes Hyperfluorescence: Folding Intermediates or Conformationally Flexible Native States?* Biophys. J. **2002**, *83*, 473-483.

(83) Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; Freeman, **1999**. ISBN:978-0716732686

(84) Koehl, P., *Electrostatics calculations: latest methodological advances.* Curr. Opin. Struct. Biol. **2006**, *16*, 142-151. dx.doi.org/doi:10.1016/j.sbi.2006.03.001

(85) Warshel, A.; Shara, P. K.; Kato, M.; Parson, W. W., *Modeling electrostatic effects in proteins.* Biochimica et Biophysica Acta **2006**, *1764*, 1647-1676.

(86) Hummer, G.; Garde, S.; García, A. E.; Pohorille, A.; Pratt, L. R., *An information theory model of hydrophobic interactions.* Proc. Natl. Acad. Sci. U.S.A **1996**, *93*, 8951-8955. dx.doi.org/doi:10.1073/Proc. Natl. Acad. Sci. U.S.A.93.17.8951

(87) Sharp, K.; Nicholls, A.; Fine, R.; Honig, B., *Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects.* Science **1991**, *252*, 106-109. dx.doi.org/doi:10.1126/science.2011744

(88) Dill, K. A., *Additivity principles in biochemistry.* J Biol Chem **1997**, *1997*, 701-704.

(89) Pearlman, D. A.; Rao, B. G., *Free energy calculations: Methods and applications.* In *Encyclopedia of computational chemistry*; Schleyer, P. v. R., Ed. **1998**; Vol. 2, p 1036-1061. doi:10.1002/0470845015.cfa011

(90) Reinhard, F.; Grubmüller, H., *Estimation of absolute solvent and solvation shell entropies via permutation reduction.* J. Chem. Phys. **2007**, *126*, 014102. dx.doi.org/doi:10.1063/1.2400220

(91) Mukherjee, A., *Entropy Balance in the Intercalation Process of an Anti-Cancer Drug Daunomycin.* J. Phys. Chem. Lett. **2011**, *2*, 3021-3026. dx.doi.org/doi:10.1021/jz2013566

(92) Numata, J.; Wan, M.; Knapp, E. W., *Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation.* Genome Inform. **2007**, *18*, 192-205. dx.doi.org/doi:10.1142/9781860949920_0019

(93) Feynman, R., *The Feynman Lectures on Physics 1-44: The Laws of Thermodynamics.* **1962**, *1*.

(94) Flory, P. J. *Principles of polymer chemistry*, **1953**. ISBN:978-0801401343

(95) Teraoka, I. *Polymer solutions: An Introduction to Physical Properties*, **2002**. ISBN:978-0-471-38929-3

(96) Urry, D.; Hugel, T.; Seitz, M.; Gaub, H.; Sheiba, L.; Dea, J.; Xu, J.; Parker, T., *Elastin: a representative ideal protein elastomer.* Philosophical Transactions of The Royal Society B **2002**, *357*, 169-184. dx.doi.org/doi:10.1098/rstb.2001.1023

(97) Urry, D. W., *Elastic biomolecular machines.* Scientific American **1995**, 64-69.

(98) Leff, H. S., *Entropy, Its Language, and Interpretation.* Found. Phys. **2007**, *37*, 1744-1766. dx.doi.org/doi:10.1007/s10701-007-9163-3

(99) Ben-Naim, A. *Entropy and the Second Law: Interpretationsss and Misss-Interpretationsss*; World Scientific Publishing Company, **2012**. ISBN:978-9814374897

(100) Ben-Naim, A. *Discover Entropy and the Second Law of Thermodynamics: A Playful Way of Discovering a Law of Nature*; World Scientific Publishing Company, **2010**. ISBN:978-9814299763

(101) H Lüllmann, A. Z., K Mohr, D Bieger *Color Atlas of Pharmacology (2nd ed.)*; Thieme: Stuttgart, **2000**. ISBN:978-3137817024

(102) DuBay, K. H.; Bothma, J. P.; Geissler, P. L., *Long-Range Intra-Protein Communication Can Be Transmitted by Correlated Side-Chain Fluctuations Alone.* PLoS Comput. Biol. **2011**, *7*, e1002168. dx.doi.org/doi:10.1371/journal.pcbi.1002168.g004

(103) Tzeng, S.-R.; Kalodimos, C. G., *Protein dynamics and allostery: an NMR view.* Curr. Opin. Struct. Biol. **2011**, *21*, 62-67. dx.doi.org/doi:10.1016/j.sbi.2010.10.007

(104) Fernández, A.; Fraser, C.; Scott, L. R., *Purposely engineered drug-target mismatches for entropy-based drug optimization.* Trends Biotech. **2011**, *In Print*. dx.doi.org/doi:10.1016/j.tibtech.2011.07.003

(105) Juneja, A.; Riedesel, H.; Hodoscek, M.; Knapp, E. W., *Bound Ligand Conformer Revealed by Flexible Structure Alignment in Absence of Crystal Structures: Indirect Drug Design Probed for HIV-1 Protease Inhibitors.* J. Chem. Theory Comput. **2009**, *5*, 659-673. dx.doi.org/doi:10.1021/ct8004886

(106) Shuman, C. F.; Hämäläinen, M. D.; Danielson, U. H., *Kinetic and thermodynamic characterization of HIV-1 protease inhibitors.* J Mol Recognit **2004**, *17*, 106-119. dx.doi.org/doi:10.1002/jmr.655

(107) Perryman, A. L.; Lin, J.-H.; McCammon, A., *HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs.* Prot Sci **2004**, *13*, 1108-1123. dx.doi.org/doi:10.1110/ps.03468904

(108) Broglia, R. A.; Levy, Y.; Tiana, G., *HIV-1 protease folding and the design of drugs which do not create resistance.* Curr. Opin. Struct. Biol. **2008**, *18*, 60-66. dx.doi.org/doi:10.1016/j.sbi.2007.10.004

(109) Numata, J. *Conformational search of Huntingtin in the early steps of aggregation (MSc. Thesis)*, Freie Universität Berlin, **2005**.

(110) Chiti, F.; Dobson, C. M., *Protein Misfolding, Functional Amyloid, and Human Disease.* Annu. Rev. Biochem. **2006**, *75*, 333-366. dx.doi.org/doi:10.1146/annurev.biochem.75.101304.123901

(111) Meyer-Luehmann, M.; Spires-Jones, T. L.; Prada, C.; Garcia-Alloza, M.; Calignon, A. d.; Rozkalne, A.; Koenigsknecht-Talboo, J.; Holtzman, D. M.; Bacskai, B. J.; Hyman, B. T., *Rapid appearance and local toxicity of amyloid beta plaques in a mouse model of Alzheimer's disease.* Nature **2008**, *451*, 720-724. dx.doi.org/doi:10.1038/nature06616

(112) Straub, J. E.; Thirumalai, D., *Toward a Molecular Theory of Early and Late Events in Monomer to Amyloid Fibril Formation.* Annu. Rev. Phys. Chem. **2011**, *62*, 437-63. dx.doi.org/doi:10.1146/annurev-physchem-032210-103526

(113) Tarus, B.; Straub, J. E.; Thirumalai, D., *Structures and Free-Energy Landscapes of the Wild Type and Mutants of the Aβ21–30 Peptide Are Determined by an Interplay between*

Intrapeptide Electrostatic and Hydrophoboc Interactions. J. Mol. Biol. **2008**, *379*, 815-829. dx.doi.org/doi:10.1016/j.jmb.2008.04.028

(114) Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J. W.; McFarlane, H. T.; Madsen, A. Ø.; Riekel, C.; Eisenberg, D., *Atomic structures of amyloid cross-bold beta spines reveal varied steric zippers.* Nature **2007**, *447*, 453-457. dx.doi.org/doi:10.1038/nature05695

(115) Joshi, A.; Vance, D.; Rai, P.; Thiyagarajan, A.; Kane, R. S., *The design of polyvalent therapeutics.* Chem. Eur. J. **2008**, *14*, 7738-7747. dx.doi.org/doi:10.1002/chem.200800278

(116) Rolland, O.; Turrin, C.-O.; Caminade, A.-M.; Majoral, J.-P., *Dendrimers and nanomedicine: multivalency in action.* New J. of Chem. **2009**, *33*, 1809-1824. dx.doi.org/doi:10.1039/B901054H

(117) Calderón, M.; Welker, P.; Licha, K.; Fichtner, I.; Graeser, R.; Haag, R.; Kratz, F., *Development of efficient acid cleavable multifunctional prodrugs derived from dendritic polyglycerol with a poly(ethylene glycol) shell.* J. Control. Release **2011**, *151*, 295-301. dx.doi.org/doi:10.1016/j.jconrel.2011.01.017

(118) Stiriba, S.-E.; Frey, H.; Haag, R., *Dendritic Polymers in Biomedical Applications: From Potential to Clinical Use in Diagnostics and Therapy.* Angew. Chem. Int. Ed. **2002**, *41*, 1329-1334. dx.doi.org/doi:10.1002/1521-3773(20020415)41:8<1329::AID-ANIE1329>3.0.CO;2-P

2 Balanced and bias-free computation of conformational entropy differences for molecular trajectories

This chapter is connected to an accepted publication:

Numata, J.; Knapp, E. W., *Balanced and bias-corrected computation of conformational entropy differences for molecular trajectories*. J. Chem. Theory Comput. **2012**, *8*(4), 1235-1245. Submitted on 20-dec-2011. Accepted on 14-mar-2012. dx.doi.org/10.1021/ct200910z

2.1 Introduction

A macrostate of a molecular system can be specified by appropriate thermodynamic variables. The conformational entropy of a molecular system is a measure of the missing information about the specific molecular conformation (microstate) adopted among the many available conformations of the macrostate. This interpretation follows Jaynes' work[1,2] and Ben-Naim's reformulation of statistical mechanics in terms of information theory[3]. The physical entropy[4] S is proportional to the dimensionless information entropy[5] $S_{inf} = -\Sigma\, p_i\, ln(p_i)$ according to $S = k_B\, S_{inf}$, where the p_i are the probabilities that the molecular system adopts a particular microstate i and k_B is the Boltzmann constant. The interplay of entropy S and the average internal energy $<U>$ is described by the free energy expression $F = <E> - TS$. The Boltzmann factor $\exp(-F/k_BT)$, involving the free energy F, provides the relative probabilities of occupation for specific macrostates at a given absolute temperature T.

> "Once in a while, engineering has contributed a great deal to [physical theory]. Two examples that come to mind [are] the analysis of heat engines by the engineer Carnot. And the other is the analysis of information theory by Shannon, recently. And both of those are closely related phenomena, it turns out."
>
> Richard Feynman, 1962
> *The Feynman Lectures on Physics 1-44: The Laws of Thermodynamics.*

Knowledge on conformational entropy differences is an essential ingredient to understand binding affinities. Conformational entropy is the missing link to a full free energy difference when using methods such as MMPB/SA (enthalpy from the Molecular Mechanics force field, solvation free energy from the Poisson-Boltzmann equation, hydrophobic effect

from the solvent accessible Surface Area)[6-9]. In the present work, we develop efficient algorithms for estimating conformational entropy differences of macro-molecular systems comprising many degrees of freedom. We apply these methods to two model systems: (1) A three-atom molecule in two different confined spaces. (2) Trialanine in implicit solvent with two conformer regimes. For both models, we generate very precise benchmark entropy values to compare with.

2.1.1 Experimental measurements of conformational entropy

Using thermodynamic relations, it is possible to separate enthalpic and entropic contributions to free energy changes measured experimentally[10]. The absolute conformational entropy of the protein backbone has also been estimated from Atomic Force Microscopy measurements[11]. However, the separation of the total entropy change into solvent and solute components is in general not straightforward[12]. The reason is that conformational entropy is a measure of the microscopic variability of conformations, a level of detail which is challenging to resolve experimentally. Recently, experimental techniques from Nuclear Magnetic Resonance (NMR) have been used to peek into the microscopic states of the solute through order parameters and to estimate solute conformational entropy changes upon binding[13-17]. Along the same line, a view of allosteric phenomena and protein recognition is emerging from NMR experiments, which support the interpretation of allostery[18] (i.e. certain spatially distant sites in a protein are strongly correlated) as a network of molecular groups undergoing concerted motions[19], and establishes conformational entropy changes as key in modulating allostery[20] and protein-ligand recognition[21].

While the motion of sequentially adjacent amino acid residues can obviously be correlated, a recent study suggests that residues as distant as 15 Å can be even more strongly correlated than residues in close spatial contact[22]. It is noteworthy that Shannon conceived the concept of information entropy[5] for communication channels. We can now quantify communication between amino acid residues[23] using his theory. Nevertheless, the interpretation of NMR data to estimate entropy often does not consider correlations among order parameters[24]. Methods from information theory such as those in the present study can be used to account for non-linear correlations[25,26] in the molecular coordinates. Since correlations always reduce the entropy[3], including them will provide a tighter upper bound to conformational entropy.

2.1.2 Theoretical estimation of conformational entropy

Macromolecules involve many degrees of freedom. Therefore, they constitute a special kind of challenge in the estimation of conformational entropy. A variety of methods have been proposed to tackle this problem[27]. The quasi-harmonic approximation (QHA)[28-33] is based on 'principal component analysis' (PCA), also known as eigenvalue decomposition, which accounts for linear correlations between pairs of coordinates. It fits the observed probability density for the eigenmode coordinates of an effective harmonic oscillator model for which statistical mechanical quantities like the entropy can be expressed analytically. More elaborate QHA approaches apply corrections in third order moments of the coordinates[34] or in pair-wise supra-linear correlations [35,36]. A further development of pair-wise supra-linear correlations is the 'minimally coupled subspace' approach[37]. It combines 'independent component analysis'[38] with 'mutual information (MI) expansion'[39] and 'adaptive kernel density estimator' approaches[37].

DNA[33] and RNA[40] display 'collective coordinates' (eigenmodes), which are close to harmonic modes in Cartesian coordinates using PCA[41] or QHA[28-33]. Hence, applying these methods in Cartesian coordinates to ribonucleotides[42], only small corrections for anharmonicity and pairwise supralinear correlations are needed using the nearest-neighbor method[35]. However, peptides and proteins possess different types of degrees of freedom, where Cartesian coordinates describing the conformations of polypeptide chains are highly correlated, even after applying PCA or QHA[43]. Internal Bond-Angle-Torsion (BAT) coordinates can avoid such correlations to a large extent[44] and can also be applied in the quasi-harmonic approximation[45-47].

Other approaches fit the observed distributions in torsional angle space to probability distributions given in closed form[48] like Gaussian and/or von-Mises kernel density estimators[49-53]. The latter approach is non-parametric and approximates the probability density as a sum of peaks for which an analytical expression of the entropy is available. Another unrelated method to compute entropy, inspired by polymer physics, is the rigorous but computationally demanding hypothetical scanning[54,55], which is based on reconstructing the macromolecular chain conformer from scratch. Methods originally devised to estimate free energy differences, like thermodynamic perturbation and integration, have also be extended to estimate entropy differences[56-58].

In this work, we employ the internal BAT coordinates combined with a histogram method to estimate entropy with the mutual information (MI) expansion[39], which is capable of accounting for supralinear correlations[59,60]. Most importantly, we introduce novel techniques that expedite convergence and compensate bias in estimating conformational entropy differences.

2.2 Analytical derivation: Configurational entropy of a macromolecule

2.2.1 Absolute and relative configurational entropies

To define entropy in the canonical ensemble of a macromolecule with N atoms, we start with the partition function of the conformer domain α

$$Q_\alpha = h^{-3N} \int d\mathbf{p}^N \int_{\Omega_\alpha} d\mathbf{r}^N \exp(-E_\alpha / k_B T) . \tag{2.1}$$

The Hamiltonian

$$E_\alpha = \sum_{n=1}^{N} p_n^2 / 2m_n + U_\alpha(\mathbf{r}^N) . \tag{2.2}$$

in eq (2.1) involves kinetic and potential energy terms of the N atom macromolecule. The symbol Ω_α signifies the domain of configurations that identify the conformer α. The potential energy $U_\alpha(\mathbf{r}^N)$ is a function in the $3N$ Cartesian coordinates denoted by the $3N$-dimensional vector $\mathbf{r}^N$. The energy $U_\alpha(\mathbf{r}^N)$ is infinite outside of the domain α that defines the conformer. Integrating over the $3N$ momenta in eq (2.1), we can write Q_α as

$$Q_\alpha = Z_\alpha \Big/ \prod_{n=1}^{N} \Lambda_n^3$$

in terms of the configuration integral

$$Z_\alpha = \int_{\Omega_\alpha} \exp[-U_\alpha(\mathbf{r}^N)/(k_B T)] \, d\mathbf{r}^N \tag{2.3}$$

and the "momentum" contribution, expressed by the $3N$-fold product of the thermal de Broglie wavelengths

$$\Lambda_n = h/\sqrt{2\pi m_n k_B T}\ , \tag{2.4}$$

where m_n is the mass of atom n, k_B is the Boltzmann constant, h is Planck's constant, and T is the absolute temperature. The free energy F_α of the conformer in domain α is

$$F_\alpha = -k_B T \ln(Q_\alpha). \tag{2.5}$$

The ensemble average of the internal energy is

$$\langle E_\alpha \rangle = k_B T^2 \left(\frac{\partial \ln Q_\alpha}{\partial T} \right)_{N,V} = \tfrac{3}{2} N k_B T + \langle U_\alpha \rangle\ , \tag{2.6}$$

where the ensemble average of the potential energy can be written as

$$\langle U_\alpha \rangle = \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) U_\alpha(\mathbf{r}^N) \tag{2.7}$$

using the probability density function

$$P_\alpha(\mathbf{r}^N) = \exp[-U_\alpha(\mathbf{r}^N)/(k_B T)] \big/ Z_\alpha\ . \tag{2.8}$$

Rearranging eq (2.8) and taking logarithms of both sides, we get

$$U_\alpha(\mathbf{r}^N) = -k_B T \ln[P_\alpha(\mathbf{r}^N) Z_\alpha]\,. \tag{2.9}$$

Substitution of eq (2.9) into eq (2.7) gives

$$\langle U_\alpha \rangle = -k_B T \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N) Z_\alpha] \ . \tag{2.10}$$

Now we define the configurational entropy of the conformer domain α as

$$S_\alpha = (<E_\alpha> - F_\alpha) / T \, . \tag{2.11}$$

Using eq (2.6) and (2.10) we can rewrite the **absolute configurational entropy**, eq (2.11), as

$$S_\alpha = \tfrac{3}{2} N k_B - k_B \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N) \prod_{n=1}^{N} \Lambda_n^3], \tag{2.12}$$

or rearranging

$$S_\alpha = \tfrac{3}{2} N k_B - 3 k_B \sum_{n=1}^{N} \ln \Lambda_n - k_B \int_{\Omega_\alpha} d\mathbf{r}^N P_\alpha(\mathbf{r}^N) \ln[P_\alpha(\mathbf{r}^N)] . \tag{2.13}$$

In the configurational entropy difference, $\Delta S_{\alpha\beta} = S_\alpha - S_\beta$, the first two terms in eq (2.13) cancel, if both entropies refer to the same temperature, yielding

$$\Delta S_{\alpha\beta} = k_B (\hat{s}_\alpha - \hat{s}_\beta), \tag{2.14}$$

where the **relative configurational entropy** is defined by

$$\hat{s}_\delta = -\int_{\Omega_\delta} d\mathbf{r}^N P_\delta(\mathbf{r}^N) \ln[P_\delta(\mathbf{r}^N)] \, , \ \delta = \alpha, \beta \, , \tag{2.15}$$

which is analog to the Shannon differential entropy[5] for the probability density $P_\delta(\mathbf{r}^N)$. If entropy differences at different temperatures are evaluated, eq (2.13) should be used. In this work we can use eq (2.15), since we compute entropy differences at the same temperature.

Eq (2.15) is the expression for a relative entropy for two reasons: (i) Its actual value varies by an additive constant term dependent on the length units (e.g. Ångström) used for the coordinates $\mathbf{r}^N$. (ii) It is a differential (continuous[61]) entropy, which may assume negative or positive values (see App. I of ref 3 and sec. 20 of Shannon[5]). Conversely, the expression (2.13) is an absolute entropy[62] because: (i) The length units used in the conformational integral cancel. (ii) Planck's constant h discretizes (quantizes) the phase space (cf. eq 7.12 of Landau & Lifshitz[63]).

Alternatively, (2.15) may be rewritten

$$\hat{s}_\delta = -\left\langle \ln(P_\delta(\mathbf{r}^N)) \right\rangle \equiv \left\langle \ln\left[\frac{1}{P_\delta(\mathbf{r}^N)} \right] \right\rangle . \tag{2.16}$$

The inverse of a probability density is the multiplicity. From (2.16), it becomes clear that entropy is a logarithmic measure of the average multiplicity of microstates, which is called degeneracy in the context of quantum mechanics.

2.2.2 Sackur-Tetrode equation as a limiting case for ideal gas

We now calculate the absolute configurational entropy for an ideal gas, where the probability distribution is uniform throughout the volume V of the container, such that

$$P_d(\mathbf{r}^N) = 1/V^N . \tag{2.17}$$

Solving eq (2.13) analytically, we get

$$S_d = \tfrac{3}{2} N k_B - 3k_B \sum_{n=1}^{N} \ln \Lambda_n - k_B N \ln V \tag{2.18}$$

Additionally, we have identical masses, so $m_n = m$, so the absolute entropy for distinguishable particles (2.18) becomes

$$S_d = Nk_B \ln\left[V\left(\frac{2\pi e\, m k_B T}{h^2} \right)^{3/2} \right] \tag{2.19}$$

For an ideal gas, we have indistinguishable particles, so

$$S_{ST} = S_d - k_B \ln N! \tag{2.20}$$

The last term in (2.20) accounts for indistinguishability, which in this work only applies to the ideal gas. Using Stirling's approximation, we have $\ln N! \approx N \ln N - N$ and may write

$$S_{ST} = Nk_B \ln\left[\frac{Ve^{5/2}}{N}\left(\frac{2\pi m k_B T}{h^2} \right)^{3/2} \right] \equiv Nk_B \ln\left[\frac{V}{N}\left(\frac{2\pi m k_B T}{h^2} \right)^{3/2} \right] + \frac{5}{2} Nk_B \,, \tag{2.21}$$

which is the familiar Sackur-Tetrode equation[64].

2.2.3 Entropy using local spherical polar (BAT) coordinates

We introduce local spherical polar coordinates[65-67], also referred to as 'bond-angle-torsion' (BAT) coordinates[59,68,69], since they represent the conformational displacements of the atoms of a macromolecule in a more natural fashion than Cartesian coordinates. BAT coordinates simplify the configurational integrals, as for instance eqs (2.3) or (2.15), since they involve bond lengths, bond angles and torsions. This coordinate system is local because the frame of reference is shifted and rotated at each new bond to accommodate the molecular topology. These local coordinates are adapted to the molecular structure by separating degrees of freedom with high flexibility (torsion angles) from those with low flexibility (bonds and bond angles). This helps avoiding strong but spurious correlations inherent in atomic Cartesian coordinates[44]. Since these correlations are large, they can mask the physically relevant correlations.

These coordinates are defined by fixing the coordinate $\mathbf{r}_1$ of the terminal atom 1 of the macromolecule at the origin of the coordinate system. All other coordinates refer to the bond vectors $\mathbf{b}_n$. The local spherical coordinates for the bond vector are $\mathbf{b}_n = (b_n, \theta_n, \varphi_n)$,

$n = 2, 3, \ldots N$, (bond length b_n, inclination angle θ_n, azimuthal angle φ_n). We begin with bond vector $\mathbf{b}_2 = \mathbf{r}_2 - \mathbf{r}_1$ of the end atom 1, using the z- and x-axes from a lab frame as a reference for rotations θ_2 and φ_2. For the second bond vector $\mathbf{b}_3 = \mathbf{r}_3 - \mathbf{r}_2$, we use $\mathbf{b}_2$ as a reference for θ_2 but still need the x-axis from the lab frame as a reference for φ_3. For a linear molecule, the bond vectors are consecutively $\mathbf{b}_n = \mathbf{r}_n - \mathbf{r}_{n-1}$, $n = 4, 5, \ldots N$. For the local spherical coordinates of bond vector $\mathbf{b}_n$ we take atom position $\mathbf{r}_{n-1}$, as the coordinate origin, the preceding bond vector $\mathbf{b}_{n-1}$ as z-axis, and the unit vector parallel to the cross product $\mathbf{b}_{n-2}\times\mathbf{b}_{n-1}$ as x-axis. In a non-linear, branched molecule, we use for all bond vectors following a branch point (atom with more than two covalent bonds) the bond vector of the preceding two bonds as reference for z- and x-axes. Independently of the degree of branching, a molecule with N atoms and no ring structure possesses $N-1$ covalent bonds. Each ring introduces an additional bond. To avoid overcompleteness, one covalent bond in each ring is ignored, which automatically transforms the molecular topology back to a branched structure. Thus, together with the coordinates $\mathbf{r}_1$ of the initial atom 1 a complete set of $3N$ BAT coordinates (Fig. 2.1) is obtained for an N atom molecule. These BAT coordinates are collected in the $3N$-dimensional supervector

$$\vec{\mathbf{b}} = (\mathbf{r}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots . \mathbf{b}_{N-1}, \mathbf{b}_N), \text{ with } \mathbf{b}_n = (b_n, \theta_n, \varphi_n), \quad n = 2, 3, \ldots N. \tag{2.22}$$

The potential energy function U_α is independent of position and orientation of the solute in the solvent. Therefore, we can separate contributions of those degrees of freedom and perform the corresponding integrations in configurational integrals as for instance eqs (2.3) or (2.15) in closed form using BAT coordinates[65,66,68,69]. The integration over $\mathbf{r}_1$ in configuration integrals like eq (2.15) can be performed directly, yielding as a result the volume V available to solvent and solute together.

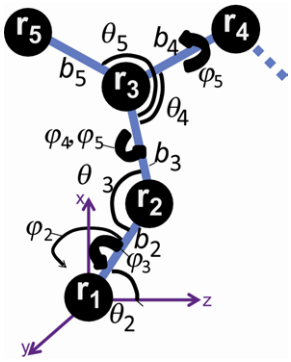

Fig. 2.1: Local spherical polar coordinates (BAT coordinates) of a branched molecule. The lab frame (purple) is the initial reference for external rotations θ_2, φ_2 and φ_3. Further up the chain, the frame of reference is local and defined by the chemical bonds.

Rotating the first bond vector $\mathbf{b}_2 = \mathbf{r}_2 - \mathbf{r}_1$ together with the whole solute molecule is described by varying the polar coordinate angles (θ_2, φ_2). Similarly the whole molecule can be rotated about the bond $\mathbf{b}_2$ described by the azimuthal angle φ_3. The potential energy function U_α of the solute does not depend on the orientation of the whole solute molecule. Hence, the integrations over θ_2, φ_2 and φ_3 in the configuration integrals like eq (2.15) can be performed directly to give the factor $8\pi^2$. As a result we have for instance for the (configurational) state sum, eq (2.3)

$$Z_\alpha = V 8\pi^2 \int db_2\, b_2^2 \int db_3\, b_3^2 \int_0^{2\pi} d\theta_3 \sin\theta_3 \prod_{n=4}^{N} \int d^{(3)}\mathbf{b}_n \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T]\,, \tag{2.23}$$

with the differential of the local spherical polar coordinates

$$d^{(3)}\mathbf{b}_n = b_n^2 db_n \sin\theta_n\, d\theta_n\, d\varphi_n \tag{2.24}$$

and the $3N$–6 BAT variables combined in the ($3N$-6)-dimensional vector

$$\vec{\mathbf{b}}' = (b_2, b_3, \theta_3, \mathbf{b}_4, \mathbf{b}_5, \ldots . \mathbf{b}_{N-1}, \mathbf{b}_N), \quad \mathbf{b}_n = (b_n, \theta_n, \varphi_n)\,. \tag{2.25}$$

with bond length b_n, inclination angle θ_n, azimuthal angle φ_n. In analogy to eq (2.25) we also define the differential form

$$d^{(3N-6)}\mathbf{b}' = {b_2}^2 db_2 \times {b_3}^2 db_3 \times \sin\theta_3 \, d\theta_3 \times \prod_{n=4}^{N} d^{(3)}\mathbf{b}_n \, . \qquad (2.26)\text{a}$$

Thus, we can write now the (configurational) state sum, eq (2.23) in the compact form

$$\frac{Z_\alpha}{V 8\pi^2} = \int d^{(3N-6)}\mathbf{b}' \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T] \equiv \breve{z}_\alpha \, , \qquad (2.27)$$

where $\breve{z}_\alpha$ is now the conformational state sum exclusive of the position and orientations of the solute. We can now define the reduced conformational probability distribution

$$\rho_\alpha(\vec{\mathbf{b}}') = \exp[-U_\alpha(\vec{\mathbf{b}}')/k_B T]/\breve{z}_\alpha \, . \qquad (2.28)$$

2.2.3.1 Relative conformational entropy in terms of BAT coordinates

The reduced relative conformational entropy, which neglects translation and orientation of the macromolecule, is

$$s_\delta = -\int_{\Omega_\delta} d^{(3N-6)}\mathbf{b}' \, \rho_\delta(\vec{\mathbf{b}}') \ln[\rho_\delta(\vec{\mathbf{b}}')] \equiv -\left\langle \ln(\rho_\delta) \right\rangle , \; \delta = \alpha, \beta. \qquad (2.29)$$

Hence, the entropy differences of a molecular system can be expressed by the dimensionless configurational entropies as is done in eq (2.14) or alternatively by the reduced dimensionless conformational entropies, eq (2.29) according to

$$\Delta S_{\alpha\beta} = k_B \, (\hat{s}_\alpha - \hat{s}_\beta) \equiv k_B \, (s_\alpha - s_\beta). \qquad (2.30)$$

In case an implicit solvent model is used, the potential energy function U_δ is explicitly defined and the configurational integral in eq (2.29) can, *in principle*, be evaluated directly. For an explicit solvent model U_α depends implicitly on the thermodynamic state of the system (i.e., pressure and temperature) and involves averaging the Boltzmann factor $\exp(-U_{solv}/(k_B T))$, where U_{solv} is the solute-solvent interaction, over "free" solvent configurations for each fixed solute

conformation[70]. In both cases, the resulting U_α incorporates the influence of the solvent on the distribution of the molecular conformations[71]. In practical applications, it is often more advantageous to sample the conformational probability density $\rho_\alpha(\vec{\mathbf{b}}')$ through simulations. It should be noted that a rigorous separation of the conformational entropy of a solute from the entropy of the embedding solvent[72,73] is not possible in the current scheme because of the correlations between the two molecular subsystems.

2.3 Numerical method to estimate conformational entropy differences

We are now prepared to numerically estimate conformational entropy differences for a macromolecule (solute) immersed in a solvent possessing two distinct conformer domains α and β using eq (2.30). These conformers are for instance conformational domains separated by torsional energy barrier or the native folded and denatured unfolded structures of a protein[74]. The entropy difference (2.30) can be expressed in terms of the reduced relative conformational entropy (2.29) for each conformer $\delta = \alpha,\beta$, corresponding to the Shannon differential entropy[5] of the probability density (2.28).

The conformational entropy, eq (2.29), of a macromolecule with N atoms involves an integral in $3N–6$ dimensions. Hence, even for a small macromolecule, solving such integrals suffers from the curse of dimensionality. It is virtually impossible to perform these integrals explicitly even for molecules of a few atoms. Alternatively, one can use sampling methods based on molecular dynamics (MD) or Monte Carlo (MC) simulations, which generate molecular conformers in the frame of a canonical or quasi-canonical ensemble. In case the computation of canonical ensemble averages (free energy, enthalpy and entropy) can be based on a single equilibrated trajectory, *importance sampling* with Metropolis MC or MD simulation[75] and energy averaging are straightforward techniques to evaluate them. We will use this method to obtain reliable benchmark data to compare our results with. This method is however not applicable if the problem requires the use of different trajectories from independent simulations, as is generally necessary for studying protein-ligand binding. Hence, other procedures are needed which can

deal with data from different trajectories. An alternative for these cases is the numerical estimation presented here.

To estimate conformational entropy differences, we use non-redundant internal BAT coordinates for the given molecular topology. The high dimensionality of the conformational space is reduced through the mutual information expansion. The probability densities of the lower-dimensional subspaces in BAT coordinates are discretized by a histogram method and used in the calculation of conformational entropy differences. The biases inherent in entropy estimation are compensated through the bias-removal and balancing methods.

2.3.1 Automated selection of BAT coordinates

For a given molecular topology, a set of non-redundant internal BAT coordinates is constructed using the procedure described in section 2.2.3 "Entropy using local spherical polar (BAT) coordinates". In practice, this translates into a tree algorithm also described by Gilson et al[59]. The PERL implementation of the BAT tree algorithm by Thomas Steinbrecher[76], which in turn uses ptraj[77], is adapted and modified to use Charmm/NAMD trajectories.

2.3.1.1 Continuity maximization for torsions

In contrast to molecular bond angles, torsion angles can vary over the whole angular regime from 0 to 2π, such that the 2π periodicity must be considered to avoid discontinuities. We apply a 'continuity maximization' algorithm to deal with this problem. For each torsion angle, its one-dimensional probability distribution is discretized with a large number of histogram bins (say 1000), many more than will finally be used for entropy computations. In this histogram, the longest continuous stretch of empty bins is detected. The end points of the angular interval for the histogram used to evaluate the entropies are placed such that they exclude this regime. If no histogram bin is empty, the original angular distribution is kept and used for the entropy evaluation. For a torsional coordinate with values that cover the whole 2π span, the choice of the end points formally has no effect, and numerically it would only have a vanishing one. However, for an angular variable that covers only part of the 2π span, this algorithm avoids considering a large number of empty (unused) histogram bins.

2.3.1.2 Phase angles

If P torsions φ_n share three atoms (for a methyl group P = 3), geometrical correlations can be reduced furthermore by transforming P − 1 torsions into phase angles[78] ϕ_n. The hydrogens of a methyl group display such behavior, for which we define a master torsion angle, say φ_i , and two phase angles[78] ϕ_k. Generally, if the torsion angles φ_i and φ_j have three atoms in common, we keep φ_i and substitute φ_j by the phase angle

$$\phi_j = \varphi_j - \varphi_i \,. \tag{2.31}$$

This transformation has a unit Jacobian and preserves a complete geometric description of the molecule. In Fig. 2.1, the atoms with coordinates $\mathbf{r}_4$ and $\mathbf{r}_5$ give rise to torsions φ_4 and φ_5. According to eq (2.31) we substitute torsion φ_5 by the phase angle $\phi_5 = \varphi_5 - \varphi_4$. Such phase angles[78] have narrower distributions than torsion angles.

In our algorithm, main chain torsions (of the polypeptide backbone) are kept as full torsions, and the ones defined at branches (describing side chain orientations) are converted into phase angles. Both phase angles and the ability to define main chain atom types are implemented in our modified version of the BAT tree algorithm.

2.3.2 Mutual Information expansion in low dimensional subspaces

The convergence of the reduced conformational entropy s, eq (2.29), suffers from the curse of dimensionality. Therefore, we approximate s by a systematic series, projecting the probability distribution function ρ from the L-dimensional space, spanned by the generalized coordinates $\vec{q}^{\,t} = (q_1, q_2, \ldots q_L)$ into subspaces of lower dimensions as defined below

$$\rho_{(3)\,i,j,k}(q_i, q_j, q_k) = -\int \rho(\vec{q})\, d^{(L-3)}q_{i,j,k} \text{ , with } \; d^{(L-3)}q_{i,j,k} = \prod_{l \neq i,j,k}^{L} dq_l \tag{2.32}a$$

and the analog expressions of two- and one-dimensional reduced probability distribution functions

$$\rho_{(2)\,i,j}(q_i,q_j) = \int \rho_{(3)\,i,j,k}(q_i,q_j,q_k)\; dq_k \qquad (2.32)\text{b}$$

and

$$\rho_{(1)\,i}(q_i) = \int \rho_{(2)\,i,j}(q_i,q_j)\, dq_j\,. \qquad (2.32)\text{c}$$

The factors J_n appearing in the conformational integrals of (2.32) are from the Jacobian determinant describing the transformation of the volume element from Cartesian to generalized coordinates. In the present application we use BAT coordinates, eq (2.26)a, where according to eq (2.26)b the Jacobian factors J_n are

$$J_n = b_n^{\,2} \qquad \text{for } q_n = b_n\,, \qquad (2.33)\text{a}$$

$$J_n = \sin\theta_n \qquad \text{for } q_n = \theta_n\,, \qquad (2.33)\text{b}$$

$$J_n = 1 \qquad \text{for } q_n = \varphi_n\,. \qquad (2.33)\text{c}$$

Individual values of the low dimensional probability densities (2.32) can be readily estimated from a finite set of simulation data. Their statistical accuracy improves the lower the dimension of the considered subspace is. With these reduced probability distributions, one can define entropy expressions in the corresponding low dimensional conformational space as for instance

$$s_{(3)\,i,j,k} = -\int \rho_{(3)\,i,j,k}(q_i,q_j,q_k)\, ln\Big(\rho_{(3)\,i,j,k}(q_i,q_j,q_k)\Big) \prod_{l=i,j,k} J_l\, dq_l\,, \qquad (2.34)\text{a}$$

for the three-dimensional subspace and analog expressions for the two- and one-dimensional subspaces

$$s_{(2)\,i,j} = -\int \rho_{(2)\,i,j}(q_i,q_j)\, ln\Big(\rho_{(2)\,i,j}(q_i,q_j)\Big) \prod_{l=i,j} J_l\, dq_l \qquad (2.34)\text{b}$$

and

$$s_{(1)\,i} = -\int \rho_{(1)\,i}(q_i)\, ln\Big(\rho_{(1)\,i}(q_i)\Big)\, J_i dq_i\;. \qquad (2.34)\text{c}$$

We are now prepared to formulate the MI expansion for entropies in an L dimensional phase space[35,39,59]

$$S_{MIE} = \sum_{i=1}^{L} S_{(1)\,i} - \sum_{i<j=1}^{L} I_{(2)\,i,j} + \sum_{i<j<k=1}^{L} I_{(3)\,i,j,k} - \ldots \,. \tag{2.35a}$$

The terms $I_{(2)\,i,j}$ and $I_{(3)\,i,j,k}$ in (2.35) are the mutual information terms of 2^{nd} and 3^{rd} order, respectively, which are defined as

$$I_{(2)\,i,j} = S_{(1)\,i} + S_{(1)\,j} - S_{(2)\,i,j} \tag{2.35b}$$

and

$$I_{(3)\,i,j,k} = S_{(1)\,i} + S_{(1)\,j} + S_{(1)\,k} - (S_{(2)\,i,j} + S_{(2)\,i,k} + S_{(2)\,j,k}) + S_{(3)\,i,j,k} \,. \tag{2.35c}$$

The MI expansion (2.35) starts with the sum of marginal entropy contributions $S_{(1)i}$ in the individual one-dimensional subspaces, neglecting correlations between them. The next terms correct for these correlations up to a given order. The MI expansion can in principle be extended to any desired order, up to full dimensionality[39]. However, higher order terms have a notoriously difficult convergence behavior. In the present work, we will use the MI expansion up to third order. According to our experience, this is sufficient to evaluate entropies from molecular trajectories reliably when internal BAT coordinates are used.

2.3.3 Discretization

To evaluate the subspace entropies according to (2.34), the molecular conformer coordinates obtained from a simulation of a canonical ensemble are discretized using histogram bins. In a three-dimensional subspace spanned by the coordinates $\vec{q}_{ijk} = (q_i, q_j, q_k)$, the bins are numbered by the integer vector $\vec{m}_{ijk} = (m_i, m_j, m_k)$, where the m_l, ($l = i, j, k$) run from 1 to M_l and their widths are given by Δq_l. If the total number of conformations belonging to conformer regime δ is $N^{(\delta)}$ and the number of conformations in the bin $\vec{m}_{ijk}$ belonging to the three-dimensional subspace spanned by $\vec{q}_{ijk}$ is $N^{(\delta)}_{(3)\,i,j,k}(\vec{m}_{ijk})$, the corresponding discretized probability is

$$p^{(\delta)}_{(3)\,i,j,k}\left(\vec{m}_{ijk}\right) = N^{(\delta)}_{(3)\,i,j,k}(\vec{m}_{ijk}) \Big/ \left(N^{(\delta)} \prod_{l=i,j,k} J^{(\delta)}_l(m_l)\,\Delta q_l \right), \tag{2.36a}$$

such that the probability density function is normalized to unity according to

$$1 \equiv \tfrac{1}{N^{(\delta)}} \sum_{m_i,m_j,m_k=1}^{M_i,M_j,M_k} N^{(\delta)}_{(3)\,i,j,k}(\vec{m}_{ijk}) \simeq \int \rho^{(\delta)}_{(3)\,i,j,k}(\vec{q}_{ijk})\, dq_i dq_j dq_k \ .$$

The $J^{(\delta)}_l(m_l)$ refer to the Jacobian factors (2.33) for the different BAT coordinates q_l. Using analog definitions, the discretized probabilities for two- and one-dimensional subspaces (normalized the same way as in the three-dimensional subspace) are

$$p^{(\delta)}_{(2)\,i,j}\left(\vec{m}_{ij}\right) = N^{(\delta)}_{(2)\,i,j}(\vec{m}_{ij}) \Big/ \left(N^{(\delta)} \prod_{l=i,j} J^{(\delta)}_l(m_l)\,\Delta q_l \right) \tag{2.36b}$$

and

$$p^{(\delta)}_{(1)\,i}\left(m_i\right) = N^{(\delta)}_{(1)\,i}(m_i) \Big/ \left(N^{(\delta)} J^{(\delta)}_i(m_i)\,\Delta q_i \right). \tag{2.36c}$$

Based on these discretized probabilities, the entropies in the three-, two- and one-dimensional subspaces spanned by $\vec{q}_{ijk}$, $\vec{q}_{ij}$ and q_i, respectively can be written as

$$s^{(\delta)}_{(3)\,i,j,k} \simeq -\left(\prod_{l=i,j,k} \Delta q_l \right) \sum_{m_i,m_j,m_k=1}^{M_i,M_j,M_k} \left(\prod_{l=i,j,k} J^{(\delta)}_l(m_l) \right) p^{(\delta)}_{(3)\,i,j,k}\left(\vec{m}_{ijk}\right) \ln\left(p^{(\delta)}_{(3)\,i,j,k}\left(\vec{m}_{ijk}\right) \right), \tag{2.37a}$$

$$s^{(\delta)}_{(2)\,i,j} \simeq -\left(\prod_{l=i,j} \Delta q_l \right) \sum_{m_i,m_j=1}^{M_i,M_j} \left(\prod_{l=i,j} J^{(\delta)}_l(m_l) \right) p^{(\delta)}_{(2)\,i,j}\left(\vec{m}_{ij}\right) \ln\left(p^{(\delta)}_{(2)\,i,j}\left(\vec{m}_{ij}\right) \right), \tag{2.37b}$$

$$s^{(\delta)}_{(1)\,i} \simeq -\Delta q_i \sum_{m_i=1}^{M_i} J^{(\delta)}_i\left(m_i\right) p^{(\delta)}_{(1)\,i}\left(m_i\right) \ln\left(p^{(\delta)}_{(1)\,i}\left(m_i\right) \right) \tag{2.37c}$$

used to evaluate the MI expansion (2.35). To account for the periodicity of the torsion angles, the histogram bins are placed appropriately, using an adaptive algorithm, as described in section 2.3.1.1 “Continuity maximization for torsions”.

2.3.4 Bias-Removal

The entropy expressions (2.37) are based on estimates of the probability density function using finite samples that represent the canonical ensemble. These are subject to fluctuations, which lead to systematic deviations (bias) that underestimate the true value of entropy[79,80]. In the limit of small probabilities to find the molecular system in one particular bin of the histogram, a simple correction (bias-removal) term can be added that compensates this bias and yields bias-free (unbiased) entropy estimates according to

$$\hat{s}^{(\delta)}_{(1)\,i} = s^{(\delta)}_{(1)\,i} + \frac{\hat{M}^{(\delta)}_{i} - 1}{2N^{(\delta)}} \ , \tag{2.38a}$$

$$\hat{s}^{(\delta)}_{(2)\,i,j} = s^{(\delta)}_{(2)\,i,j} + \frac{\hat{M}^{(\delta)}_{ij} - 1}{2N^{(\delta)}} \ , \tag{2.38b}$$

$$\hat{s}^{(\delta)}_{(3)\,i,j,k} = s^{(\delta)}_{(3)\,i,j,k} + \frac{\hat{M}^{(\delta)}_{ijk} - 1}{2N^{(\delta)}} \ . \tag{2.38c}$$

The $\hat{M}^{(\delta)}$ count only the occupied bins of the histograms, (with $p^{(\delta)}(\vec{m}) > 0$), such that $\hat{M}^{(\delta)}_{ijk} \le \prod_{l=i,j,k} M_l$ and $N^{(\delta)}$ is the total number of frames (molecular conformations) in the sample. Evidently, the corrections are larger for entropy terms in higher dimensional subspaces[81,82]. The bias-removal corrections (2.38) depend only on parameters characterizing the evaluation of the data ($\hat{M}^{(\delta)}$, $N^{(\delta)}$) and not on the particular system considered ($p^{(\delta)}$).

2.3.5 Balancing

In practice, we are often interested in entropy differences between different states, for example between two conformer regimes (see eq (2.30)) or between the bound and unbound states of a protein-ligand system. Using a finite number of frames (molecular conformations from

simulation), the entropy difference may already have converged, although the entropies of individual conformer regimes have not. We have noticed that convergence of entropy differences is most efficient when the same number of effectively independent frames ($N^{(\alpha)} \approx N^{(\beta)}$) is used for both conformer regimes (α, β). When the two conformer regimes are simulated in a single-trajectory, an imbalance occurs if $N^{(\alpha)} > N^{(\beta)}$. Discretizing these data in a histogram, eqs (2.36), the systematic errors for the subspace entropies, eqs (2.37), will differ for the two conformers. Hence, the systematic errors will not cancel in the entropy difference, eq (2.30), using these subspace entropies. To avoid this problem, the set of data are balanced keeping all frames of the minority conformer regime β, while reducing the number of frames of the majority conformer regime α by randomly deleting frames of the conformer α. If instead only a contiguous part of the trajectory is used to reduce the number of frames of the majority conformer, the two conformer regimes are no longer explored under the same conditions. Effects of such a nonequivalent exploration are discussed in section 2.5.6.1"Importance of choosing frames at random in the balancing method".

Here we provide an explanation for the observation that with less data for the majority conformer regime better estimates of entropy differences are obtained. Smooth probability distributions have higher entropy than rough distributions. For example, a perfectly smooth Gaussian probability distribution provides the maximum entropy for fixed variance[83]. Alternatively, a rough, multi-peaked probability density of the same variance contains more information, since the multi-peaked distribution 'classifies' data in more detail. It is well established[82] that the statistical bias originates from statistical variations in the bin values of the histogram p, eq (2.36) representing the true probability density ρ, eq (2.32). It is evident that histograms will on average become smoother the more data are used to estimate the distribution. We conjecture that balancing works well because it produces histograms with comparable roughness in both conformer regimes. Thus, the bias from the histogram roughness cancels in the entropy difference (2.30). This behavior is detailed under section 2.5.6 "Convergence of the entropy estimates".

Hence, we recommend applying balancing when the conformers of both regimes are taken from the same trajectory. However, balancing will also work, if different trajectories of the same

molecular system are simulated under equivalent conditions. In contrast to the bias-removal correction that applies to entropies of individual states (conformer regimes), the balancing correction applies only to entropy differences. While the bias-removal correction term (2.38) removes systematic deviations (biases) connected solely with the evaluation procedure, balancing accounts also for systematic deviations that depend on the particular system under study.

2.3.6 Generating molecular conformations in a canonical ensemble

Data that represent a canonical ensemble of molecular conformations can be generated by MD. We use Langevin dynamics as implemented in CHARMM35b1 as thermostat. To avoid slowing down of dynamics as observed in ref 84 a friction constant of $\gamma_{\mathrm{Lang}} = 1\ \mathrm{ps}^{-1}$ is used[85]. However, other thermostats such as the Andersen thermostat[86] or Nosé-Hoover chains[87] may also be appropriate[88].

Some implicit solvent models such as GBMV with standard parameters are known not to conserve energy[89] in microcanonical (NVE) MD simulations because of the complexity of the molecular surface of the solute used to approximate the Poisson-Boltzmann solvation free energy. As a consequence, these models combined with a thermostat may generate imperfect canonical ensembles. Therefore, we prefer to use the energy conserving implicit solvent model FACTS[90], defined purely on the basis of pairwise distances between atoms.

2.3.7 Benchmark Entropy

The free energy change between two molecular conformer regimes can be calculated from a single trajectory if equilibrated simulation data reflecting Boltzmann statistics are available. $N^{(\alpha)}$ and $N^{(\beta)}$ are the number of conformations (frames) for the conformer regimes α and β obtained from a simulation using *importance sampling* with MD or Metropolis MC[75]. This is sometimes called the *counting method*[27] to obtain the configurational free energy difference

$$\Delta F_{\alpha\beta} = F_\alpha - F_\beta = -k_B T \ln(N^{(\alpha)} / N^{(\beta)})\,. \tag{2.39}$$

The entropy difference between two conformer domains α and β of a macromolecule with N atoms can be written

$$\Delta S_{\alpha\beta\,,bench} = (\Delta\langle E_{\alpha\beta}\rangle - \Delta F_{\alpha\beta})/T\,. \tag{2.40}$$

Using eq (2.6) $\langle E_\delta\rangle = {}^3/_2\, Nk_BT + \langle U_\delta\rangle$ and given conformers α and β at identical temperature T, the difference of the internal energy

$$\Delta\langle E_{\alpha\beta}\rangle = \Delta\langle U_{\alpha\beta}\rangle = \langle U_\alpha\rangle - \langle U_\beta\rangle\,, \tag{2.41}$$

can be evaluated from MD simulation data by averaging the potential energies U_δ, $\delta = \alpha, \beta$, over all frames of conformer regime α and β, respectively. Entropy differences computed from data of a single trajectory based on (2.40) converge more rapidly than an MI expansion. Therefore, they can be used as a benchmark to test the MI expansion method. Conversely, the MI expansion can also be applied to compute entropy differences for situations where the conformer regimes need to be generated by independent MD simulations where relation (2.39) cannot be applied to compute the free energy difference. Such independent trajectories are required for instance to evaluate the binding affinities of ligand-receptor or protein-protein complexes.

Using MD simulation data to evaluate $\Delta F_{\alpha\beta}$ according to (2.39), $\Delta F_{\alpha\beta}$ converges more rapidly with the length of the trajectory than the evaluation of $\Delta S_{\alpha\beta}$ and $\Delta U_{\alpha\beta}$ based on (2.40) and (2.41), respectively. This convergence behavior has been reported previously[56] and is discussed at length for our simulations in section 2.5.4 “Convergence of benchmark entropy, energy and free energy”. The simulation data are obtained with *importance sampling* based on Boltzmann statistics such that for an evaluation of $\Delta F_{\alpha\beta}$ all frames are used with equal weights, while for evaluation of $\Delta U_{\alpha\beta}$ the frames need to be reweighed using the potential energy terms $U^{(\delta)}$, $\delta = \alpha, \beta$. Hence, the effective number of frames available for the latter case is smaller, resulting in larger statistical errors. As a consequence, the convergence of the benchmark value of $\Delta S_{\alpha\beta}$ according to (2.40) is limited by the convergence behavior of $\Delta U_{\alpha\beta}$.

2.4 Model system 1: Monte Carlo simulation of a three-atom molecule in a cage

2.4.1 Simulation procedure

The first model system that we investigate is a three-atom molecule whose conformations are generated by a continuous random walk starting at the origin with fixed step size (bond length). The first atom is considered to be fixed at the minimum of a wall opened toward the positive z-axis defining a cage (see Fig. 2.2). The wall surface obeys the relation

$$z_{\text{wall}}(x,y) = \varepsilon(x^2+y^2)^{1/2} \quad , \varepsilon > 0. \tag{2.42}$$

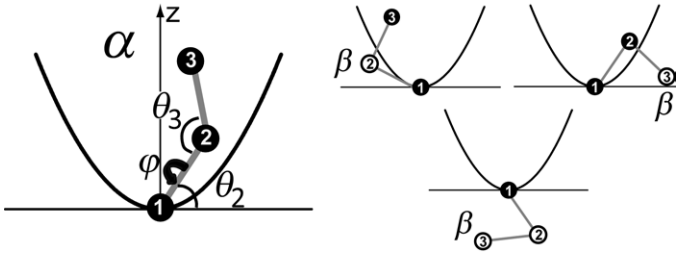

Fig. 2.2: Three-atom molecule modeled as continuous random walk with fixed bond lengths. Each conformation is defined by two bond angles (θ_2, θ_3) and one torsion angle φ. The conformer regime α is the ensemble of conformations where atom 2 and 3 are both above the depicted parabolic wall, eq (2.42), while conformer β comprises all other conformers (right part).

The second atom can change its position by varying its angle θ_2 relative to the plane rectangular to the *z*-axis. The third atom can move by rotating around the axis formed by atoms 1 and 2 by the azimuthal angle φ and by varying the bond angle θ_3 of the three atoms. Rotations of the molecule around the *z*-axis do not matter, since they do not change the configuration of wall and molecule, due to the rotational symmetry of the wall surface, eq (2.42). The set of angular variables (θ_2, θ_3, φ) are analog to the BAT coordinates. They are the internal coordinates of the molecule fixed with atom 1 at the wall.

2.4.1.1 Monte Carlo algorithm

A simple Monte Carlo (MC) algorithm is used to generate 5×10^7 frames of the free, unrestricted 3-atom molecule by a Random Walk, (RW). To generate each frame in Cartesian coordinates, we proceed as follows:

1. Place the first atom at the origin: $\mathbf{r}_1 = (0, 0, 0)$
2. Place atom 2 at $\mathbf{r}_2 = \mathbf{r}_1 + \mathbf{b}_2$.
3. Place atom 3 at $\mathbf{r}_3 = \mathbf{r}_2 + \mathbf{b}_3$.

Here, $\mathbf{b}_n$ is a vector whose tip is uniformly randomly distributed on a sphere of radius b, where b is the fixed bond length. This is accomplished through an algorithm due to Marsaglia[91], which is an optimized version of von Neumann's algorithm[92]. The independent, identically distributed pseudorandom numbers required by Marsaglia's algorithm[91] are generated by the pseudorandom number generator Taus088 due to L'Ecuyer[93].

The ensemble of free conformations is now subject to restriction by a hard wall described by, eq (2.42) with $\varepsilon = 0.612$. The constant ε is chosen arbitrarily to provide a positive curvature and divide the conformers unevenly. The conformer regime α comprises the frames where all atoms are above z_{wall}. The rest, where any or all atoms are below z_{wall}, is denominated β.

2.4.2 Clustering of conformations

The locations of atoms 2 and 3 relative to the wall surface determine the conformer regime (α or β) to which the molecule belongs (see Fig. 2.2). If both atoms are above the wall, the molecule belongs to conformer regime α. If one or both atoms are below the wall, the molecule is in conformer regime β. In this way we have constructed a molecular model with an asymmetric distribution between the two conformer regimes. Choosing the parameter value $\varepsilon = 0.612$ for the wall surface, eq (2.42), yields an asymmetry of 1 to 10.4 in the proportion of conformations between regime α and β. A simple MC procedure is used to generate 5×10^7 free molecular conformations. Then the wall surface, eq (2.42), is introduced and the molecular conformers are assigned to one of the two conformer regimes.

2.4.3 Entropy estimation

Here, we estimate the entropy change between the two conformers for the three-atom molecule in a cage. A two step continuous unconstrained random walk starting at the coordinate origin can be considered as a three-atom model where the first atom is fixed at the origin. The Cartesian, as well as the three internal coordinates (θ_2, θ_3, φ) (for a definition see Fig. 2.2) of such a molecular model are by construction uncorrelated. By introducing a wall to divide the ensemble of conformers into two regimes, correlations between the coordinates are introduced. All three internal coordinates are supralinearly[26] correlated, as evidenced by non-vanishing pairwise $I_{(2)\,i,j}$ and third order $I_{(3)\,1,2,3}$ MI terms[39].

For the chosen value of the curvature ε = 0.612 of the quadratic wall, eq (2.42), we obtain for 5×10^7 random walks (frames) $N^{(\alpha)} = 4.38 \times 10^6$ conformers of type α and $N^{(\beta)} = 45.6 \times 10^6$ conformers of type β (Fig. 2.2). Since β is the majority conformer, applying balancing means to randomly select N_α frames of conformer regime β for the entropy difference computation. The benchmark entropy, eq (2.40), may be used as a standard. It converges quickly with the number of frames (solid line in Fig. 2.3) to the value $\Delta S_{\alpha\beta} = -\Delta F_{\alpha\beta}/T = k_\mathrm{B} \ln(N^{(\alpha)} / N^{(\beta)}) =$ −19.5 J/(mol K), since $\Delta U_{\alpha\beta} = 0$.

Among the estimators, the slowest convergence is observed (Fig. 2.3) when neither balancing nor bias-removal is applied, equivalent to the original method by Gilson et al[59,60]. The fastest convergence is achieved by applying both balancing and bias-removal. When separating the effects, the balancing method alone provides a stronger improvement for a small number of frames (N_frames), while bias-removal provides a stronger improvement for larger number of frames N_frames. In practical applications, it is advisable to apply both balancing and bias-removal, as they work synergistically to accelerate convergence.

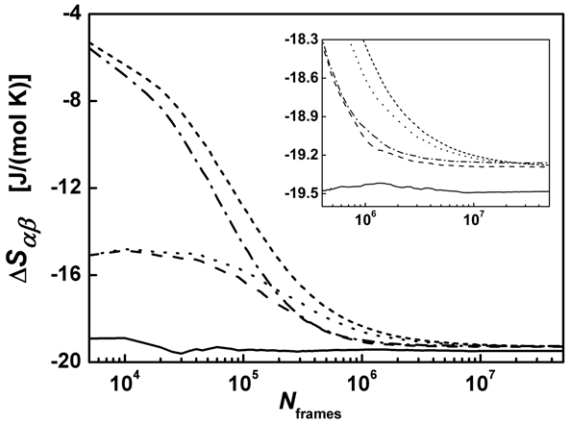

Fig. 2.3: Entropy difference $\Delta S_{\alpha\beta}$ for the three-atom molecule as described in section 2.4.3, probing the correction methods: balancing and bias-removal. The computations are based on a total of 5×10^7 conformers. The solid line —— is the benchmark entropy difference, eq (2.40). All entropy estimators use the third order MI expansion with M=35 histogram bins, eqs (2.35)–(2.37). Dashed line – – balanced & bias-free; dotted line · · · · balanced & biased; dash-dotted line –·– unbalanced & bias-free (reflected to be above the benchmark); short dashed line ----· unbalanced & biased (reflected to be above the benchmark). The latter two curves have been reflected about their asymptotic values $\Delta S_{\alpha\beta}(\infty)$ according to $\Delta S_{\alpha\beta} = \Delta S_{\alpha\beta}(\infty) - \Delta S_{\alpha\beta}$ for ease of comparison. The inlay zooms into the last phase of convergence. The fastest convergence among estimators is achieved by applying both methods: bias-free and balancing.

The abscissa in Fig. 2.3 and Fig. 2.4 is the total number of frames in the simulation. Because in the balancing method we actually use only a subset of those frames, the CPU requirements of the entropy evaluation are reduced by one order of magnitude, while at the same time improving the convergence. Nevertheless, we use the same abscissa to allow comparison between the methods. The number of histogram bins M chosen is the resolution at which the conformational space and the correlations between the different variables will be sampled. The dependence on M is plotted in Fig. 2.4 for the uncorrected (a), and for the balanced and bias-free methods (b). If we choose M too large, there will not be enough data to fill the bins, and convergence will be slower and incomplete for the given amount of data, which is 5×10^7 conformers. If we choose too small an M, the resolution will not be high enough to capture the correlations. The values of M between 20 and 35 are most suitable for the third order MI expansion using 5×10^7 conformers. There is, however, a dependence of $\Delta S_{\alpha\beta}$ on M, which is reduced by using balancing and bias-removal, but not completely eliminated. In summary, most of the entropy estimates have reached their asymptotic value when using balancing and bias-removal (Fig. 2.4b). Conversely the corresponding results are far from being converged, if the MI expansion method is used without corrections, i.e. unbalanced and biased (Fig. 2.4a).

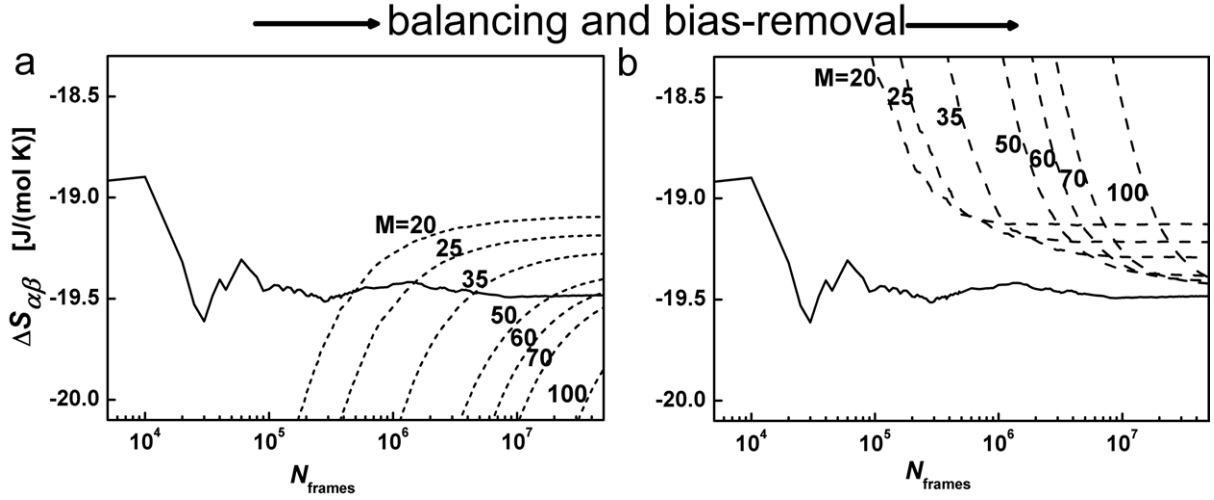

Fig. 2.4: Entropy difference $\Delta S_{\alpha\beta}$ for the three-atom molecule probing different histogram sizes using **a:** uncorrected estimates (unbalanced and biased) and **b:** corrected entropy estimates (balanced and bias-free) with improved asymptotic convergence. The computations are based on a total of 5×10^7 conformations. Solid line is the benchmark entropy according to eq (2.40). All entropy estimators (dashed lines) use the third order MI expansion varying the number of histogram bins M. Note that the curves with M = 35 are identical to Fig. 2.3 except for the reflection.

2.5 Model system 2: Molecular Dynamics simulation of trialanine

2.5.1 Simulation procedure

Simulations of trialanine were performed with 13 different conditions, each one spanning a 1 μs trajectory. The canonical ensemble was approximated using the Langevin thermostat with coupling constant γ_{Lang} = 1 ps^{-1}. The time propagation step was 1 fs. No SHAKE constraints were used to account also for entropy contributions from hydrogen atom bond vibrations. Conformations were saved every 0.2 ps for a total of $N_{frames} = 5 \times 10^6$. The CHARMM22[94] force field was used together with the implicit solvation model FACTS[90] with parameters κ = 8 and dielectric constant ε = 1.0 implemented in CHARMM35b1. In order to generate a total of 13 simulations with different entropies, we varied the hydrophobic "surface tension" term $\gamma_{H\phi}$ and scaled the attractive $1/r^6$ term of the Lennard Jones potential by the dimensionless factor ε_{attr}. For vanishing surface tension ($\gamma_{H\phi}$ = 0.0), we used $\varepsilon_{attr}(j) = 0.00 + 0.25\ j$, j = 0, 1, 2, . . . 6, and for $\gamma_{H\phi}$ = 0.025 cal/(mol K Å^2) and $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2), we used ε_{attr} = 0.00; 0.50; and 1.0.

2.5.2 Clustering of conformations

The molecular conformations of the trajectories were post-processed to generate two conformers by using a geometric criterion. The main anharmonic motion in trialanine is about the dihedral angle ψ_2 of the middle residue[95,96], which we have chosen as our 'order parameter' (see Fig. 2.6). We separate two conformers of trialanine by searching for two minima in occupation in the torsion angle ψ_2. As a result we obtain a conformer regime α with dihedral angles similar to an α-helix and a conformer regime β with torsion angles similar to polyglycine 3_1-helix (P_{II}).

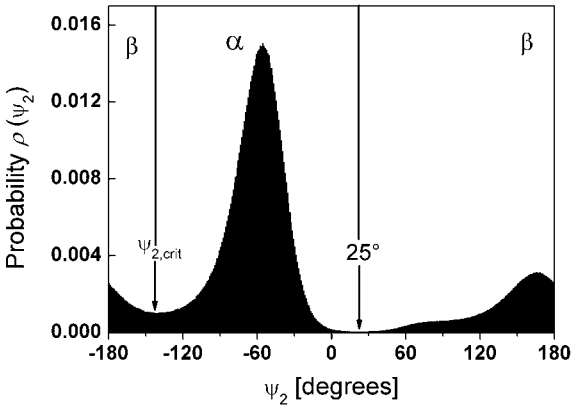

Fig. 2.5: Probability density for the Ramachandran dihedral ψ_2 in simulation condition 8. The torsion angle ψ_2 (see Fig. 2.6) is used as order parameter dividing the conformers α and β. This circular variable delimits the conformers at two positions: $\psi_{2,crit}$ is computed as the region with minimum population near ψ_2=-140°, which varies according to the simulation conditions (See Table 2.1). The second cut position is fixed at ψ_2=25°, since it depends on the repulsive wing of the Lennard Jones potential and is identical for all 13 simulation conditions.

Our trialanine model consists of N = 34 atoms. Its geometry can be described with N–1 = 33 bonds, N–2 = 32 angles and N–3 = 31 torsions yielding a total of 96 BAT coordinates. Furthermore, the torsion angles can be divided into 13 main torsions and 18 associated phase angles.

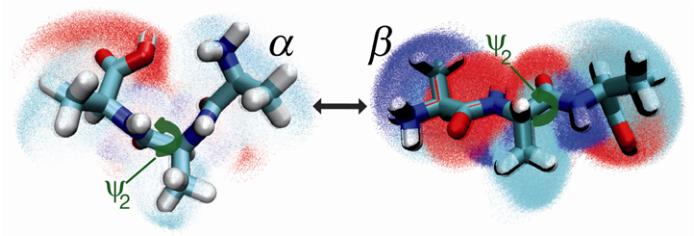

Fig. 2.6: The compact (α) and extended (β) conformers of trialanine. The Ramachandran torsion ψ_2 is used as an order parameter to define the conformers. The compact conformer α has a lower (more favorable) potential energy $U_\alpha < U_\beta$, but also lower (more unfavorable) entropy $s_\alpha < s_\beta$, eq (2.29), than the extended conformer. By how much U and s differ is a function of the surface tension ($\gamma_{H\phi}$) and the scaled $1/r^6$ attractive term of the Lennard Jones potential (ε_{attr}), which were varied in each of the 13 simulations.

2.5.3 Entropy estimation

For the MD simulations of trialanine, larger values of ε_{attr} enhance the attractive wing of the Lennard Jones potential. This leads to more compact conformations ($N^{(\alpha)} > N^{(\beta)}$), and a larger entropy difference $\Delta S_{\beta\alpha}$. Larger surface tension ($\gamma_{H\phi}$) up to a value of 0.045 cal/(mol K Å^2) had a smaller and opposite effect on $\Delta S_{\beta\alpha}$. By varying ε_{attr} and $\gamma_{H\phi}$, different simulation conditions are created, which are then used to test the entropy estimator based on the MI expansion. The order parameter ψ_2 serves to cluster the conformers α and β (see Fig. 2.5 and Fig. 2.6).

2.5.4 Convergence of benchmark entropy, energy and free energy

In order to test our numerical method to estimate entropy, we need reliable benchmarks. Here we show that the trialanine benchmark values are appropriately converged. In Fig. 2.7, we observe that the free energy difference $\Delta F_{\beta\alpha}$ converges the fastest among thermodynamic variables. The energy difference $\Delta U_{\beta\alpha}$ is slower in convergence, and $\Delta S_{\beta\alpha,bench}$, being calculated as a difference, is the slowest one to converge. The simulation condition ID is assigned by ascending values of $\Delta F_{\beta\alpha}$. The order in the values of the energy difference $\Delta U_{\beta\alpha}$ and the entropy difference $\Delta S_{\beta\alpha,bench}$ differs somewhat with respect to the ascending $\Delta F_{\beta\alpha}$ order (see bars on the right of Fig. 2.7.) The free energy $\Delta F_{\beta\alpha}$ is the result of the interplay of energetic and entropic contributions, which are related but not identical. A given potential energy surface (which varies among simulation conditions 1 to 13) determines which microstates are accessible to each conformer at temperature

T. The energetic component results from the microstates' average energy (average "funnel depth"), and the entropic component from their multiplicity (average "funnel width"), adapting the concepts of Wolynes[97] to our system. Thus, there is no a priori reason to believe that the ascending order of the values of energy, entropy and free energy differences should be identical. See color labels in Fig. 2.7a, b and c.

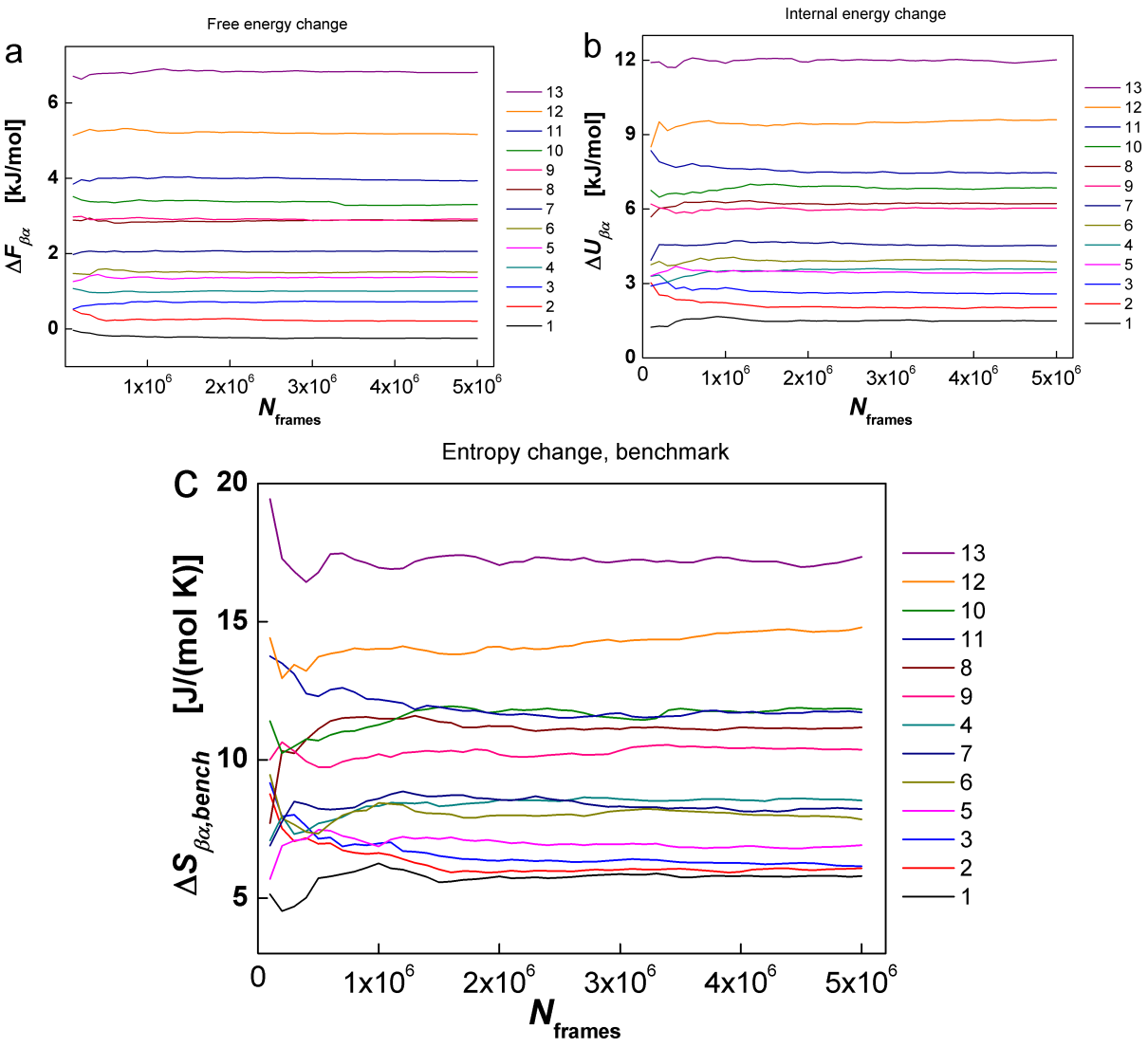

Fig. 2.7: Convergence of the thermodynamic variables in the 1 μs trialanine simulation using $5x10^6$ frames. The frames are used in time order. **a:** Free energy change. **b:** Internal energy change. **c:** Entropy change benchmark.

Krivov et al. simulated tetraalanine[98] with the PARAM19 force field of CHARMM[99] and the ACS[100] implicit solvent model. To evaluate entropy, the tetraalanine conformers were clustered using not a geometric, but a kinetic criterion. The simulation was done both with Langevin dynamics and with a method that confines and explores conformations in a given conformer

basin. They also find that the extended β conformer has higher average energy but is stabilized by entropy. The entropy difference between the helical α and extended β conformations of tetraalanine was found to be $\Delta S_{\beta\alpha}$ = 20.4 J/(mol K)[98], comparable to our results for trialanine, which range from about 5.8 to 17.3 J/(mol K) depending on the simulation conditions.

In Table 2.1 the final asymptotic values for the thermodynamic variables are provided. For each simulation condition, the numerical values for the hydrophobic "surface tension" term $\gamma_{H\phi}$ and the $1/r^6$ attractive Lennard Jones potential scaling factor ε_{attr} used in each simulation can be read. Also, the critical value of ψ_2, a Ramachandran dihedral angle of the middle residue of trialanine[95,96], which we use as order parameter, is provided. $\psi_{2,crit}$ is the value of that angle at which the ensemble population is the lowest, and used to divide the conformers α and β. The second value at which the circular variable ψ_2 is cut is fixed at 25°, a value identical for all simulations. It is the consequence of the repulsive wing of the Lennard Jones potential (identical in all 13 simulations) and physically interpretable as a steric constraint. See Fig. 2.5 for an example of the probability distribution $\rho(\psi_2)$ corresponding to simulation condition ID 8.

Table 2.1: Converged values of the thermodynamic variables for trialanine simulation with 13 different conditions.

Simulation condition	ε_{attr} [dimless]	$\gamma_{H\phi}$ [cal/(mol K Å^2)]	$\Delta F_{\beta\alpha}$ [kJ/mol]	$\Delta U_{\beta\alpha}$ [kJ/mol]	$\Delta S_{\beta\alpha,bench}$ [J/(mol K)]	$\psi_{2,crit}$ [degrees]
1	0.00	0.045	-0.25	1.49	5.80	-134.5
2	0.00	0.025	0.21	2.03	6.08	-138.5
3	0.00	0.000	0.73	2.58	6.15	-140.5
4	0.50	0.045	1.01	3.57	8.54	-135.5
5	0.25	0.000	1.37	3.44	6.92	-139.5
6	0.50	0.025	1.51	3.86	7.85	-139.5
7	0.50	0.000	2.06	4.53	8.23	-141.5
8	1.00	0.045	2.87	6.22	11.17	-140.5
9	0.75	0.000	2.92	6.03	10.37	-140.5
10	1.00	0.025	3.30	6.85	11.83	-141.5
11	1.00	0.000	3.93	7.45	11.72	-141.5
12	1.25	0.000	5.16	9.60	14.79	-144.5
13	1.50	0.000	6.81	12.01	17.34	-144.5

2.5.5 Detailed results for entropy estimates using the MI expansion (MIE)

Here we present the results for entropy estimation with MIE1 (Fig. 2.8), MIE2 (Fig. 2.9) and MIE3 (Fig. 2.10). The four panels of these figures demonstrate the effect of using different correction methods. Upper left: unbalanced, biased; upper right: unbalanced, bias-free; lower left: balanced, biased; lower right: balanced, bias-free. In these figures, we consider all 96 BAT degrees of freedom of the trialanine model: bonds, angles, torsions and if necessary phase angles that replace corresponding torsion angles. The lower right panel (d) presents the best results using both correction methods: balancing and bias-removal. Also shown in each panel is the average and standard deviation of the estimate-to-benchmark ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$. Average and standard deviation for this ratio are calculated over all 13 simulation conditions and all five of histogram schemes with different numbers of bins M.

2.5.5.1 MIE1 using all BAT coordinates

The first order MI expansion (MIE1) in Fig. 2.8 is well converged. Nevertheless, the converged value does not agree well with the benchmarks, as can be seen by the deviation of the computed results from the dashed diagonal line representing the perfect agreement. In MIE1, the individual entropies are estimated as the sum of the marginal entropies (first term of (2.35)). Compensating the bias according to eq (2.38) yields for MIE1 a small correction only, which results in no noticeable change from a→b and c→d in Fig. 2.8. The size of the correction is small because in the 1st order MI expansion the number of histogram bins is small such that the bins are well filled and exhibit small fluctuations. This contrasts with MIE2 and MIE3 having quadratically and cubically as many histogram bins, respectively. Thus, for MIE1 the major correction comes from balancing (a→c and b→d). The balancing method narrows the spread between the estimators for the different number of histogram bins M (different symbols), but as expected cannot correct for the lack of correlation in MIE1.

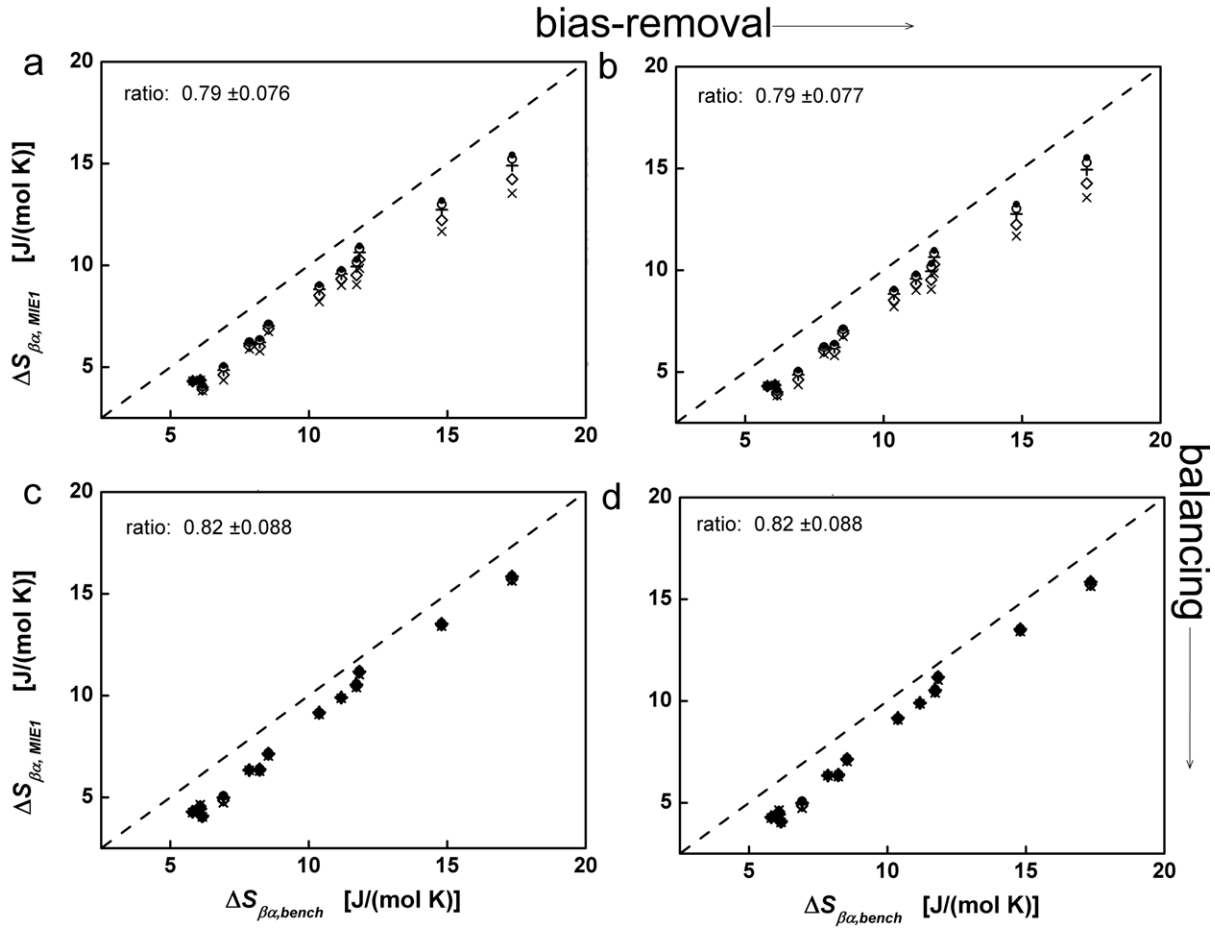

Fig. 2.8: Results with first order MI expansion (MIE1). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇ M=25; + M=35; ○ M=50; ● M=100. The arrows show application of the correction methods: none (a), either (b, c) or both (d). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

2.5.5.2 MIE2 using all BAT coordinates

In Fig. 2.9, we see a large and beneficial effect of the balancing method (a→c and b→d). The bias-removal acts to fine-tune the entropy differences in c→d. It becomes evident that balancing and bias-removal act synergistically to improve the accuracy of the entropy estimates.

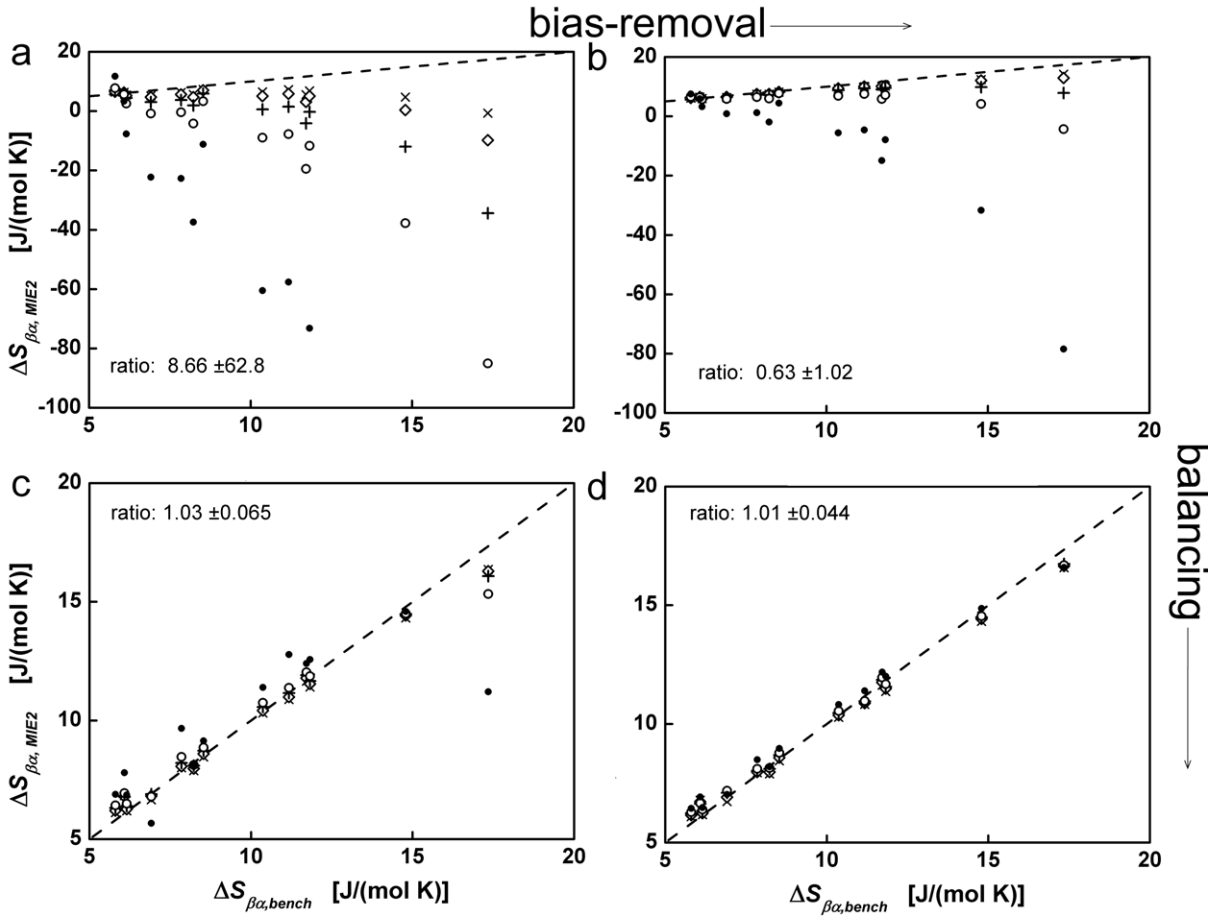

Fig. 2.9: Results with second order MI expansion (MIE2). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇ M=25; + M=35; ○ M=50; ● M=100. The arrows show application of the correction methods: none (**a**), either (**b**, **c**) or both (**d**). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

2.5.5.3 MIE3 using all BAT coordinates

The MIE3 entropy difference estimates in Fig. 2.10 show poor agreement with the benchmarks. There is definite improvement by using bias-removal and balancing, but even Fig. 2.10d where both methods have been used is far from optimal. From this we conclude that we need more frames than the 5×10^6 frames used here to obtain well converged MIE3 estimates.

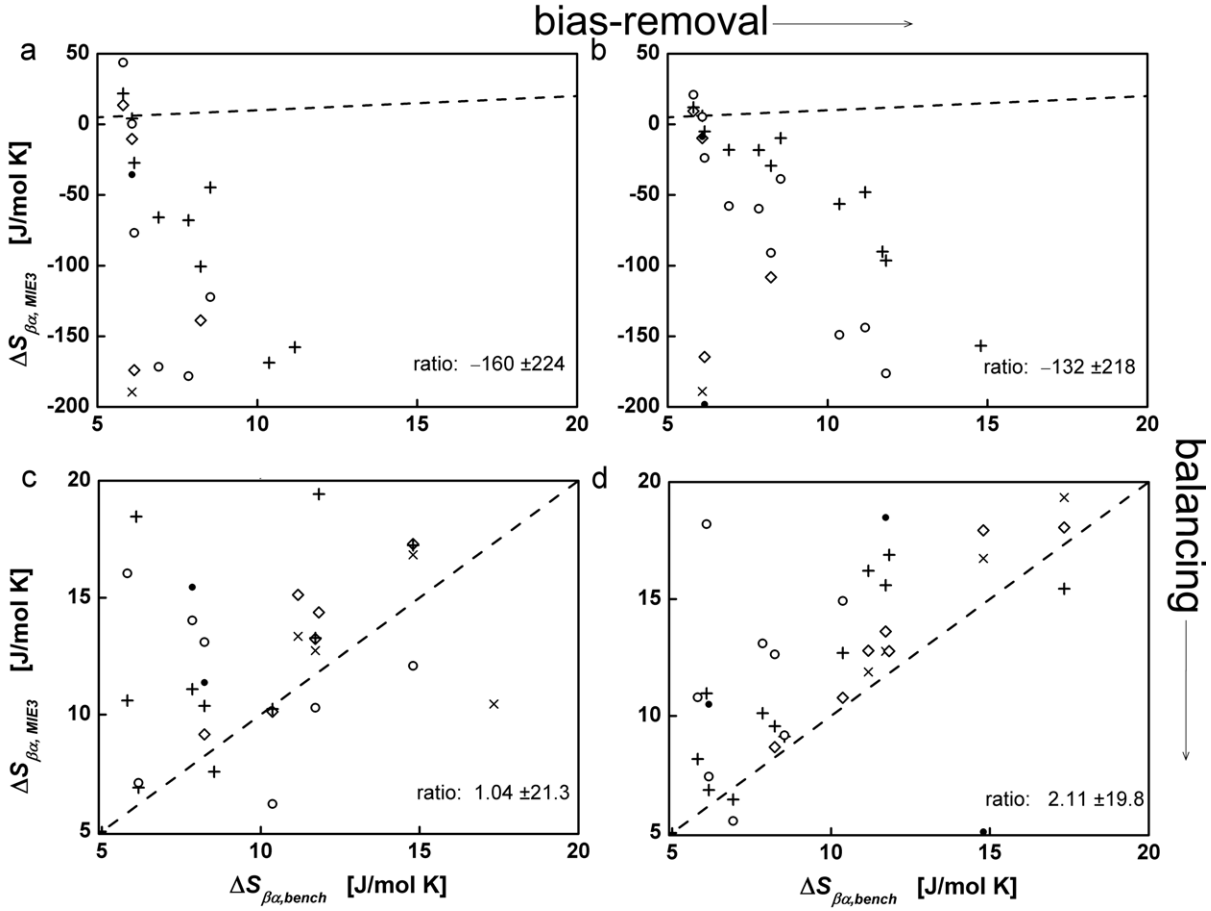

Fig. 2.10: Results with third order MI expansion (MIE3). Entropy difference estimates $\Delta S_{\beta\alpha}$ (abscissa) between the two conformers β and α for the trialanine model are compared with benchmark entropies (ordinate). All 96 BAT degrees of freedom are used. The symbols stand for the number of histogram bins used: × M=20; ◇ M=25; + M=35; ○ M=50; ● M=100. The arrows show application of the correction methods: none (a), either (b, c) or both (d). Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0.

2.5.5.4 MIE Using only soft degrees of freedom

In recent work of Brüschweiler et al.[51], it was suggested to employ the main torsion angles ('soft degrees of freedom') only and to neglect the 'hard degrees of freedom', including phase angles. In their work, Brüschweiler et al. only account for the momenta contributions (cf. second term of eq (2.13)) of the hard degrees of freedom, which is required because the entropy difference is estimated for conformers at two different temperatures (T = 380 K and T = 270 K). They assume that the Jacobian determinant (which only arises from hard degrees of freedom) will be conformation-independent and thus cancel. The validity of this assumption is discussed in section 2.6.1.2, "Approximate cancellation of the Jacobian term of the entropy". The momenta

contributions and the constant Jacobian are embodied into eq (2) of ref 51. Using only torsions as soft degrees of freedom resulted in estimate-to-benchmark ratios between 0.87 and 0.96 when testing entropy differences of dipeptide conformers at two different temperatures (see Table I, last column, of Brüschweiler et al.[51]).

Furthermore, Brüschweiler et al. studied the conformational entropy change between the bound and unbound conformers of a protein[52]. They found that linear correlations (as obtained from the covariance matrix[35]) between torsion angles are fairly similar in the bound and unbound states. Based on this fact, Brüschweiler et al. suggested[52] to neglect correlations between the torsion angles (as estimated from mutual information, which includes non-linear correlations[26]). In defining 'soft degrees of freedom' Brüschweiler et al. considered only one main torsion angle per shared pair of bonds. This is confirmed in the statement that the alanine dipeptide "has a total of 7 soft degrees of freedom"[51]. Translated to our definition of BAT coordinates, trialanine has 13 main torsions. However, trialanine also has 18 associated phase angles, which may or may not count as 'soft degrees of freedom'. The remaining 33 bond lengths and 32 bond angles are considered stiff or 'hard degrees of freedom'. Although their entropy estimation employs different numerical methods[51,52], their results are on similar footing with ours since: (i) They employ (a subset of) BAT coordinates. (ii) Their data are naturally balanced, as their conformers belong to two independent simulations involving the same simulation conditions, from which they likely take the same number of frames for their analysis.

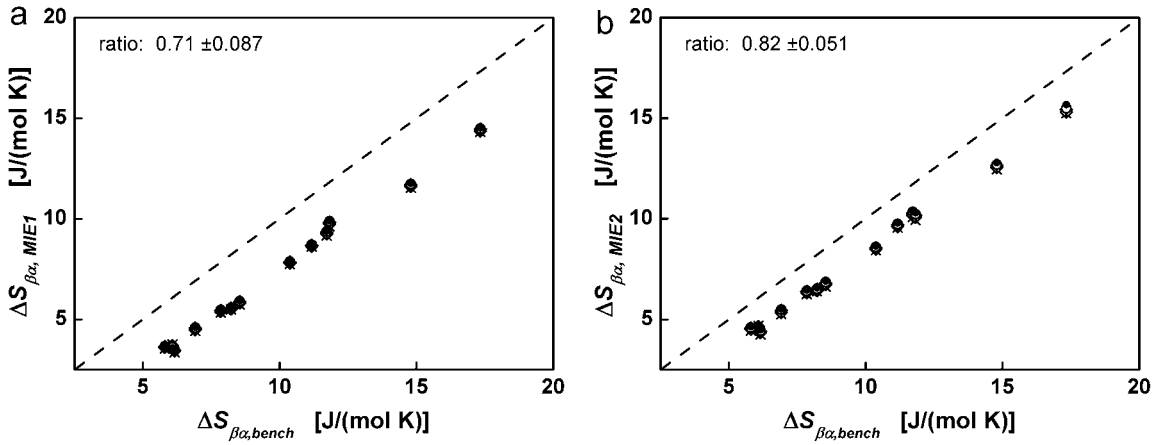

Fig. 2.11: Entropy estimates for the trialanine model using only the main 13 torsion angles as 'soft degrees of freedom', and neglecting the conformational variations of phase angles, bond angles and bond lengths. Both correction methods (balancing and bias-removal) are used, as they yield the best results. Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and

the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0. **a:** First order MI expansion (MIE1); **b:** Second order MI expansion (MIE2).

We applied their suggestions to our trialanine model. In Fig. 2.11a, we follow both suggestions. Using only the main 13 torsions with the 1st order MI expansion (MIE1) yields a low value of the estimate-to-benchmark ratio of 0.71 ±0.087. In Fig. 2.11b, we switch to the 2nd order (MIE2), obtaining a larger estimate-to-benchmark ratio of 0.82 ±0.051. If we now alter the definition of soft degrees of freedom to include all 31 torsion and phase angles, we obtain a ratio of 0.81 ±0.069 for MIE1 and a ratio of 0.97 ±0.027 for MIE2 (Fig. 2.12).

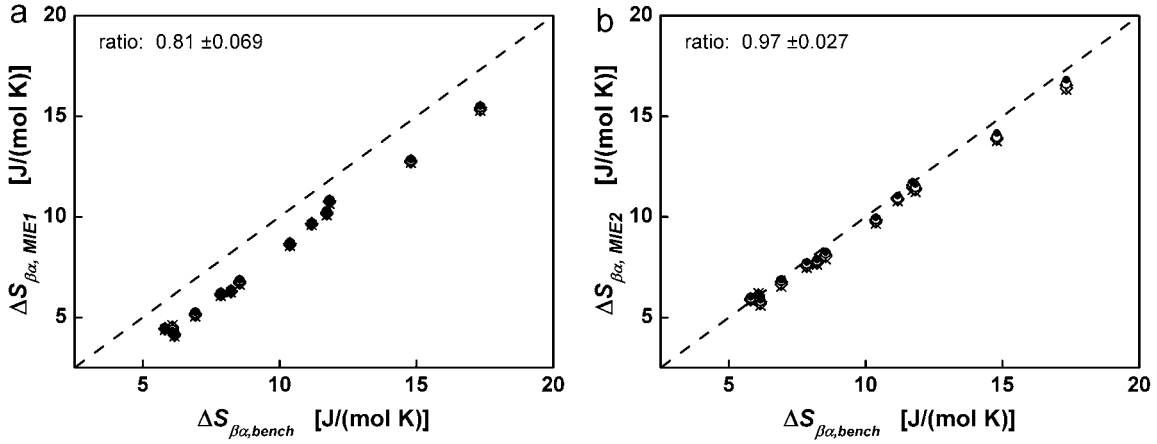

Fig. 2.12: Entropy estimates for the trialanine model using 31 'soft degrees of freedom' (13 torsions and 18 phase angles), and neglecting the conformational variations of angles and bonds. Both correction methods (balancing and bias-removal) are used, as they yield the best results. Also given are average and standard deviations for the ratio $\Delta S_{\beta\alpha,MIE1}$ / $\Delta S_{\beta\alpha,bench}$ of all 13 simulation conditions and the five histogram schemes with different numbers of bins M. The optimal result is 1.0 ±0.0. **a:** First order MI expansion (MIE1); **b:** Second order MI expansion (MIE2).

In summary, the best estimates for trialanine are obtained when applying both correction methods: balancing and bias-removal in the 2nd order MI expansion. Furthermore, using all 96 BAT coordinates with M = 35 bins histogram (Fig. 2.9d) leads to the best estimate-to-benchmark ratio of 1.01 ±0.037. The second best results are obtained using only the 'soft degrees of freedom' defined as the torsion and phase angles (Fig. 2.12b). Note that most data points in Fig. 2.12b are below the identity line (ratios below 1.0), pointing to a slight systematic underestimation of the entropy differences due to small contributions from the hard degrees of freedom.

2.5.6 Convergence of the entropy estimates

In this section, we analyze the convergence properties of entropy and entropy difference estimates using the 2nd order MI expansion (MIE2) and employing both correction methods (balancing and bias-removal). For the sake of clarity, only the final converged benchmark values of the entropy difference $\Delta S_{\beta\alpha,bench}$ (eq (2.40)) are shown as dashed lines in Fig. 2.13 and Fig. 2.14. The convergence properties of the entropy difference for the benchmark values was treated separately in the section 2.5.4 "Convergence of benchmark entropy, energy and free energy".

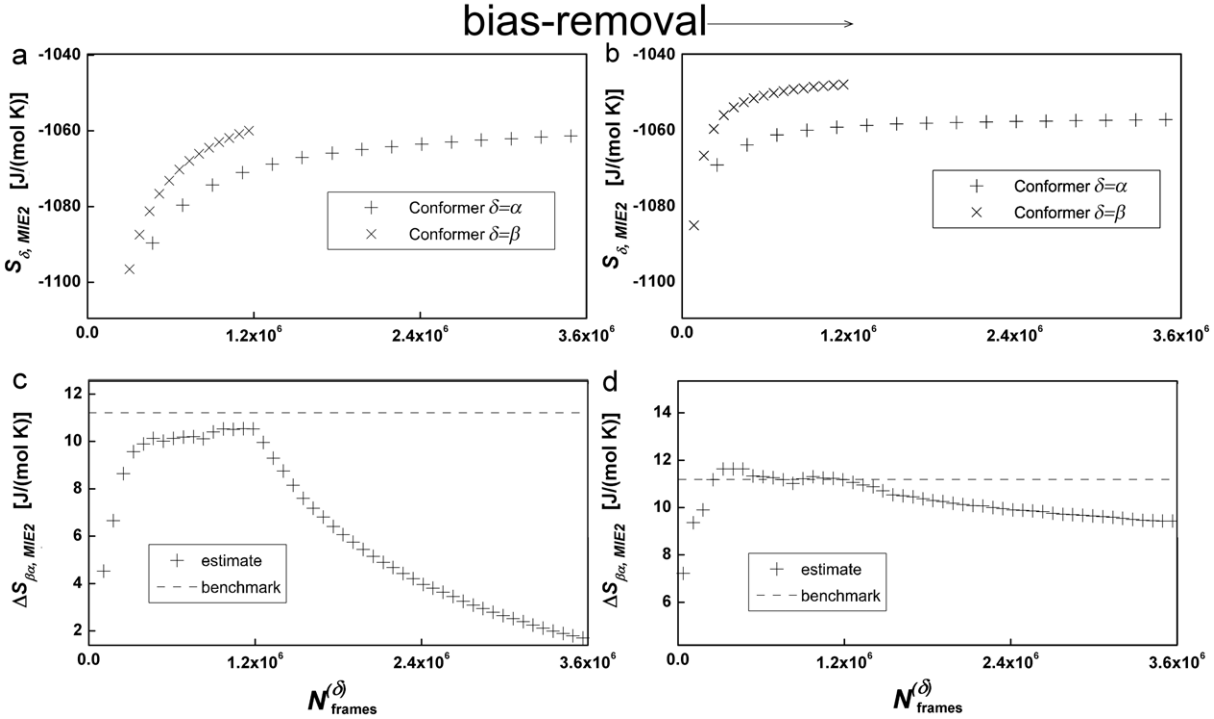

Fig. 2.13: Unbalanced number of frames, 2nd order estimator (MIE2) used to plot individual (relative) entropies S_α, S_β and the entropy difference $\Delta S_{\alpha\beta}$. Convergence of the entropy estimates versus number of frames used for the trialanine simulation condition 8 (parameters: $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00). Frames are used in time order. The abscissa denotes with $N^{(\delta)}_{frames}$ the effective number of frames used for $\delta=\alpha$, β. This differs from N_{frames} used elsewhere, which refers to all frames of the simulation. The dashed line marks the final benchmark value. **a:** Individual conformer entropies without bias-removal. **b:** Individual conformer entropies using bias-removal. **c:** Entropy difference without bias-removal. d: Entropy difference using bias-removal.

In the following discussion we will use the example of trialanine with simulation condition 8 (parameters $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00) using the 2nd order MIE expansion (MIE2). The individual entropies S_δ are not fully converged, whether unbalanced (Fig. 2.13a, b) or balanced data are used (Fig. 2.14a, b), and independently of whether bias removal a→b is

applied or not. As matter of fact, balancing will slow down the convergence of entropies of the majority conformer S_α. However, our main focus is on computing entropy differences $\Delta S_{\beta\alpha}$. There, we observe a beneficial effect of balancing. Without balancing, the entropy difference $\Delta S_{\beta\alpha}$ diverges (Fig. 2.13c, d), while with balancing the entropy difference converges (Fig. 2.14c, d). This is due to the fact that after balancing the individual conformer entropies ($S_{\alpha,MIE2}$ and $S_{\beta,MIE2}$) possess similar systematic errors, which cancel in the entropy difference $\Delta S_{\beta\alpha}$. The bias-removal method c→d provides an additional beneficial fine-tuning for the entropy difference.

2.5.6.1 Importance of choosing frames at random in the balancing method

In the balancing method, only a subset of the frames of the majority conformer is used. It is important to choose those frames at random[101] instead of simply taking a contiguous subset of the trajectory, since that results in a nonequivalent exploration of the phase space. While the convergence of the individual entropies $S_{\delta,MIE2}$ using time order or random order is indistinguishable to the eye due to the large magnitude of the individual entropies (Fig. 2.14a, b), the consequences for the convergence of the entropy difference $\Delta S_{\beta\alpha,MIE2}$ are clearly visible. In Fig. 2.14c, d we see that the convergence of the entropy difference is accelerated by choosing the frames at random. The reason for this does not lie in the numerical properties of the bias of the histogram method, but rather in the fact that the randomly ordered conformations result in a more complete phase space exploration at a given number of frames. Choosing the frames at random is important for MD and MC simulations, where the frames are correlated with each other.

The convergence behavior of $\Delta S_{\beta\alpha,MIE2}$ for all 13 simulation conditions using balancing and bias-removal is presented in Fig. 2.15.

2.5.7 Summary of results for model system 2

We conclude that the best results for estimators of the entropy differences $\Delta S_{\beta\alpha}$ between the two the conformers (α, β) of the trialanine model are obtained using all BAT coordinates in the 2nd order MI expansion. These entropy estimates are well converged () and agree best with the benchmark (Fig. 2.9).

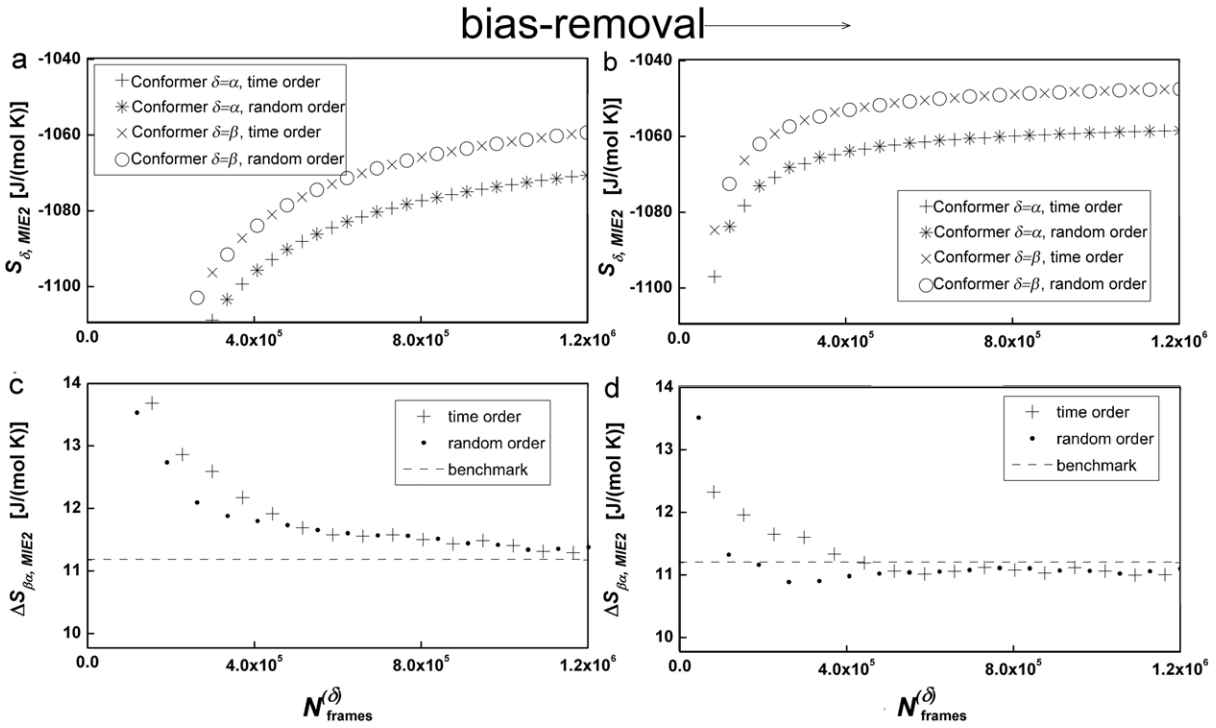

Fig. 2.14: Balanced number of frames, 2nd order estimator (MIE2) used to plot individual entropies S_α, S_β and the entropy difference $\Delta S_{\alpha\beta}$. Convergence of the entropy estimates versus number of frames used for the trialanine simulation condition 8 (parameters: $\gamma_{H\phi}$ = 0.045 cal/(mol K Å^2) and ε_{attr} = 1.00). Frames are used in time and random order as indicated in the figure. The abscissa denotes with $N^{(\delta)}_{frames}$ the effective number of frames used, which is identical for $\delta=\alpha, \beta$ when applying the balancing method. This differs from N_{frames} used elsewhere, which refers to all frames of the simulation. The dashed line marks the final benchmark value. **a:** Individual conformer entropies without bias-removal. **b:** Individual conformer entropies using bias-removal. **c:** Entropy difference without bias-removal. d: Entropy difference using bias-removal.

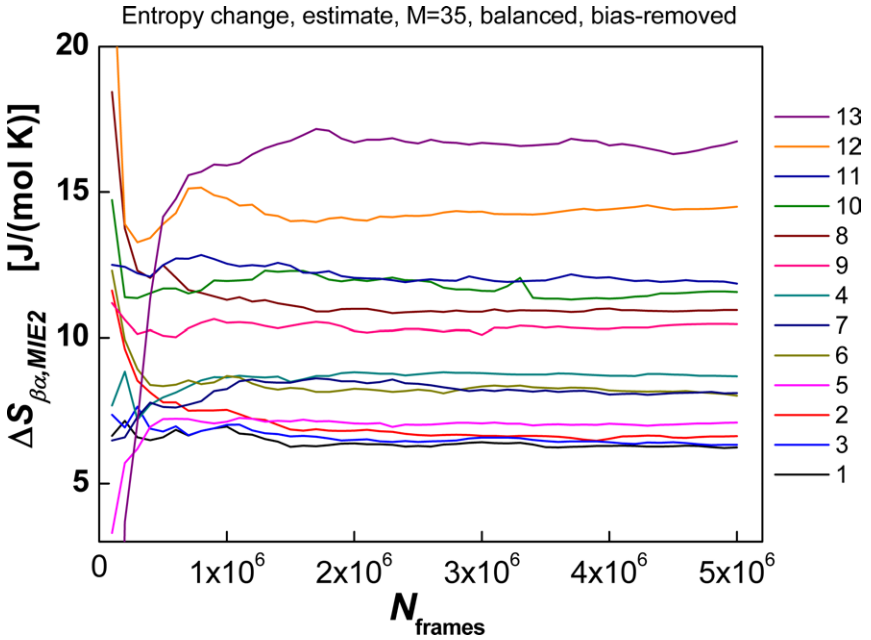

Fig. 2.15: Convergence of the entropy estimates with the second order MI expansion, using balancing and bias-removal. The frames are used in time order.

We show the deviation of the entropy estimates from the benchmark values in Fig. 2.16. The simulation conditions are labeled as 1 to 13, ordered by increasing $\Delta F_{\beta\alpha}$. The simulations with conditions 1 and 2 have vanishing $\Delta F_{\beta\alpha}$, so that $K_{eq} = N^{(\beta)} / N^{(\alpha)} = \exp(-\Delta F_{\beta\alpha}/k_B T) \approx 1$, i.e. the numbers of frames are equal. In other words, the molecular system is naturally balanced. In contrast, the simulation with condition 13 is very unbalanced, with $\Delta F_{\beta\alpha} >> 0$ and $K_{eq} \approx 0.07$, such that there is room for improvement using the balancing method. See caption of Fig. 2.16 for values of parameters and thermodynamic variables for the 13 simulation conditions.

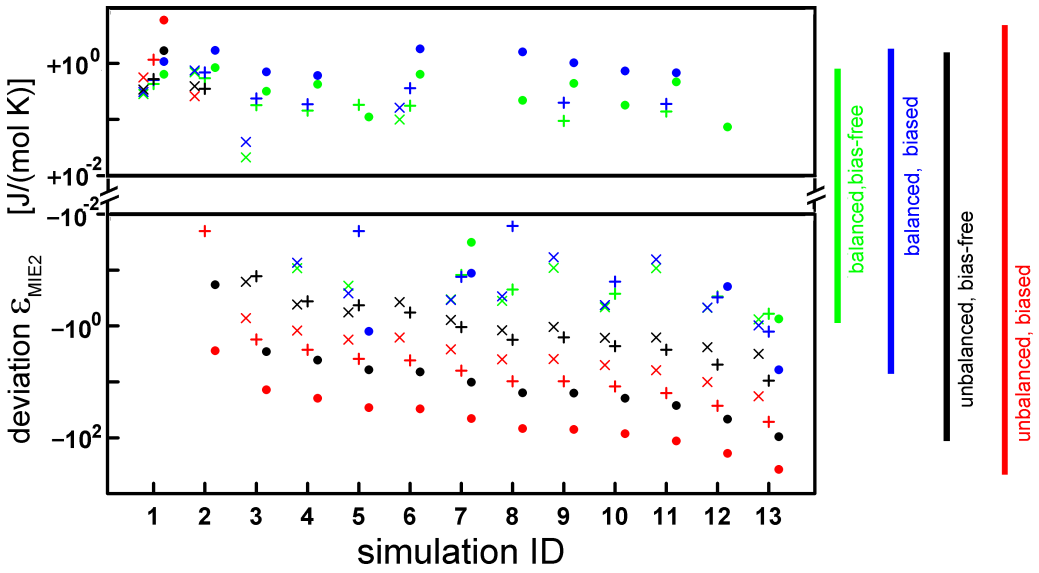

Fig. 2.16: Deviation of the estimated conformational entropy difference $\Delta S_{\beta\alpha}$, eqs (2.35)–(2.37), using the 2nd order MI expansion (MIE2) with all 96 BAT coordinates from the benchmark value, eq (2.40). Based on MD simulations of 1 μs with 5×10^6 frames (coordinate sets) for trialanine. Smaller deviations are for symbols near the center of the discontinuous logarithmic ordinate. The MD simulations with 13 different conditions are ordered by increasing $\Delta F_{\beta\alpha}$. The color labels the correction methods used (see bars on the right). The symbols label the number of bins used in histograms: × M = 20; + M = 35; ● M = 100. It is apparent that the deviation of the estimated $\Delta S_{\beta\alpha}$ is smallest when the estimates are both balanced and bias-free (green). Details of the MD simulation are given in section 2.5.1. Correspondence between simulation condition ID and parameters is as follows: ID 1 (ε_{attr} =0.0, $\gamma_{H\phi}$=0.045 kcal/(mol Å^2)); 2 (0.00, 0.025); 3 (0.00, 0.000); 4 (0.50, 0.045); 5 (0.25, 0.000); 6 (0.50, 0.025); 7 (0.50, 0.000); 8 (1.00, 0.045); 9 (0.75, 0.000); 10 (1.00, 0.025); 11 (1.00, 0.000); 12 (1.25, 0.000); 13 (1.50, 0.000).

The deviations in Fig. 2.16 are plotted for the 2nd order MI expansion including all 96 BAT coordinates. Using the balancing method (green and blue symbols) results in the smallest deviations of $\Delta S_{\beta\alpha}$ from the benchmark values. In particular, combining balancing with bias-removal (green) results in an average absolute deviation of less than 0.3 J/(mol K). Using the balancing method without bias-removal (blue) results in an average deviation of 0.7 J/(mol K), about twice as large. The unbalanced $\Delta S_{\beta\alpha}$ values (black and red) have generally large, negative deviations and a systematic, spurious dependency on $\Delta F_{\beta\alpha}$. When only bias-removal is applied (black), but no balancing, the absolute deviation becomes 7.5 J/(mol K). The red symbols in Fig. 2.16 represent the estimates of $\Delta S_{\beta\alpha}$, where no corrections are applied, corresponding to the original method by Gilson et al[59,60]. In this case, the entropy difference has not converged using all available $N^{(\alpha)} + N^{(\beta)} = 5 \times 10^6$ frames. The average absolute deviation of the estimated entropy difference from the benchmark value in absence of any corrections is 32 J/(mol K), which is about 100 times larger than the corresponding results obtained applying both correction methods.

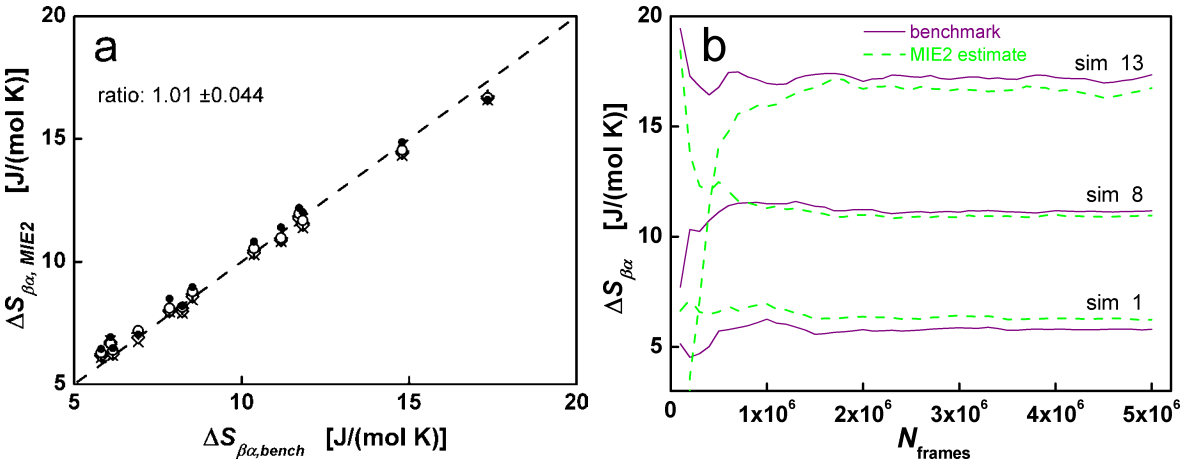

Fig. 2.17: Entropy difference $\Delta S_{\beta\alpha}$ for the trialanine model system using all 96 BAT coordinates (bonds, bond angles, torsion angles and phase angles) and considering 5×10^6 frames, which are in time order. **a:** Influence of the number of histogram bins, M, on the estimated entropy difference of the 2nd order MI expansion (MIE2), plotted versus the corresponding benchmark values for the 13 different MD simulation conditions applying both corrections: balancing and bias-removal. The number of histogram bins M was varied: × M = 20; ◇ M = 25; + M = 35; ○ M = 50; ● M = 100. The dashed diagonal line corresponds to perfect agreement between benchmark and estimate of entropy difference. Also given is the average and standard deviation of the ratio $\Delta S_{\beta\alpha,MIE} / \Delta S_{\beta\alpha,bench}$ over all five M values and data from all 13 simulation

conditions (optimal result is 1.0 ±0.0). **b:** Convergence of the benchmark and the MIE2 estimate as a function of the number of frames N_{frames} using M = 35 histogram bins. For the sake of clarity, only three representative simulations (1, 8, 13) are shown. See caption of Fig. 2.16 for correspondence between simulation (sim) condition ID and their parameters.

To obtain close agreement with the benchmark values for the trialanine model system (see Fig. 2.17a) it was necessary to include all 96 BAT coordinates and pair correlations between them, as implemented in the 2nd order MI expansion of the entropy differences. The estimate-to-benchmark ratio $\Delta S_{\beta\alpha,MIE2}$ / $\Delta S_{\beta\alpha,bench}$ (see Fig. 2.17a) was found to be 1.01 ± 0.044 when averaged over all five histogram schemes M (M= 20,25, 35, 50, 100) and all 13 simulation conditions. In Fig. 2.17b, we see that both benchmark (solid line) and estimated (dashed line) entropy differences are asymptotically converged, with the benchmark converging more quickly. This is shown for three examples in Fig. 2.17b (and for all examples in Fig. 2.7 and Fig. 2.15). The 1st order MI expansion (Fig. 2.8) converges much more quickly than the 2nd order, but the entropies $\Delta S_{\beta\alpha}$ obtained with the 1st order MI expansion have an estimate-to-benchmark ratio of 0.82 ± 0.051 (corresponding to 1 − 0.82 = 18% average underestimation; see Table 2.2). The 3rd order MI expansion does not converge for the available 5×10^6 frames, and would likely require at least one order of magnitude more frames (Fig. 2.10). For more information, see section 2.5.5 “Detailed results for entropy estimates using the MI expansion (MIE)”.

We follow the suggestions of Brüschweiler et al. and employ only the main torsion angles[51] in the first order MI expansion[52]. For more details, see section 2.5.5.4 “MIE Using only soft degrees of freedom”. Neglecting 33 bonds, 32 bond angles and 18 phase angles, 13 main torsion angles remain for the trialanine model involving 34 atoms. Applying the 1st order MI expansion with M = 35, the estimate-to-benchmark ratio averaged over all 13 simulation conditions is 0.71 ±0.089. Including also pairwise correlations by using the 2st order MI expansion raises the average ratio to 0.82 ±0.051. If we now redefine the ‘soft degrees of freedom’ to include not only the main 13 torsions but also the 18 phase angles, we obtain ratios of 0.81 ± 0.071 (for MIE1) and 0.97 ± 0.024 (for MIE2), which are much closer to unity. All data are summarized in Table 2.2.

Table 2.2: Averages and standard deviations for the estimate-to-benchmark ratio $\Delta S_{\beta\alpha,MIE} / \Delta S_{\beta\alpha,bench}$ over all 13 simulation conditions using histograms with M = 35 bins. The optimal result is 1.0 ± 0.0. The entropy estimates were computed using the 1st and 2nd order MI expansion (MIE1 and MIE2) applying both correction methods (balancing and bias-removal). The estimate-to-benchmark ratios vary for the different coordinate sets and orders of the MI expansion used. Best results are obtained with MIE2 using all 96 BAT coordinates, and second best results are for the 31 'soft degrees of freedom' (13 torsions and 18 phase angles).

coordinate set	order of MI expansion	
	MIE1	MIE2
13 main torsion angles	0.71 ± 0.089	0.82 ± 0.051
13 torsion and 18 phase angles	0.81 ± 0.071	0.97 ± 0.024
all 96 BAT coordinates	0.82 ± 0.091	1.01 ± 0.037

2.6 Discussion

2.6.1.1 BAT coordinates represent phase space compactly

Internal BAT coordinates allow a compact representation of the available conformational volume of a molecule. An alternative and complementary view of entropy to the *missing information* is *a measure of the phase space volume* occupied by a certain state (see Sec. 27.3 in ref 102) in the canonical ensemble. The mobility of hydrogen atoms of a methyl group in internal BAT coordinates is characterized mainly by a single dihedral angle, while the remaining two phase angles are less important, since they belong to the stiffer degrees of freedom. All other degrees of freedom of the methyl group describe small amplitude vibrations in three bond angles and three bond lengths. Alternatively, the Cartesian representation requires nine geometrically highly correlated coordinates, all of which involve large amplitude motions. Even after applying PCA OR QHA, such correlations persist[43] for polypeptide chains. Internal BAT coordinates avoid such spurious correlations inherent to Cartesian coordinates and are therefore more suitable to describe the relevant correlations of motion in a molecule, thus yielding improved entropy estimates if the MI expansion is used.

2.6.1.2 Approximate cancellation of the Jacobian term of the entropy

In our calculations we consistently employ the factors based on the Jacobian determinant, so that the entropy in terms of internal BAT coordinates, eq (2.29), yields formally the same entropy as in Cartesian coordinates, eq (2.15). The influence of the Jacobian does not in general cancel for entropy differences (see theorem 1.3.2 in ref 61). Nevertheless, we have noticed that there is an approximate cancellation of the Jacobian contributions in the calculation of entropy differences between conformers of the same molecule if the MI expansion is used. This can be rationalized as follows: (1) The Jacobian term plays no role in 2nd, 3rd or higher order terms of the MI expansion (see appendix of ref 103). (2) Most of the entropy difference is due to torsions, for which the Jacobian is unity (see eq (2.33)). (3) The only Jacobian contributions to entropy differences are due to the 1-dimensional bond and bond angle entropies. The probability densities for these coordinates experience only small changes between conformers, so the Jacobian term in the entropy difference will vanish. Similar conclusions about bond lengths and angles have been reached by others[54].

The case of non-covalent bonds in the calculation of the entropy of receptor-ligand binding[60] should nevertheless be treated separately. The relative motion of ligand and receptor (residual translation and libration in the bound conformation) can be characterized using three torsions, two angles and one bond degrees of freedom. The factors from the Jacobian determinant associated with the angles and the bond may be extremely different in the bonded and non-bonded states, as the variability of such bonds and angles is much wider than in a covalent bond. As such, these Jacobian determinants should always be treated explicitly, as their influence does not cancel in entropy differences. This was indeed the case in section 2.4 “Model system 1: Monte Carlo simulation of a three-atom molecule in a cage”, where bond angles have large variations and the Jacobian term is essential for obtaining a thermodynamically relevant[104] entropy difference.

2.6.1.3 Entropy estimation in signal processing versus molecular simulations

The signal processing community has designed a wealth of approaches to estimate entropy from samples of time series. They include histogram methods[80,105], kernel density estimators[106] and the k-nearest-neighbor approach[26,35,103,107-109]. For a finite number of samples, all entropy estimators

suffer from statistical and systematic biases[79,80]. The systematic bias can be understood intuitively because entropy is a sensitive measure of the *variability* of a probability density, and a finite sample will tend to underestimate this variability. A major focus in signal processing[110] is to estimate entropy with mutual information (MI) estimators for a small number of variables (around 10) and a small number of samples (about 10^3). Entropy estimation for molecular simulation data presents a different type of challenge, since we compute entropy differences considering molecular systems involving 10^2 or more atomic coordinates, where one needs samples of 10^5 or more independent coordinate frames. In this study, we provide evidence that adequately bias-free and balanced histogram-based entropy estimators work best for data from molecular simulation. At the same time, its simplicity makes this method computationally more efficient than others like the k-nearest-neighbor approach.

2.7 Conclusion

In this work, trialanine, a small test model molecule, was used to prove that the 2nd order MI expansion, in conjunction with balancing and bias-removal corrections, allows for proper convergence of the entropy difference $\Delta S_{\beta\alpha}$. This is the case even though the individual conformational entropies s_α and s_β, eq (2.29), are not converged (see section 2.5.6 "Convergence of the entropy estimates"). Notwithstanding, the estimated values of $\Delta S_{\beta\alpha}$ are in excellent agreement with the corresponding benchmark values.

The use of local spherical polar coordinates[65-67], the so-called BAT coordinates[69], enables a clear-cut separation of global translation and rotation from the internal degrees of freedom. In the quasi-harmonic approximation[30,31] and other approaches[53], the rigid rotor approximation[111] is often used to remove the translational and rotational degrees of freedom. Unfortunately, the rigid rotor introduces spurious mass dependencies[32,112] and correlations between external and internal degrees of freedom, which can be avoided by using BAT coordinates. The BAT coordinates are also adept at describing internal motions of polypeptides for numerical computations of entropy, since this coordinate system minimizes spurious geometric correlations between molecular coordinates.

Without the balancing and bias-removal corrections, the method has been used before[59,60]. Here, we demonstrated that the uncorrected estimate converges when using a much larger sample size (see Fig. 2.3). However, balancing and bias-removal corrections accelerate convergence in a synergistic fashion and enable a more efficient use of the available simulated frames. The balancing method allows for a more efficient systematic cancellation of sampling errors in entropy differences, which works well even if the individual entropy contributions are poorly converged. Applied simultaneously with balancing, the bias-removal method compensates systematic bias due to a limited sample size.

However, just paying attention to the convergence of an entropy estimator does not guarantee that the algorithm works properly and that the results are reliable. In the test phase of an algorithm to compute entropies, a careful comparison with benchmark values is necessary before one can consider applying it to larger macromolecules, where benchmark values are not easily available. This is the purpose of the present study. Such comparisons have been done before[47,51,59], proceeding then to calculate entropy for large molecular systems. We show here that the converged entropy differences obtained with the balanced and bias-free histogram method agree well with thermodynamic benchmarks. For the trialanine model system, the conformational entropy estimates agree with benchmarks to an average deviation of 0.3 J/(mol K), or alternatively an estimate-to-benchmark ratio of 1.01 ±0.037 (see Table 2.2). A small standard deviation and an average estimate-to-benchmark ratio close to unity together indicate converged estimates and a thermodynamically relevant result.

We tested the suggestions of Brüschweiler et al.[51,52] of just using the main torsion angles (excluding phase angles, bond angles and bonds), as well as using the 1st order MI expansion only. For trialanine, this resulted in an estimate-to-benchmark ratio of only 0.71 ±0.089. However, following this line of thought and using only the main torsions and phase angles in the 2nd order MI expansion yielded results almost as good as those of the full BAT coordinate set, with an estimate-to-benchmark ratio of 0.97 ±0.024. This reduced set of coordinates is only about 1/3 of the full set of BAT coordinates. As a consequence, the computational cost for the 2nd order MI expansion is reduced to about 1/9th.

When estimating conformational entropy differences with the methods presented here, the following guidelines are important: (i) The molecular dynamics trajectories need to be long enough to provide a Boltzmann distribution of equilibrated microstates representative for the considered conformer macrostates; (ii) The full set of BAT coordinates, or alternatively only the torsion and phase angles (plus any external bonds and angles analog to the ones we defined in Fig. 2.2) need to be considered; (iii) To perform a 2nd order MI expansion, necessary to achieve a sufficient accuracy of a few percent, a trajectory with a large number of independent frames is needed (at least 10^5 frames with random frame selection for balancing); (iv) Avoid using more frames for the dominant molecular conformer, since the best bias cancellation in entropy differences occurs when the number of frames is balanced; (v) When balancing requires considering only a subset of the total number of frames, select the frames randomly from the whole trajectory to utilize the conformational space explored by the simulation as completely as possible.

Since entropy estimators are generally biased, the balancing method presented here is likely also applicable for algorithms estimating entropy differences by methods other than histogram binning. The complete method, including 1st to 3rd order MI expansion, balancing and bias-removal can be performed with the program ENTROPICAL, which can be obtained from the author and used with CHARMM and NAMD topologies and trajectories.

> "X-ray structures of proteins are like a tree in winter, beautiful in its stark outline but lifeless in appearance. Molecular dynamics gives life to this structure by clothing the branches with leaves that flutter because of the thermal winds."
>
> Claude Poyart, 1988 paraphrased by Martin Karplus, *Spinach on the Ceiling: A Theoretical Chemist's Return to Biology (Autobiography).* Annu. Rev. Biophys. Biomol. Struct. **2006**, *35*, 1- 47.

2.8 References for Chapter 2

(1) Jaynes, E. T., *Where do we stand on maximum entropy?* In *The Maximum Entropy Formalism -R. Levine, M. Tribus (Eds.)*; MIT Press, Cambridge, MA: **1979**. ISBN:978-0262120807

(2) Jaynes, E. T., *Information Theory and Statistical Mechanics (part 1).* Phys. Rev. **1957**, *106*, 620-630. dx.doi.org/doi:10.1103/PhysRev.106.620

(3) Ben-Naim, A. *A Farewell To Entropy: Statistical thermodynamics based on information*; World Scientific Publishing Company: Singapore, **2008**. ISBN:978-9812707079

(4) Clausius, R., *Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie.* Ann. der Physik **1865**, *201*, 353-400. dx.doi.org/doi:10.1002/andp.18652010702

(5) Shannon, C. E.; Weaver, W., *A mathematical theory of communication.* Bell Syst. Tech. J **1948**, *27*, 379-423 dx.doi.org/doi:10.1145/584091.584093

(6) Srinivasan, J.; Cheatham_III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., *Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices.* J. Am. Chem. Soc. **1998**, *120*, 9401-9409. dx.doi.org/doi:10.1021/ja981844+S0002-7863(98)01844-7

(7) Watanabe, H.; Tanaka, S.; Okimoto, N.; Hasegawa, A.; Taiji, M.; Tanida, Y.; Mitsui, T.; Katsuyama, M.; Fujitani, H., *Comparison of binding affinity evaluations for FKBP ligands with state-of-the-art computational methods: FMO, QM/MM, MM-PB/SA and MP-CAFEE approaches.* Chem-Bio Inf. J. **2010**, *10*, 32-45. dx.doi.org/doi:10.1273/cbij.10.32

(8) Polyansky, A. A.; Zubac, R.; Zagrovic, B., *Estimation of Conformational Entropy in Protein–Ligand Interactions: A Computational Perspective.* Meth. Mol. Biol. **2012**, *819*, 327-353. dx.doi.org/doi:10.1007/978-1-61779-465-0_21

(9) Homeyer, N.; Gohlke, H., *Free Energy Calculations by the Molecular Mechanics Poisson–Boltzmann Surface Area Method.* Mol. Inf. **2012**, *In Print.* dx.doi.org/doi:10.1002/minf.201100135

(10) Salwiczek, M.; Samsonov, S.; Vagt, T.; Nyakatura, E.; Fleige, E.; Numata, J.; Cölfen, H.; Pisabarro, M. T.; Koksch, B., *Position dependent effects of fluorinated amino acids on hydrophobic core formation of a coiled coil heterodimer.* Chem. Eur. J. **2009**, *15*, 7628-7636. dx.doi.org/doi:10.1002/chem.200802136

(11) Thompson, J. B.; Hansma, H. G.; Hansma, P. K.; Plaxco, K. W., *Backbone Conformational Entropy of Protein Folding: Experimental Measures from Atomic Force Microscopy.* J. Mol. Biol. **2002**, *322*, 645-652. dx.doi.org/doi:10.1016/S0022-2836(02)00801-X

(12) Makhatadze, G. I.; Privalov, P. L., *Hydration effects in protein unfolding.* Biophys. Chem. **1994**, *51*, 291-309. dx.doi.org/doi:10.1016/0301-4622(94)00050-6

(13) Stone, M. J., *NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding.* Acc. Chem. Res. **2001**, *34*, 379-388. dx.doi.org/doi:10.1021/ar000079c

(14) Brüschweiler, R.; Case, D. A., *Collective NMR relaxation model applied to protein dynamics.* Phys. Rev. Lett. **1994**, *72*, 940-943. dx.doi.org/doi:10.1103/PhysRevLett.72.940

(15) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J., *Conformational entropy in molecular recognition by proteins.* Nature **2007**, *448*, 325-330. dx.doi.org/doi:10.1038/nature05959

(16) Diehl, C.; Engstrm, O.; Delaine, T.; Håkansson, M.; Genheden, S.; Modig, K.; Leffler, H.; Ryde, U.; Nilsson, U. J.; Akke, M., *Protein Flexibility and Conformational Entropy in Ligand Design Targeting the Carbohydrate Recognition Domain of Galectin-3.* J. Am. Chem. Soc. **2010**, *132*, 14577-14589. dx.doi.org/doi:10.1021/ja105852y

(17) Marlow, M. S.; Dogan, J.; Frederick, K. K.; Valentine, K. G.; Wand, A. J., *The role of conformational entropy in molecular recognition by calmodulin.* Nat. Chem. Biol. **2010**, *6*, 352-358. dx.doi.org/doi:10.1038/nchembio.347

(18) Süel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat. Struct. Biol. **2003**, *10*, 56-59. dx.doi.org/doi:10.1038/nsb881

(19) Hammes-Schiffer, S., *Enzyme Motions Inside and Out.* Science **2006**, *312*, 208-209. dx.doi.org/doi:10.1126/science.1127654

(20) DuBay, K. H.; Bothma, J. P.; Geissler, P. L., *Long-Range Intra-Protein Communication Can Be Transmitted by Correlated Side-Chain Fluctuations Alone.* PLoS Comput. Biol. **2011**, *7*, e1002168. dx.doi.org/doi:10.1371/journal.pcbi.1002168.g004

(21) Tzeng, S.-R.; Kalodimos, C. G., *Protein dynamics and allostery: an NMR view.* Curr. Opin. Struct. Biol. **2011**, *21*, 62-67. dx.doi.org/doi:10.1016/j.sbi.2010.10.007

(22) Fenwick, R. B.; Esteban-Martin, S.; Richter, B.; Lee, D.; Walter, K. F. A.; Milovanovic, D.; Becker, S.; Lakomek, N. A.; Griesinger, C.; Salvatella, X., *Weak Long-Range Correlated Motions in a Surface Patch of Ubiquitin Involved in Molecular Recognition.* J. Am. Chem. Soc. **2011**, *133*, 10336-10339. dx.doi.org/doi:10.1021/ja200461n

(23) Brüschweiler, R., *Protein dynamics: Whispering within.* Nat. Chem. **2011**, *3*, 665-666. dx.doi.org/doi:10.1038/nchem.1124

(24) Calandrini, V.; Abergel, D.; Kneller, G. R., *Protein dynamics from a NMR perspective: Networks of coupled rotators and fractional Brownian dynamics.* J. Chem. Phys. **2008**, *128*, 145102. dx.doi.org/doi:10.1063/1.2894844

(25) Lange, O. F.; Grubmüller, H., *Generalized Correlation for Biomolecular Dynamics.* Proteins: Struct., Funct., Bioinf. **2006**, *62*, 1053-1061. dx.doi.org/doi:10.1002/prot.20784

(26) Numata, J.; Ebenhöh, O.; Knapp, E. W., *Measuring correlations in metabolomic networks with mutual information.* Genome Inform. **2008**, *20*, 112-122. dx.doi.org/doi:10.1142/9781848163003_0010

(27) Meirovitch, H., *Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation.* Curr. Opin. Struct. Biol. **2007**, *17*, 181-186. dx.doi.org/doi:10.1016/j.sbi.2007.03.016

(28) Stern, O., *Ueber eine Methode zur Berechnung der Entropie von Systemen elastisch gekoppelter Massenpunkte.* Ann. der Physik **1916**, *356*, 237-260. dx.doi.org/doi:10.1002/andp.19163561902

(29) Schlitter, J., *Estimation of absolute and relative entropies of macromolecules using the covariance matrix.* Chem. Phys. Lett. **1993**, *215*, 617-621. dx.doi.org/doi:10.1016/0009-2614(93)89366-P

(30) Schäfer, H.; Mark, A. E.; van_Gunsteren, W. F., *Absolute entropies from molecular dynamics simulation trajectories.* J. Chem. Phys. **2000**, *113*, 7809-7817. dx.doi.org/doi:10.1063/1.1309534

(31) Andricioaei, I.; Karplus, M., *On the calculation of entropy from covariance matrices of the atomic fluctuations.* J. Chem. Phys. **2001**, *115*, 6289-6292. dx.doi.org/doi:10.1063/1.1401821

(32) Carlsson, J.; Åqvist, J., *Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding.* J. Phys. Chem. B **2005**, *109*, 6448-6456. dx.doi.org/doi:10.1021/jp046022f

(33) Harris, S. A.; Laughton, C. A., *A simple physical description of DNA dynamics: Quasi-harmonic analysis as a route to the configurational entropy.* J. Phys.: Condens. Matter **2007**, *19*, 076103. dx.doi.org/doi:10.1088/0953-8984/19/7/076103

(34) Rojas, O. L.; Levy, R. M.; Szabo, A., *Corrections to the quasiharmonic approximation for evaluating molecular entropies.* J. Chem. Phys. **1986**, *85*, 1037-1043. dx.doi.org/doi:10.1063/1.451296

(35) Numata, J.; Wan, M.; Knapp, E. W., *Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation.* Genome Inform. **2007**, *18*, 192-205. dx.doi.org/doi:10.1142/9781860949920_0019

(36) Baron, R.; Hünenberger, P. H.; McCammon, J. A., *Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties.* J. Chem. Theory Comput. **2009**, *5*, 3150-3160. dx.doi.org/doi:10.1021/ct900373z

(37) Hensen, U.; Lange, O. F.; Grubmüller, H., *Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach.* PLoS One **2010**, *5*, e9179. dx.doi.org/doi:10.1371/journal.pone.0009179

(38) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; Wiley-Interscience: New York, **2001**. ISBN:978-0471405405

(39) Matsuda, H., *Physical nature of higher-order mutual information: Intrinsic correlations and frustration* Phys. Rev. E **2000**, *3*, 3096-3102. dx.doi.org/doi:10.1103/PhysRevE.62.3096

(40) Noy, A.; Pérez, A.; Lankas, F.; Luque, F. J.; Orozco, M., *Relative Flexibility of DNA and RNA: a Molecular Dynamics Study.* J. Mol. Biol. **2004**, *343*, 627-638. dx.doi.org/doi:10.1016/j.jmb.2004.07.048

(41) Amadei, A.; Linssen, A.; Berendsen, H., *Essential dynamics of proteins.* Proteins 1993 **1993**, *17*, 412-425. dx.doi.org/doi:10.1002/prot.340170408

(42) Mukherjee, A., *Entropy Balance in the Intercalation Process of an Anti-Cancer Drug Daunomycin.* J. Phys. Chem. Lett. **2011**, *2*, 3021-3026. dx.doi.org/doi:10.1021/jz2013566

(43) Chang, C.-E.; Chen, W.; Gilson, M. K., *Evaluating the Accuracy of the Quasiharmonic Approximation.* J. Chem. Theory Comput. **2005**, *1* 1017-1028. dx.doi.org/doi:10.1021/ct0500904

(44) Mendez, R.; Bastolla, U., *Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins.* Phys. Rev. Lett. **2010**, *104*, 228103. dx.doi.org/doi:10.1103/PhysRevLett.104.228103

(45) Karplus, M.; Kushick, J. N., *Method for estimating the configurational entropy of macromolecules.* Macromolecules **1981**, *14*, 325-332. dx.doi.org/doi:10.1021/ma50003a019

(46) Nola, A. D.; Berendsen, H. J. C.; Edholm, O., *Free energy determination of polypeptide conformations generated by molecular dynamics.* Macromolecules **1984**, *17*, 2044-2050. dx.doi.org/doi:10.1021/ma00140a029

(47) Harpole, K. W.; Sharp, K. A., *Calculation of Configurational Entropy with a Boltzmann-Quasiharmonic Model: The Origin of High-Affinity Protein-Ligand Binding.* J. Phys. Chem. B **2011**, *115*, 9461-9472. dx.doi.org/doi:10.1021/jp111176x

(48) Darian, E.; Hnizdo, V.; Fedorowicz, A.; Singh, H.; Demchuk, E., *Estimation of the Absolute Internal-Rotation Entropy of Molecules with Two Torsional Degrees of Freedom from Stochastic Simulations.* J. Comput. Chem. **2005**, *26*, 651-660. dx.doi.org/doi:10.1002/jcc.20198

(49) Wang, J.; Brüschweiler, R., *2D Entropy of Discrete Molecular Ensembles.* J. Chem. Theory Comput. **2006**, *2*, 18-24. dx.doi.org/doi:10.1021/ct050118b

(50) Li, D.-W.; Khanlarzadeh, M.; Wang, J.; Huo, S.; Brüschweiler, R., *Evaluation of Configurational Entropy Methods from Peptide Folding-Unfolding Simulation.* J. Phys. Chem. B **2007**, *111*, 13807-13813. dx.doi.org/doi:10.1021/jp075220e

(51) Li, D. W.; Brüschweiler, R., *In silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides.* Phys. Rev. Lett. **2009**, *102*, 118108. dx.doi.org/doi:10.1103/PhysRevLett.102.118108

(52) Li, D.-W.; Showalter, S. A.; Bruschweiler, R., *Entropy Localization in Proteins.* J. Phys. Chem. B **2010**, *114*, 16036-16044. dx.doi.org/doi:10.1021/jp109908u

(53) Suárez, E.; Díaz, N.; Suárez, D., *Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations.* J. Chem. Theory Comput. **2011**, *7*, 2638-2653. dx.doi.org/doi:10.1021/ct200216n

(54) Cheluvaraja, S.; Meirovitch, H., *Calculation of the entropy and free energy of peptides by molecular dynamics simulations using the hypothetical scanning molecular dynamics method.* J. Chem. Phys. **2006**, *125*, 024905. dx.doi.org/doi:10.1063/1.2208608

(55) Meirovitch, H., *Methods for calculating the absolute entropy and free energy of biological systems based on ideas from polymer physics.* J. Mol. Recognit. **2010**, *23*, 153-172. dx.doi.org/doi:10.1002/jmr.973

(56) Pearlman, D. A.; Rao, B. G., *Free energy calculations: Methods and applications.* In *Encyclopedia of computational chemistry*; Schleyer, P. v. R., Ed. **1998**; Vol. 2, p 1036-1061. doi:10.1002/0470845015.cfa011

(57) Peter, C.; Oostenbrink, C.; Dorp, A. v.; van_Gunsteren, W. F., *Estimating entropies from molecular dynamics simulations.* J. Chem. Phys. **2004**, *120*, 2652-2661. dx.doi.org/doi:10.1063/1.1636153

(58) Chipot, C.; Pohorille, A., *Calculating Free Energy Differences Using Perturbation Theory.* In *Free energy calculations: Theory and applications in chemistry and biology*; Springer: Berlin, Heidelberg, **2007**. ISBN: 978-3540384472

(59) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K., *Extraction of configurational entropy from molecular simulations via an expansion approximation.* J. Chem. Phys. **2007**, *127*, 024107. dx.doi.org/doi:10.1063/1.2746329

(60) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K., *Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101*

Ubiquitin E2 Variant Domain with an HIV-Derived PTAP Nonapeptide. J. Mol. Biol. **2009**, *389*, 315-335. dx.doi.org/doi:10.1016/j.jmb.2009.04.003

(61) Ihara, S. *Information Theory for Continuous Systems*; World Scientific Publishing, **1993**. ISBN:978-9810209858

(62) Planck, M., *Absolute Entropie und chemische Konstante.* Ann. der Physik **1922**, *371*, 365-372. dx.doi.org/doi:10.1002/andp.19223712105

(63) Landau, L. D.; Lifshitz, E. M. *Statistical Physics Part 1-Vol. 5*; 3rd ed. ed.; Oxford: Pergamon Press, **1980**.

(64) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row, **1973**. ISBN:978-1891389153

(65) Pitzer, K. S., *Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation: II. Unsymmetrical Tops Attached to a Rigid Frame.* J. Chem. Phys. **1946**, *14*, 239. dx.doi.org/doi:10.1063/1.1932193

(66) Herschbach, D. R.; Johnston, H. S.; Rapp, D., *Molecular Partition Functions in Terms of Local Properties.* J. Chem. Phys. **1959**, *31*, 1652-1661. dx.doi.org/doi:10.1063/1.1730670

(67) Gō, N.; Scheraga, H. A., *On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation.* Macromolecules **1976**, *9*, 535-542. dx.doi.org/doi:10.1021/ma60052a001

(68) Potter, M. J.; Gilson, M. K., *Coordinate Systems and the Calculation of Molecular Properties.* J. Phys. Chem. A **2002**, *106*, 563-566. dx.doi.org/doi:10.1021/jp0135407

(69) Chang, C.-E.; Potter, M. J.; Gilson, M. K., *Calculation of Molecular Configuration Integrals.* J. Phys. Chem. B **2003**, *107*, 1048-1055. dx.doi.org/doi:10.1021/jp027149c

(70) Kirkwood, J. G., *Statistical Mechanics of Fluid Mixtures.* J. Chem. Phys. **1935**, *3*, 300-313. dx.doi.org/doi:10.1063/1.1749657

(71) Lazaridis, T.; Karplus, M., *Thermodynamics of protein folding: a microscopic view.* Biophys. Chem. **2003**, *100*, 367-395. dx.doi.org/doi:10.1016/S0301-4622(02)00293-4

(72) Zhou, H.-X.; Gilson, M. K., *Theory of Free Energy and Entropy in Noncovalent Binding.* Chem. Rev. **2009**, *109*, 4092-4107. dx.doi.org/doi:10.1021/cr800551w

(73) Reinhard, F.; Grubmüller, H., *Estimation of absolute solvent and solvation shell entropies via permutation reduction.* J. Chem. Phys. **2007**, *126*, 014102. dx.doi.org/doi:10.1063/1.2400220

(74) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E., *How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?* Biophys. J. **2011**, *100*, L47-L49. dx.doi.org/doi:10.1016/j.bpj.2011.03.051

(75) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., *Equation of State Calculations by Fast Computing Machines* J. Chem. Phys. **1953**, *21*, 1087-1092. dx.doi.org/doi:10.1063/1.1699114

(76) Steinbrecher, T., *amber2accent v0.4*, **2007**. ambermd.org/amber2accent/

(77) Case, D. A.; III, T. E. C.; Darden, T.; Gohlke, H.; Luo, R.; Merz_Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., *The Amber biomolecular simulation programs.* J Comp. Chem. **2005**, *26*, 1668-1688. dx.doi.org/doi:10.1002/jcc.20290

(78) Abagyan, R.; Totrov, M.; Kuznetsov, D., *ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation.* J. Comput. Chem. **1994**, *15*, 488-506. dx.doi.org/doi:10.1002/jcc.540150503

(79) Paninski, L., *Estimation of Entropy and Mutual Information.* Neural Comput. **2003**, *15*, 1191-1253. dx.doi.org/doi:10.1162/089976603321780272
(80) Schürmann, T., *Bias analysis in entropy estimation.* J. Phys. A **2004**, *37*, L295-L301. dx.doi.org/doi:10.1088/0305-4470/37/27/L02
(81) Steuer, R.; Kurths, J.; Daub, C. O.; Weise, J.; Selbig, J., *The mutual information: Detecting and evaluating dependencies between variables.* Bioinformatics **2002**, *18 Suppl. 2*, S231-S240. dx.doi.org/doi:10.1093/bioinformatics/18.suppl_2.S231
(82) Herzel, H.; Schmitt, A. O.; Ebeling, W., *Finite sample effects in sequence analysis.* Chaos Solitons Fractals **1994**, *4*, 97-113. dx.doi.org/doi:10.1016/0960-0779(94)90020-5
(83) Jaynes, E. T.; Bretthorst, G. L. *Probability Theory: The Logic of Science*, **2003**. ISBN:978-0521592710
(84) Juneja, A.; Numata, J.; Nilsson, L.; Knapp, E. W., *Merging Implicit with Explicit Solvent Simulations: Polyethylene Glycol.* J. Chem. Theory Comput. **2010**, *6*, 1871-1883. dx.doi.org/doi:10.1021/ct100075m
(85) Bussi, G.; Parrinello, M., *Accurate sampling using Langevin dynamics.* Phys. Rev. E **2007**, *75*, 056707. dx.doi.org/doi:10.1103/PhysRevE.75.056707
(86) Andersen, H. C., *Molecular dynamics simulations at constant pressure and/or temperature.* J. Chem. Phys. **1980**, *72*, 2384-2393. dx.doi.org/doi:10.1063/1.439486
(87) Martyna, G. J.; Klein, M. L.; Tuckerman, M., *Nosé-Hoover chains: The canonical ensemble via continuous dynamics.* J. Chem. Phys. **1992**, *97*, 2635-2643. dx.doi.org/doi:10.1063/1.463940
(88) Rosta, E.; Buchete, N.-V.; Hummer, G., *Thermostat Artifacts in Replica Exchange Molecular Dynamics Simulations.* J. Chem. Theory Comput. **2009**, *5*, 1393-1399. dx.doi.org/doi:10.1021/ct800557h
(89) Chocholouová, J.; Feig, M., *Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations.* J. Comput. Chem. **2006**, *27*, 719-729. dx.doi.org/doi:10.1002/jcc.20387
(90) Haberthür, U.; Caflisch, A., *FACTS: Fast Analytical Continuum Treatment of Solvation.* J. Comput. Chem. **2007**, *29*, 701-715. dx.doi.org/doi:10.1002/jcc.20832
(91) Marsaglia, G., *Choosing a Point from the Surface of a Sphere.* Ann. Math. Statist. **1972**, *43*, 645-646. dx.doi.org/doi:10.1214/aoms/1177692644
(92) Neumann, J. v., *Various Techniques Used in Connection with Random Digits.* NBS Appl. Math. Ser. **1951**, *12*, 36-38.
(93) L'Ecuyer, P., *Maximally equidistributed combined Tausworthe generators.* Mathematics of computation **1996**, *65*, 203-213. dx.doi.org/www.jstor.org/stable/2153840
(94) MacKerell_Jr, A.et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins.* J. Phys. Chem. B **1998**, *102*, 3586-3616.
(95) Hamm, S. W. a. P., *Structure Determination of Trialanine in Water Using Polarization Sensitive Two-Dimensional Vibrational Spectroscopy.* J. Phys. Chem. B **2000**, *104*, 11316-11320. dx.doi.org/doi:10.1021/jp001546a
(96) Schweitzer-Stenner, R.; Eker, F.; Huang, Q.; Griebenow, K., *Dihedral Angles of Trialanine in D2O Determined by Combining FTIR and Polarized Visible Raman Spectroscopy.* J. Am. Chem. Soc. **2001**, *123*, 9628-9633.
(97) Wolynes, P. G., *Energy landscapes and solved protein-folding problems.* Phil. Trans. R. Soc. A **2005**, *363*, 453-467. dx.doi.org/doi:10.1098/rsta.2004.1502

(98) Krivov, S.; Chekmarev, S. F.; Karplus, M., *Potential Energy Surfaces and Conformational Transitions in Biomolecules: A Successive Confinement Approach Applied to a Solvated Tetrapeptide.* Phys. Rev. Lett. **2002**, *88*, 038101. dx.doi.org/doi:10.1103/PhysRevLett.88.038101

(99) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem. **1983**, *4*, 187-217. dx.doi.org/doi:10.1002/jcc.540040211

(100) Schaefer, M.; Bartels, C.; Karplus, M., *Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model.* J. Mol. Biol. **1998**, *284*, 835-848. dx.doi.org/doi:10.1006/jmbi.1998.2172

(101) Matsumoto, M.; Nishimura, T., *Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator.* ACM T. Model. Comput. S. **1998**, *8*, 3-30. dx.doi.org/doi:10.1145/272991.272995

(102) Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*; Vintage Books USA, **2005**. ISBN:978-0679776314

(103) Kraskov, A.; Stögbauer, H.; Grassberger, P., *Estimating mutual information.* Phys. Rev. E **2004**, *69*, 066138. dx.doi.org/doi:10.1103/PhysRevE.69.066138

(104) Hnizdo, V.; Gilson, M. K., *Thermodynamic and Differential Entropy under a Change of Variables.* Entropy **2010**, *12*, 578-590. dx.doi.org/doi:10.3390/e12030578

(105) Moddemeijer, R., *On estimation of entropy and mutual information of continuous distributions.* Signal Processing **1989**, *16*, 233-248. dx.doi.org/doi:10.1016/0165-1684(89)90132-1

(106) Beirlant, J.; Dudewicz, E. J.; Györfi, L.; Meulen, E. C. v. d., *Nonparametric entropy estimation: an overview.* Intern. J. Math. Stat. Sci. **1997**, *6*, 17-39.

(107) Hnizdo, V.; Darian, E.; Federowicz, A.; Demchuk, E.; Li, S.; Singh, H., *Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules.* J. Comput. Chem. **2007**, *28*, 655-668. dx.doi.org/doi:10.1002/jcc.20589

(108) Hnizdo, V.; Singh, H.; Misra, N.; Fedorowicz, A.; Demchuk, E., *Nearest neighbor estimates of entropy.* Amer. J. Math. Management Sci. **2003**, *23*, 301-321.

(109) Nilsson, M., *On the Estimation of Differential Entropy From Data Located on Embedded Manifolds.* IEEE T. Inform. Theory **2007**, *53*, 2330-2341. dx.doi.org/doi:10.1109/TIT.2007.899533

(110) Bercher, J. F.; Vignat, C., *Estimating the entropy of a signal with applications.* IEEE T. Signal Process. **2000**, *48*, 1687-1694. dx.doi.org/doi:10.1109/78.845926

(111) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover, **1955**. ISBN:978-0486639413

(112) Carlsson, J.; Åqvist, J., *Calculations of solute and solvent entropies from molecular dynamics simulations.* Phys. Chem. Chem. Phys. **2006**, *8*, 5385-5395. dx.doi.org/doi:10.1039/b608486a

3 Influence of Spacer-Receptor Interactions on the Stability of Bivalent Ligand-Receptor Complexes

This chapter is based on an accepted publication. A copy of the original submitted manuscript can be found attached at the end of the printed version of this thesis.

Numata, J.; Juneja, A.; Diestler, D. J.; Knapp, E. W., *Influence of Spacer-Receptor Interactions on the Stability of Bivalent Ligand-Receptor Complexes*. J. Phys. Chem. B **2012**, *116*(8), 2595-2604. Submitted on 23-nov-2011. Accepted on 25-jan-2012. dx.doi.org/10.1021/jp211383s

3.1 Introduction

In the publication that this chapter introduces, we explore the thermodynamics of spacer-mediated bivalent ligand binding to a protein receptor. We combine a minimalistic model of the protein receptor surface with a random walk model for the flexible polymer serving as spacer (Fig. 3.1). We build upon the statistical mechanical fundamental theory of the multivalent enhancement effect developed in our group[1,2]. To our knowledge, this is the first time that the interaction between the polymer spacer and the protein receptor is included into a model of multivalent interactions, except for one recent study[3]. In light of our findings, spacer–receptor interactions are an essential ingredient to understand the statistical mechanics of multivalent enhancement.

3.1.1 Biological and pharmaceutical relevance of multivalency

Multivalent binding is a widespread strategy in nature. It allows otherwise weak binders to act in concert and generate strong binding. Properly designed multivalent ligands can bind to their corresponding multivalent receptors often by orders of magnitude more efficiently than their monovalent analogs[4-19]. The influenza virus infects by multivalently attaching to multiple sialic acid-containing oligosaccharides expressed on the surface of respiratory cells. Recently, a sialic-acid-functionalized nanoparticle was synthesized within my cooperative research group (SFB765) that mimics these interactions and binds to the virus thanks to multivalent interactions[20]. An

anticancer multivalent ligand was also recently designed by another group, packing together several copies of a tumor targeting moiety peptide with the drug camptothecin[21].

3.1.2 Polymer spacer-receptor interactions

When monovalent ligands are tethered together using flexible polymer spacers, the interactions of the spacer with the target protein receptor may affect the thermodynamics of binding significantly. The protein receptor constrains the polymer's conformational freedom, causing a large conformational entropy loss. This may be compensated by favorable enthalpic and hydrophobic interactions. We quantify these thermodynamic variables using our minimalistic models, which have the advantage of focusing on the overall protein surface topography and not depending on the precise crystal structure of the receptor involved.

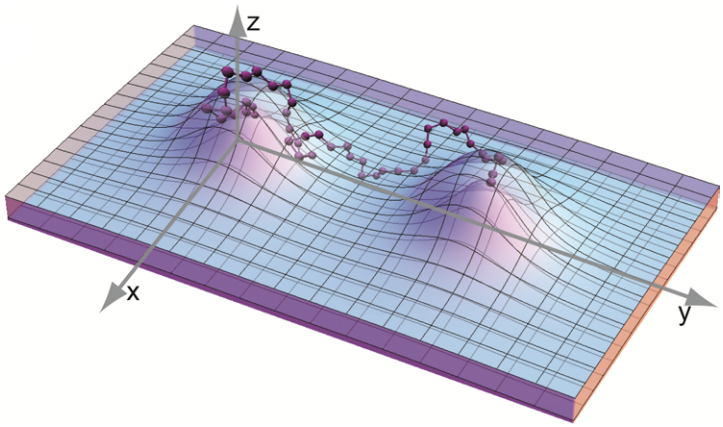

Fig. 3.1: Three-dimensional perspective of the receptor surface, concave between the binding sites. This minimalistic model captures the general features of RET channel[4] receptor and the polyethylene glycol spacer. Shown are the receptor hard surface (solid) and the attractive interacting surface (transparent) with bivalent ligand (including the polymer spacer) bound to binding sites located on the top of the two Lorentzian hills.

3.1.3 Comparison of our model to experiment

To test the validity of our theoretical model, we compare our results to the experimental bivalent enhancement of a successful synthetic bivalent ligand. Cyclic guanidine mono-phosphate (cGMP) ligands activate nucleotide-gated ion channels in bovine rod photoreceptor cells (RET)[4]. It was shown that synthesizing a bivalent ligand by tethering two cGMP ligating units with a polyethylene glycol spacer enhances the binding affinity to RET by about two orders of magnitude[4].

3.2 Conclusions

3.2.1 Receptor topography: concave, planar or convex

Our results show that for non-concave receptors, the enhancement effect can be reduced by several orders of magnitude, even to the point of rendering the binding of monomeric ligands more efficient than that of bivalent ligands. The reason is the huge loss in spacer conformational entropy from the constraints imposed by the receptor. For convex receptors, this entropy loss can no longer be compensated by other thermodynamic driving forces like enthalpy or the hydrophobic effect. The reduction of the enhancement effect resulting from conformational entropy loss is most pronounced for convex receptor surfaces, very large for planar receptor surfaces and still significant for concave receptor surfaces (like the 14-3-3 protein, Fig. 3.2). The estrogen receptor[22,23], Fig. 3.3, presents a non-concave, corrugated surface that our model predicts will abolish enhancement of bivalent ligand binding when using flexible polymers as tethers.

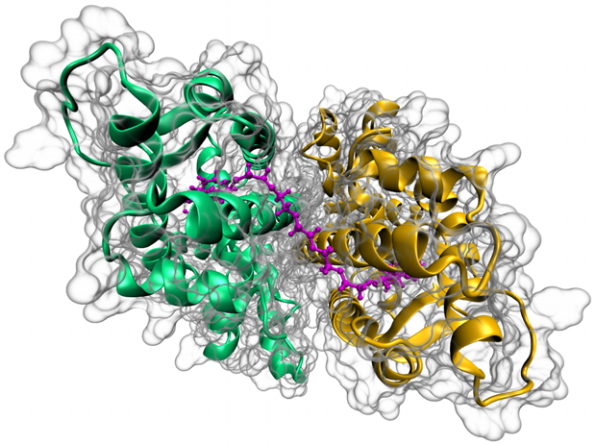

Fig. 3.2: Corrugated R–surface between the two binding sites of the bivalent homodimeric estrogen receptor (ER), shown binding a bivalent ligand consisting of two diethylstilbestrol (DES) ligating units connected by a PEG spacer of 21 main-chain atoms. The spacer geometry is modelled on the basis of the ER crystal structure (PDB id 3ERD)[24]. The PEG spacer needs to circumvent the two α-helices, which protrude from the ER surface. The backbones of the two polypeptides are traced by rubber bands in green and yellow. The transparent gray shaded area pictures the protein volume including the side chains.

3.2.2 Interaction thermodynamics: repulsive or attractive

In our study, we find that the conformational entropy loss needs to be compensated by favourable interactions between spacer and receptor, either of enthalpic or hydrophobic origin, for enhancement to occur. This is consistent with the physicochemical properties of polyethylene glycol, which is known to interact weakly but favourably with proteins. Such weak attractions have been shown to exist between protein surfaces and polyethylene glycol (PEG)[25-31], often used

as spacer material. We model the PEG–receptor surface interaction as an attractive layer next to the hard receptor surface (see Fig. 3.1), and find that a weak attraction only one tenth the magnitude of that between hydrophobic aliphatic carbon atoms can reproduce the enhancement observed in the experiment[4].

The model was programmed in around 5000 lines of original code in C++ for efficient generation and testing of a large number of configurations (up to 2×10^{11}) to ensure numerical precision. For more details about the model, results and comparison to experiment see the article's full text and supporting information attached at the end of this thesis.

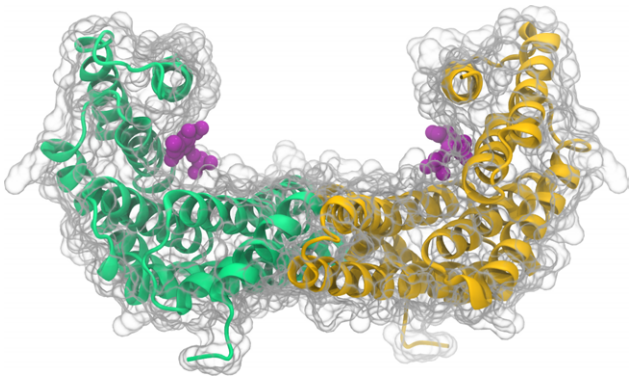

Fig. 3.3: Concave receptor surface between the two binding sites of the bivalent 14-3-3 dimeric receptor (PDB 3RDH) pictured binding two FOBISIN101 ligands (magenta)[32]. The same molecular representation for the protein is used as in Fig. 3.2.

3.3 References for Chapter 3

(1) Diestler, D. J.; Knapp, E. W., *Statistical Thermodynamics of the Stability of Multivalent Ligand-Receptor Complexes.* Phys. Rev. Lett. **2008**, *100*, 178101. dx.doi.org/doi:10.1103/PhysRevLett.100.178101

(2) Diestler, D. J.; Knapp, E. W., *Statistical Mechanics of the Stability of Multivalent Ligand–Receptor Complexes.* J. Phys. Chem. C **2010**, *114*, 5287-5304. dx.doi.org/doi:10.1021/jp904258c

(3) Wang, S.; Dormidontova, E. E., *Nanoparticle Design Optimization for Enhanced Targeting: Monte Carlo Simulations.* Biomacromolecules **2010**, *11*, 1785-1795. dx.doi.org/doi:10.1021/bm100248e

(4) Kramer, R. H.; Karpen, J. W., *Spanning binding sites on allosteric proteins with polymer-linked ligand dimers.* Nature **1998**, *395*, 710-713. dx.doi.org/doi:10.1038/27227

(5) Blaustein, R. O.; Cole, P. A.; Williams, C.; Miller, C., *Tethered blockers as molecular 'tape measures' for a voltage-gated K+channel.* Nat. Struct. Biol. **2000**, *7*, 309-311. dx.doi.org/doi:10.1038/74076

(6) Loidl, G.; Groll, M.; Musiol, H.-J.; Huber, R.; Moroder, L., *Bivalency as a principle for proteasome inhibition.* Proc. Natl. Acad. Sci. U.S.A **1999**, *96*, 5418-5422. dx.doi.org/doi:10.1073/pnas.96.10.5418

(7) Kitov, P. I.; Sadowska, J. M.; Mulvey, G.; Armstrong, G. D.; Ling, H.; Pannu, N. S.; Read, R. J.; Bundle, D. R., *Shiga-like toxins are neutralized by tailored multivalent carbohydrate ligands.* Nature **1999**, *403*, 669-672. dx.doi.org/doi:10.1038/35001095

(8) Fan, E.; Zhang, Z.; Minke, W. E.; Hou, Z.; Verlinde, C. L. M. J.; Hol, W. G. J., *High-affinity pentavalent ligands of escherichia coli heat-labile enterotoxin by modular structure-based design.* J. Am. Chem. Soc. **2000**, *122*, 2663-2664. dx.doi.org/doi:10.1021/ja993388a

(9) Gargano, J. M.; Ngo, T.; Kim, J. Y.; Acheson, D. W. K.; Lees, W. J., *Multivalent inhibition of AB5 toxins.* J. Am. Chem. Soc. **2001**, *123*, 12909-12910. dx.doi.org/doi:10.1021/ja016305a

(10) Zhang, Z.; Merritt, E. A.; Ahn, M.; Roach, C.; Hou, Z.; Verlinde, C. L. M. J.; Hol, W. G. J.; Fan, E., *Solution and crystallographic studies of branched multivalent ligands that inhibit the receptor-binding of cholera toxin.* J. Am. Chem. Soc. **2002**, *124*, 12991-12998. dx.doi.org/doi:10.1021/ja027584k

(11) Kitov, P. I.; Shimizu, H.; Homans, S. W.; Bundle, D. R., *Optimization of tether length in nonglycosidically linked bivalent ligands that target sites 2 and 1 of a Shiga-like toxin.* J. Am. Chem. Soc. **2003**, *125*, 3284-3294. dx.doi.org/doi:10.1021/ja0258529

(12) Mulder, A.; Auletta, T.; Sartori, A.; Ciotto, S. D.; Casnati, A.; Ungaro, R.; Huskens, J.; Reinhoudt, D. N., *Divalent binding of a bis(adamantyl)-functionalized calix[4]arene to β-cyclodextrin-based hosts: An experimental and theoretical study on multivalent binding in solution and at self-assembled monolayers.* J. Am. Chem. Soc. **2004**, *126*, 6627-6636. dx.doi.org/doi:10.1021/ja0317168

(13) Trevitt, C. R.; Craven, C. J.; Milanesi, L.; Syson, K.; Mattinen, M.-L.; Perkins, J.; Annila, A.; Hunter, C. A.; Waltho, J. P., *Enhanced ligand affinity for receptors in which components of the binding site are independently mobile.* Chem. and Biol. **2005**, *12*, 89-97. dx.doi.org/doi:10.1016/j.chembiol.2004.11.007

(14) Farrera, J.-A.; Hidalgo-Fernández, P.; Hannink, J. M.; Huskens, J.; Rowan, A. E.; Sommerdijk, N. A. J. M.; Nolte, R. J. M., *Divalent ligand for intramolecular complex formation to streptavidin.* Org. Biomol. Chem. **2005**, *3*, 2393-2395. dx.doi.org/doi:10.1039/B505700K

(15) Krishnamurthy, V. M.; Semetey, V.; Bracher, P. J.; Shen, N.; Whitesides, G. M., *Dependence of effective molarity on linker length for an intramolecular protein-ligand system.* J. Am. Chem. Soc. **2007**, *129*, 1312-1320. dx.doi.org/doi:10.1021/ja066780e

(16) Shewmake, T. A.; Solis, F. J.; Gillies, R. J.; Caplan, M. R., *Effects of linker length and flexibility on multivalent targeting.* Biomacromolecules **2008**, *9*, 3057-3064. dx.doi.org/doi:10.1021/bm800529b

(17) Kane, R. S., *Thermodynamics of multivalent interactions: influence of the linker.* Langmuir **2010**, *26*, 8636-8640. dx.doi.org/doi:10.1021/la9047193

(18) Huskens, J.; Mulder, A.; Auletta, T.; Nijhuis, C. A.; Ludden, M. J. W.; Reinhoudt, D. N., *A model for describing the thermodynamics of multivalent host-guest interactions at interfaces.* J. Am. Chem. Soc. **2004**, *126*, 6784-6797. dx.doi.org/doi:10.1021/ja049085k

(19) Zhou, H.-X., *Quantitative Relation between Intermolecular and Intramolecular Binding of Pro-Rich Peptides to SH3 Domains.* Biophys. J.. **2006**, *91*, 3170-3181. dx.doi.org/doi:10.1529/biophysj.106.090258

(20) Papp, I.; Sieben, C.; Ludwig, K.; Roskamp, M.; Böttcher, C.; Schlecht, S.; Herrmann, A.; Haag, R., *Inhibition of Influenza Virus Infection by Multivalent Sialic-Acid-Functionalized Gold Nanoparticles.* Small **2010**, *6*, 2900-2906. dx.doi.org/doi:10.1002/smll.201001349

(21) Khandare, J. J.; Chandna, P.; Wang, Y.; Pozharov, V. P.; Minko, T., *Novel Polymeric Prodrug with Multivalent Components for Cancer Therapy.* J Pharmacology and experimental therapeutics **2006**, *317*, 929-937. dx.doi.org/doi:10.1124/jpet.105.098855

(22) LaFratea, A. L.; Carlsona, K. E.; Katzenellenbogen, J. A., *Steroidal bivalent ligands for the estrogen receptor: Design, synthesis, characterization and binding affinities.* Bioorg. Med. Chem. **2009**, *17*, 3528-3535. dx.doi.org/doi:10.1016/j.bmc.2009.04.016

(23) Abendroth, F.; Bujotzek, A.; Shan, M.; Haag, R.; Weber, M.; Seitz, O., *DNA-Controlled Bivalent Presentation of Ligands for the Estrogen Receptor.* Angew. Chem. Int. Ed. **2011**, *50*, 8592-8596. dx.doi.org/doi:10.1002/anie.201101655

(24) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L., *The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen.* Cell **1998**, *95*, 927-937. dx.doi.org/10.1016/S0092-8674(00)81717-1

(25) Sheth, S.; Leckband, D. E., *Measurements of attractive forces between proteins and end-grafted poly(ethylene glycol) chains.* Proc. Natl. Acad. Sci. U.S.A **1997**, *94*, 8399-8404. dx.doi.org/PMID:9237988

(26) Israelachvili, J., *The different faces of poly(ethylene glycol).* Proc. Natl. Acad. Sci. U.S.A **1997**, *94*, 8378-8379.

(27) Vivarès, D.; Belloni, L.; Tardieu, A.; Bonneté, F., *Catching the PEG-induced attractive interaction between proteins.* Eur. Phys. J. E **2002**, *9*, 15-25. dx.doi.org/doi:10.1140/epje/i2002-10047-7

(28) Sheth, S.; Efremova, N.; Leckband, D. E., *Interactions of poly(ethylene oxide) brushes with chemically selective surfaces.* J. Phys. Chem. B **2000**, *104*, 7652-7662. dx.doi.org/doi:**10.1021/jp000298f**

(29) Leckband, D.; Israelachvili, J., *Intermolecular forces in biology.* Q. Rev. Biophys. **2001**, *34*, 105-267. dx.doi.org/doi:10.1017/S0033583501003687

(30) Zhou, H.-X.; Rivas, G.; Minton, A. P., *Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences.* Annu. Rev. Biophys. **2008**, *37*, 375-397. dx.doi.org/doi:10.1146/annurev.biophys.37.032807.125817

(31) Dhakshnamoorthy, B.; Raychaudhury, S.; Blachowicz, L.; Roux, B., *Cation-selective pathway of OmpF porin revealed by anomalous X-ray diffraction.* J. Mol. Biol. **2010**, *396*, 293-300. dx.doi.org/doi:10.1016/j.jmb.2009.11.042

(32) Zhao, J.; Du, Y.; Horton, J. R.; Upadhyay, A. K.; Lou, B.; Bai, Y.; Zhang, X.; Du, L.; Li, M.; Wang, B.; Zhanga, L.; Barbieri, J. T.; Khuri, F. R.; Cheng, X.; Fu, H., *Discovery and structural characterization of a small molecule 14-3-3 protein-protein interaction inhibitor.* Proc. Natl. Acad. Sci. U.S.A **2011**, *108*, 16212-16216. dx.doi.org/doi:10.1073/pnas.1100012108

4 Summary

4.1 Abstract in English

Although thermodynamics was born from the desire to optimize industrial processes, its wide applicability has recently afforded it a place in biology. Accurate estimation of thermodynamic variables for processes involving biological macromolecules is an important goal in theoretical chemistry. For macromolecules and soft matter in general, understanding of the driving forces that comprise a given free energy or binding constant requires consideration of flexibility for all molecular degrees of freedom. The entropic contribution to free energy is thus an essential ingredient, whether explicitly quantified or included in a model in the form of a correct enumeration of the multiplicity of conformations. This doctoral thesis offers two contributions to the problem of computing entropy: (1) A numerical method for estimating conformational entropy differences for macromolecules was developed. It uses techniques borrowed from information theory and applies them to statistical mechanics. The method is applicable to conformational transitions and protein-ligand binding. (2) A model to describe the enhancement of the binding affinity for a bivalent ligand tethered with a polymer spacer was expounded. The novelty of the model consists in the inclusion of spacer-receptor interactions.

4.1.1 Balanced and bias-free computation of conformational entropy differences for molecular trajectories

The mutual information expansion (MIE) is applied to estimate conformational entropy differences of macromolecules applicable to molecular dynamics or Monte Carlo simulation data on oligopeptides, polymers, proteins and ligands. The MIE serves to reduce the high dimensionality of the probability density of the conformational space of a macromolecule. The individual terms of the MIE are evaluated with a histogram method. Internal bond-angle-torsion (BAT) coordinates are used to avoid spurious correlations present when using Cartesian coordinates, which would demand using higher order terms in the MIE.

Practically all entropy estimation methods from finite samples suffer from an inherent systematic error or bias. Two approaches are applied that compensate for systematic errors that occur with a histogram method: (1) Simulation data are balanced by using the same number of coordinate sets (frames) for both conformer domains. Balancing puts fluctuations of the histogram bin contents on the same level for both conformers, allowing for efficient error cancellation. (2) Bias-removal corrects for systematic deviations due to finite number of frames per bin. Applying both corrections improves the precision of entropy differences enormously. Estimates of entropy differences are compared to thermodynamic benchmarks of polymer and peptide models, where excellent agreement is found. For trialanine as model system, the average error for the estimated conformational entropy difference is only 0.3 J/(mol K), which is 100 times smaller than without applying the two corrections. Guidelines are provided for efficiently estimating conformational entropies. The complete method, including 1st to 3rd order MI expansion, balancing and bias-removal can be performed with the program ENTROPICAL. It can be obtained from the author and used with CHARMM and NAMD topologies and trajectories.

4.1.2 Influence of spacer-receptor interactions on the stability of bivalent ligand-receptor complexes

Experiments show that a ligand-receptor complex formed by binding a bivalent ligand (D) in which the two ligating units are joined covalently by a flexible polymeric spacer (S) can be orders of magnitude more stable than the corresponding complex formed with monomeric ligands. Up until now, the molecular models that have been proposed to rationalize this “enhancement effect” neglect spacer-receptor (S–R) interactions. These interactions can nevertheless substantially influence the relative stability of complexes. Here, the results of a computational study designed to assess the impact of S–R interactions in the prototypic bivalent complex are presented and compared with results of experiments. The S–R interactions mimicking general features of biological systems are modeled by contoured R surfaces with hills (or depressions) at the binding sites. In the fictitious limit of vanishing S–R interactions, the enhancement is pronounced. This enhancement is in line with the experimental observations, although the S–R interactions, which surely occur in reality, were neglected. For strictly repulsive S–R interactions (hard R surface) the enhancement vanishes, or even reverses. This is particularly the case if the R surface is convex (i.e.

rising between the binding sites), while the enhancement is only moderately reduced if the R surface is concave. Alternatively, a weak S–R attraction close to the R surface can increase the enhancement. It is concluded that large enhancement should be observed only if both features are present: a concave R surface plus a weak S–R attraction. The latter occurs for spacer material such as polyethylene glycol (PEG), which is weakly hydrophobic and thus attracted by protein surfaces. It is shown that the enhancement of bivalent binding can be characterized by a single key parameter, which may also provide guidelines for the design of multivalent complexes with large enhancement effect.

4.2 Zusammenfassung in deutscher Sprache

Obwohl die Thermodynamik ursprünglich zur Optimierung industrieller Prozesse entwickelt wurde, hat sie sich durch ihre breiten Anwendungsmöglichkeiten in letzter Zeit auch einen Platz in der Biologie gesichert. Ein wichtiges Ziel in der theoretischen Chemie stellt die genaue Abschätzung thermodynamischer Variablen für Prozesse dar, an denen biologische Makromoleküle beteiligt sind. Das Verständnis der Triebkräfte, die eine gewisse freie Energie bei Makromolekülen und allgemein weicher Materie ausmachen, bedarf der Miteinbeziehung der Flexibilität aller molekularen Freiheitsgrade. Der entropische Beitrag zur freien Energie ist also ein unentbehrlicher Bestandteil, unabhängig davon, ob er explizit quantifiziert wird oder bei einem Modell in Form einer richtigen Aufzählung der Vielfachheit der Konformationen mit einbezogen wird. Die vorliegende Dissertation liefert zu dem Problem der Entropieberechnung zwei Beiträge: (1) Eine numerische Methode zur Berechnung von Entropiedifferenzen bei Makromolekülen wurde entwickelt. Sie entlehnt Techniken aus der Informationstheorie und wendet sie in der statistischen Mechanik an. Die Methode ist bei Konformationsänderungen und Protein-Ligandbindung verwendbar. (2) Zur Beschreibung der Verstärkung der Bindungsaffinität bei bivalenten Liganden, die mit einem Polymer-Spacer verknüpft sind, wurde ein geeignetes Modell entwickelt. Neu bei diesem Modell ist die Einbeziehung von Spacer-Rezeptor-Wechselwirkungen.

4.2.1 Ausbalancierte und von systematischen Fehlern bereinigte Berechnung konformationeller Entropiedifferenzen für molekulare Trajektorien

Die Reihenentwicklung der wechselseitigen Information (*mutual information expansion,* MIE) wird benutzt, um Differenzen in konformationeller Entropie bei Makromolekülen zu berechnen. Die Methode ist auf Moleküldynamik- oder Monte-Carlo-Simulationsdaten von Polymeren, Proteinen und Liganden anwendbar. Die MIE dient der Dimensionsreduktion der Wahrscheinlichkeitsdichte des konformationellen Raums eines Makromoleküls. Die einzelnen Entwickungsterme der MIE werden mit Hilfe einer Histogrammmethode ausgewertet. Ein internes Koordinatensystem (Bindungslänge, Bindungswinkel und Torsionswinkel, das

sogenannte BAT-System) wird benutzt, um die, bei kartesischen Koordinaten anwesenden, störenden Korrelationen zu vermeiden, die Entwickungsterme höherer Ordnung in der MIE benötigen würden. Praktisch alle Entropieschätzungsmethoden, die über eine endliche Menge von Daten verfügen, leiden an systematischen Fehlern (Bias). Zwei Korrekturmethoden werden eingesetzt, um diese systematischen Fehler von Histogrammmethoden auszugleichen: (1) Die Simulationsdaten werden ausbalanciert, indem die gleiche Anzahl von Koordinatensätzen (Einzelbildern) für beide Konformerdomänen benutzt wird. Durch dieses Ausbalancieren werden die Schwankungen der Belegungen einzelner Histogrammsäulen für beide Konformere im Mittel gleich groß. Dies führt zu einem effizienten Fehlerausgleich. (2) Die Bereinigung der systematischen Fehler (Bias) kompensiert Abweichungen, die auf Grund der endlichen Menge von Daten pro Histogrammsäule entstanden sind. Die gleichzeitige Verwendung beider Korrekturen verbessert die Genauigkeit der Abschätzung von Entropiedifferenzen erheblich. Die geschätzten Entropiedifferenzen werden mit thermodynamischen Bezugswerten für Polymer- und Peptidmodelle verglichen und stimmen mit diesen ausgezeichnet überein. Für das Modellsystem Trialanin betrug der durchschnittliche Fehler für die geschätzte konformationelle Entropiedifferenz nur 0.3 J (mol K), welcher 100-mal kleiner ist als bei Weglassen beider Korrekturmethoden. Leitlinien zur effizienten Berechnung konformationeller Entropie werden angegeben. Die komplette Methode, einschließlich MIE 1. bis 3. Grad, Ausbalancieren und Bereinigung von systematischen Fehlern, kann mit Hilfe des Programms ENTROPICAL ausgeführt werden. Das Programm arbeitet auf CHARMM- und NAMD-Topologien und -Trajektorien und wird vom Autor auf Anfrage zur Verfügung gestellt.

4.2.2 Einfluss von Spacer-Rezeptor-Wechselwirkungen auf die Stabilität von bivalenten Ligand-Rezeptor-Komplexen

Experimente zeigen, dass ein durch einen bivalenten Liganden (D) gebildeter Ligand-Rezeptor-Komplex, in dem beide bindenden Einheiten mit Hilfe eines flexiblen Polymer-Spacers kovalent verknüpft werden, um Größenordnungen stabiler sein kann als der entsprechende, durch monomere Liganden gebildete Komplex. Bislang haben molekulare Modelle der bivalenten Bindung den Verstärkungseffekt erklärt, ohne die Wechselwirkungen zwischen Spacer und Rezeptor (S–R) zu berücksichtigen. Letztere können aber die relative Stabilität der Komplexe

entscheidend beeinflussen. Wir haben Computersimulationen an geeigneten, prototypischen Modellsystemen für den bivalenten Komplexe durchgeführt, um die Auswirkungen der S–R-Wechselwirkungen auf die Bindungseffizienz zu untersuchen und mit experimentellen Ergebnissen verglichen. Die modellierten S–R-Wechselwirkungen bilden die allgemeinen Merkmale biologischer Systeme nach und werden als R-Oberfläche mit Bergen (bzw. Tälern) an den Bindungsstellen modelliert. Im fiktiven Grenzfall verschwindender S–R-Wechselwirkungen ist die Verstärkung der Bindungseffizienz groß. Dies deckt sich mit experimentellen Beobachtungen, obwohl die in der Realität sicher auftretenden S–R-Wechselwirkungen vernachlässigt wurden. Bei rein abstoßenden S–R-Wechselwirkungen (harter R-Oberfläche) verschwindet die Verstärkung oder kehrt sich gar um. Das ist insbesondere bei konvexer (also zwischen den Bindungsstellen gewölbter) R-Oberfläche der Fall, wobei die Verstärkung bei einer konkaven Oberfläche nur unwesentlich verringert ist. Alternativ kann eine schwache S–R-Anziehung nahe der R-Oberfläche die Verstärkung erhöhen. Es wird geschlussfolgert, dass nur in dem Fall, dass beide Merkmale anwesend sind, eine hohe Verstärkung zu erwarten ist: d. h. bei einer konkaven R-Oberfläche und einer schwachen S–R-Anziehung. Letzteres tritt bei Spacermaterialien wie Polyethylenglycol (PEG) auf, welches geringfügig hydrophob ist und aus diesem Grund von Proteinoberflächen angezogen wird. Es wird gezeigt, dass die Verstärkung bivalenter Bindung mit einem einzelnen Parameter gekennzeichnet werden kann, woraus sich Leitlinien für den Entwurf multivalenter Komplexe mit hoher Verstärkung gewinnen lassen.

Suppose that we were asked to arrange the following in two categories:

– *distance, mass, electric force, entropy, beauty, melody.*

I think there are the strongest grounds for placing entropy alongside beauty and melody, and not with the first three. Entropy is only found when the parts are viewed in association, and it is by viewing or hearing the parts that beauty and melody are discerned. All three are features of arrangement. It is a pregnant thought that one of these three associates should be able to figure as a commonplace quantity of science. The reason why this stranger can pass itself off among the aborigines of the physical world is that it is able to speak their language, viz., the language of arithmetic. It has a measure-number associated with it and so is made quite at home in physics.

Eddington, 1935 in *The Nature of the Physical World.*

Influence of Spacer-Receptor Interactions on the Stability of Bivalent Ligand-Receptor Complexes

Revised Manuscript from 23-jan-2012. This manuscript is the version included in the submitted PhD Thesis. Final version is http://dx.doi.org/10.1021/jp211383s

Jorge Numata[1], Alok Juneja[1,2], Dennis J. Diestler[1,3] and Ernst-Walter Knapp[1]*

[1] Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Fabeckstrasse 36A, D-14195 Berlin, Germany
[2] Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden
[3] University of Nebraska-Lincoln, Lincoln, Nebraska 68583, USA

* corresponding author: knapp@chemie.fu-berlin.de

ABSTRACT: Experiments show that a ligand-receptor complex formed by binding a bivalent ligand (D) in which the two ligating units are joined covalently by a flexible polymeric spacer (S) can be orders of magnitude more stable than the corresponding complex formed with monomeric ligands. Although molecular models rationalizing this "enhancement effect" have been proffered, they ignore spacer-receptor (S–R) interactions, which can substantially influence the relative stability of complexes. Here, the results of a computational study designed to assess the impact of S–R interactions in the prototypic bivalent complex are presented and compared with results of experiments. The S–R interactions mimicking general features of biological systems are modeled by contoured R surfaces with hills (or depressions) at the binding sites. In the fictitious limit of vanishing S–R interactions, the enhancement is pronounced, as observed in experiments. For strictly repulsive S–R interactions (hard R surface) the enhancement vanishes, or even reverses. This is particularly the case if the R surface is convex (i.e. rising between the binding sites), while the enhancement is only moderately reduced if the R surface is concave. Alternatively, a weak S–R attraction close to the R surface can increase the enhancement. It is concluded that large enhancement should be observed only if both features are present: a concave R surface plus a weak S–R attraction. The latter occurs for spacer material such as polyethylene glycol (PEG), which is weakly hydrophobic and thus attracted by protein surfaces. It is shown that the enhancement of

bivalent binding can be characterized by a single key parameter, which may also provide guidelines for the design of multivalent complexes with large enhancement effect.

Abbreviations: CA, carbonic anhydrase; cGMP, cyclic guanidine mono-phosphate; D, bivalent (or divalent) ligand; ER, estrogen receptor; *K*, binding constant; M, monomeric ligand; *q*, canonical molecular partition function; PEG, polyethylene glycol; R, bivalent receptor; $RD^{(1)}$, $RD^{(2)}$ and $RD_2^{(1)}$ denote complexes in which, respectively, one binding site of R is occupied by one ligating unit of D, both sites of R are occupied by units of a single D, and both sites are occupied by units of separate Ds; RET, nucleotide-gated ion channel in bovine rod photoreceptor cells; R_F, Flory radius (i.e. root-mean-square end-to-end distance between ligating units of free D); R_{hard} model, R model with hard surface; R_{planar} model, R model with planar surface and S-R attraction; R_{soft} model, R model with S-R attraction; S, spacer; SASA, solvent accessible surface area.

1. INTRODUCTION

Multivalent ligand-receptor complexes consist of associations of molecules held together through multiple, simultaneous, non-covalent bonds. They play essential roles in many natural (biological) processes[1-4], in medicinal chemistry for the design of new therapeutics[5,6] as well as in the synthesis of artificial supramolecular systems[7-11]. To delineate the nature of multivalent complexes, we focus on a prototypal molecular system, namely the bivalent complex in which a *bivalent* (or divalent) ligand (D) binds to a *bivalent* receptor (R). We assume that D is constructed by covalently joining two *monovalent* ligands (Ms) via a *spacer* (S) (e.g., a polymer chain). Note that we replace the term "linker" used in earlier works[12,13] by "spacer" (denoted "S") to emphasize that S imposes a constraint on the distance between the monomeric ligating units. The resulting D then consists of two ligating units connected by S. The bivalent complex forms as the units ensconce themselves in the two binding sites of R, which one can visualize as "pockets" formed by groups of atoms in special configurations so as to conform to the ligating units (i.e., the unit and the binding site are physicochemically complementary; they accommodate each other through *specific*, unique interactions). We further assume that the ligating units, which in essence are Ms modified by virtue of a covalent bond to S, practically do not differ from the original (free) Ms in their chemical nature.

This revised manuscript is included as part of the submitted doctoral thesis. For citation purposes please refer to the final and published version: **J. Phys. Chem. B 2012 116(8) http://dx.doi.org/10.1021/jp211383s**

By a judicious choice of S, one can construct a multivalent complex whose thermodynamic stability is far greater than that of its monovalent counterpart[14-29]. To be specific, we note that the stability of the prototype, in which both ligating units of D bind to the two binding sites of R, can be enhanced relative to that of the monovalent complex, in which two Ms independently bind to the two sites of R. This "enhancement effect" has an important practical implication, which is that a desired effect initiated by formation of a bivalent complex can be accomplished at concentrations of D much lower than those required of the monovalent counterpart.

It should be noted that the term enhancement effect, which implies that the stability of the bivalent complex increases relative to that of the monovalent counterpart, reflects the prejudice that one should achieve a desired effect using as little of the presumably precious (or toxic, if it involves undesirable side effects) M as possible. In principle, if D is designed poorly, the stability of the bivalent complex may just as well be less than that of its monovalent counterpart.

In previous articles[12,13] we developed a fundamental theory of the enhancement effect and compared its predictions with the results of experimental studies of the binding of cyclic guanidine mono-phosphate (cGMP) ligands, which activate nucleotide-gated ion channels in bovine rod photoreceptor cells (RET)[14]. Applying classical statistical mechanics to the prototypal molecular system, we derived closed expressions for the binding constants in terms of molecular properties. Among the approximations that were introduced to permit derivation of analytic formulas was the complete neglect of all spacer-receptor (S-R) interactions as is generally done in theoretical descriptions of multivalent binding. In spite of these simplifications the agreement between theory and experiment is reasonably good.

Recently a very simple model for S-R interactions (purely repulsive, planar R) was used in a Monte Carlo study of multivalent binding of functionalized nanoparticles[30]. The purpose of the current article is to present the results of a systematic study of a realistic topographical model designed to assess the influence of S-R interactions on the enhancement effect. The topography of the R surface is varied between convex and concave and weak S-R attractions are considered. Such weak attraction has been shown to exist between protein surfaces and polyethylene glycol (PEG)[31-37], often used as spacer material. We find that S-R interactions can play a significant role in altering the relative

stability of complexes, which can also have a strong impact on the ligand's biological activity. In the light of our present study, the reasonable agreement between experiment and our model[12,13] that ignores S-R interactions must be regarded as fortuitous, resulting from cancellations of errors. In the present study we attempt to clarify why a generally expected enhancement of bivalent ligand binding relative to the monovalent case often fails to appear for unfavorable choices of R and S.

2. THEORY

2.1. Description of Prototypal Model. We adopt the "local" nomenclature employed in earlier work[12,13]. We take the binding sites of R to be equivalent and the ligating units of D to be equivalent. The symbol $RD^{(1)}$ stands for the complex in which a single unit of one D is bound to a site of R, $RD^{(2)}$ for the R complex in which the two units of the same D bind to both sites of R, and $RD^{(1)}_2$ for the complex in which units of two different Ds bind to the two sites of R.

We treat R as an extended *rigid* body and M (or the ligating unit) as an *atom-like* point mass, neglecting internal degrees of freedom of both. We envisage the two binding sites of R to be pockets formed by atoms in fixed configurations. The Ms (or ligating units) are planted in the sites to form the complex. D is constructed by covalently connecting two Ms via a generally flexible S-chain (henceforth, S is taken to be a polymer chain). We assume that the physicochemical character of the ligating units in D is practically the same as that of the Ms alone. Except for the binding interaction, additional (non-specific) interactions of the ligating units themselves with R are ignored. This approximation is justified, since the ligating units of D and the corresponding Ms are subject to the same non-specific interactions with R. The S-R interaction $U^{(SR)}$ is based on an R model with an atomically smooth, hard surface, impenetrable to atoms of S, with an attractive layer above it. The R surface has hills or depressions at the sites, so that the topography between the sites can be characterized as concave (for hills) or convex (for depressions).

2.2. Canonical Molecular Partition Functions (*q*) and Binding Constants (*K*). We restrict our consideration to solutions sufficiently dilute that *intermolecular* interactions between different Ds or different Rs can be neglected. Interpreting $U^{(SR)}$ as potential of mean force, we implicitly account for the influence of solvent. Under these

conditions, the binding constants of the complexes can be expressed simply in terms of canonical partition functions (q) of the *isolated* species. In prior work[13], we derived formulas for these qs under assumptions that differ from the present study in several respects. Here, we focus on the prototypic bivalent rather than a tetravalent R and take M (and the ligating unit) to be *atom-like* rather than an extended rigid body, as we continue to treat R. Though these approximations may appear to be severe, they do not affect the essential conclusions of the study. Previously, we formally accounted for symmetries of the different molecular species[13], but ignored them in comparing the predictions of the model with results of the experiment[14]. Here, the molecular symmetries are considered explicitly. Finally, and most important, the S-R interaction $U^{(SR)}$, which were neglected in former treatments, are included in the present one. Indeed, our focus is on the effects of $U^{(SR)}$ on the stability of the bivalent complex relative to its monovalent counterpart.

In Appendix A of the Supporting Information we derive formulas for $q(RD^{(1)})$, $q(RD^{(2)})$, and $q(RD^{(1)}{}_2)$ that take into account the S-R interactions. In Appendix B we generate expressions for the binding constants (Ks) based on the qs. These are given in the fourth column of Table S1 of the Supporting Information.

2.3. Binding Efficiency of Bivalent Complex. The quantity typically employed as a measure of the potency of the ligand to induce a desired effect is EC_{50}, which is the effective concentration of the ligand needed to induce one-half the maximum effect. Equivalently, $EC_{50} = [D]_{1/2}$ is the concentration of the ligand D when the following condition holds:

$$[R]_{sat} / [R]_{all} = ½. \quad (1)$$

In eq (1) $[R]_{sat}$ stands for the total concentration of all saturated complexes (i.e., both binding sites of R are occupied by ligating units) and $[R]_{all}$ for the total concentration of all species involving R (i.e., including also Rs with unoccupied sites). In preceding work[12,13] following Hill[38] we invoked the "all-or-none" hypothesis to simplify the theoretical treatment, as well as to be in accord with the analysis of experimental data[14], which is also based on it. The all-or-none hypothesis, which assumes that either *all* sites of a given R are occupied by ligating units at once *or none* of the sites is occupied, is equivalent to ignoring the role of complexes of intermediate degrees of saturation.

In the present article, we eschew the all-or-none simplification. Hence, for the binding of bivalent D to bivalent R eq (1) can be written explicitly

$$\frac{[\mathrm{RD}_2^{(1)}]+[\mathrm{RD}^{(2)}]}{[\mathrm{R}]+[\mathrm{RD}^{(1)}]+[\mathrm{RD}_2^{(1)}]+[\mathrm{RD}^{(2)}]}=\frac{1}{2} \tag{2}$$

In terms of the overall binding constants K(X) of the complexes X we can recast eq (2) as

$$K(\mathrm{RD}_2^{(1)})[\mathrm{D}]_{1/2}^2+\{K(\mathrm{RD}^{(2)})-K(\mathrm{RD}^{(1)})\}[\mathrm{D}]_{1/2}-1=0\,, \tag{3}$$

where $[\mathrm{D}]_{1/2} = \mathrm{EC}_{50}$ is the concentration at which the Rs are 50% saturated. Substituting theoretical expressions for the Ks, which now account for molecular symmetry, and solving the resulting quadratic equation, we obtain:

$$[\mathrm{D}]_{1/2}=\frac{1}{4\nu_{\mathrm{M}}\eta_1}\left[2-\nu_{\mathrm{M}}C_{\mathrm{eff}}(R)\eta_2\eta_1^{-1}+\sqrt{(2-\nu_{\mathrm{M}}C_{\mathrm{eff}}(R)\eta_2\eta_1^{-1})^2+4}\right]. \tag{4}$$

In eq (4), $C_{\mathrm{eff}}(R)$ is the effective concentration of the unbound ligating unit of D at distance R from the other unit bound to a site of R (see eq (B.6) of Supporting Information); ν_{M} stands for the effective volume available to a ligating unit bound in a site of R (see eq (A13a)). The quantities $\eta_{\mathrm{j}} \equiv \langle\exp(-U^{(\mathrm{SR})}/k_{\mathrm{B}}T)\rangle_{\mathrm{j}}$ (where k_B is Boltzmann's constant and T is the absolute temperature) are averages of the Boltzmann factor over sub-ensembles (designated by index j) of conformations of the free S-chain restricted so that either one ligating unit (j = 1) or both units (j = 2) of the same D are bound to R (see eqs (B7) and (B9)).

In the limit of short S-chains, when $C_{\mathrm{eff}}(R)\to 0$, since the second unit of D cannot bind to the other site of R, and $\eta_1 \to 1$, since S-R interactions become negligible, eq (4) simplifies to

$$[\mathrm{D}(C_{\mathrm{eff}} = 0)]_{1/2} = (1+2^{½}) / (2\,\nu_{\mathrm{M}}). \tag{5}$$

Given the value of $[\mathrm{D}(C_{\mathrm{eff}} = 0)]_{1/2}$ in eq (5), one can estimate the value of ν_{M}, an important parameter characterizing the binding of the ligating unit to R. If the all-or-none hypothesis is invoked, as in our previous works[12,13], eq (5) reads either $[\mathrm{D}(C_{\mathrm{eff}} = 0)]_{1/2} = 1 / \nu_{\mathrm{M}}$ if symmetry numbers are ignored, or $[\mathrm{D}(C_{\mathrm{eff}} = 0)]_{1/2} = 1 / (2\nu_{\mathrm{M}})$ if symmetry numbers are included (see Table S2 in Supporting Information). Hence, refraining from the all-or-none assumption and accounting for molecular symmetry yields

a value of ν_M larger by a factor of $(1+2^{½})/2 \approx 1.2$ than the value used earlier[12,13] by invoking the all-or-none hypothesis and neglecting molecular symmetry.

Because the synthesis of very short PEG chains bound to cGMP is impractical, $[M]_{1/2}$ was measured for RET[14] and used instead of $[D(C_{eff} = 0)]_{1/2}$. The two concentrations can be related by noting that D carries two ligating units and is therefore twice as efficient at binding as M (equivalent to a single unit) at the same concentration, in the limit of very short S-chains. Hence, we have

$$[D(C_{eff} = 0)]_{1/2} \equiv [M]_{1/2} / 2, \quad (6)$$

which we take as the formal reference value of $[D(C_{eff} = 0)]_{1/2}$ for vanishing S-chain length. Therefore, to obtain a realistic estimate of the parameter ν_M, we use for RET the value $[D(C_{eff} = 0)]_{1/2} = 36$ μM rather than the value 72 μM that was estimated by Kramer and Karpen[14] and used by us previously[12,13]. With $[D(C_{eff} = 0)]_{1/2} = 36$ μM and the minimum value $[D]_{1/2}(Min) = 0.4$ μM, measured for RET[14], the maximum enhancement effect is $[D(C_{eff} = 0)]_{1/2} / [D]_{1/2}(Min) = 90$ rather than 180.

3. MODELS AND METHODS

3.1. Receptor. In light of the enormous variety and complexity of real molecular systems and lack of structural knowledge of the ligand-receptor complex for which experimental measurements of EC_{50} are available, we refrain from specifying atomic details and rather tailor a model of R with a few essential features that may characterize a broad class of multivalent Rs. We suppose R to possess a smooth surface (impenetrable to S-chain atoms) comprising hills (or depressions) with the binding sites on top (or at the center) (see Fig. 1). Far from the peaks of the hills the surface tends to a basal plane that extends formally to infinity, so that the effect of the membrane in which R is embedded is included. The topography between the sites is then either concave or convex, if the sites are on the peaks of the hills or at the centers of the depressions, respectively. For example, RET, a homo-tetrameric protein complex whose sites are located on each of the four monomers that surround a central ion channel, likely exhibits a concave topography (i.e., the entrance to the ion channel is in a valley created by the surrounding proteins; see Fig. 1 of ref 14).

We assume the two hills (depressions) carrying the binding sites possess cylindrical symmetry with Lorentzian profiles of height $h > 0$ (depth, $h < 0$) and full width at half height of 0.25 ρ, where ρ is the distance of the two sites. We take the origin of the coordinate system to be in the basal x-y plane of R and the z axis to coincide with the axis of the hill bearing site α (see Fig. 1a). Thus, the two sites α and β are located at $\mathbf{r}_\alpha = h\,\mathbf{e}_z$ and $\mathbf{r}_\beta = \rho\,\mathbf{e}_y + h\,\mathbf{e}_z$, respectively, where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ are Cartesian unit vectors. Hence, the surface of R is specified by the locus of points $\mathbf{r} = (x, y, z)$ fulfilling the condition

$$z = H_{\text{hard}}(x, y) = H_\alpha(x, y) + H_\beta(x, y), \tag{7a}$$

where

$$H_\lambda(x, y) = h\,\rho^2/[\rho^2 + 16\,x^2 + 16\,(y - y_\lambda)^2],\ y_{\lambda=\alpha} = 0,\ y_{\lambda=\beta} = \rho. \tag{7b}$$

We designate this model as R_{hard}.

3.2. Polymeric Spacer, Bivalent Ligand and Ligand-Receptor Complex. Several of the key experiments on multivalent ligand binding employ PEG chains as spacers[14,16,25,26]. A PEG chain (S-chain) comprising N_M monomers can be represented by the formula $H\,(–O–CH_2–CH_2)_{N_M}\,–OH$. The bivalent ligand D is constructed by covalently bonding the end atoms of the S-chain to the monovalent ligands M. A ligating unit binding to R is considered to be buried in the binding site, such that the corresponding end atom of the S-chain is constrained to the top of the hill bearing that site. Thus, $RD^{(1)}$ is constructed by fixing one end atom of the S-chain at position $\mathbf{r}_\alpha = h\,\mathbf{e}_z$ where site α is situated; $RD^{(2)}$ is constructed by fixing also the second end atom of the S–chain at site β at position $\mathbf{r}_\beta = \rho\,\mathbf{e}_y + h\,\mathbf{e}_z$ (see Fig. 1a). We model the PEG chain as a freely jointed chain of $3N_M + 1 = N_S$ identical concatenated beads that represent the O atoms and CH_2 groups of PEG. The C–C and O–C covalent bonds of the PEG chain, are taken to be equivalent and of the same length b = 1.53 Å.

3.3. Spacer-Receptor Interaction. Assuming that all non-hydrogen atoms (beads) of S interact independently with R according to the same potential ψ, we can express the S-R interaction as the sum

$$U^{(\text{SR})}(\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_{N_S}) = \sum_{i=1}^{N_S} \psi(\mathbf{r}_i)\,. \tag{8}$$

The potential energy $\psi(\mathbf{r}_i)$ of interaction of the ith atom of S with R is given by

$$\psi(\mathbf{r}_i) = \begin{cases} \infty, & z_i \le H_{\text{hard}}(x_i, y_i) \\ -W, & H_{\text{hard}}(x_i, y_i) < z_i \le H_{\text{soft}}(x_i, y_i) \\ 0, & z_i > H_{\text{soft}}(x_i, y_i) \end{cases}, \tag{9}$$

where $z = H_{\text{hard}}(x, y)$, eq (7a), describes the impenetrable "hard" surface of R. In a distance range $H_{\text{hard}} < z \le H_{\text{soft}}$ defined by a second, "soft" surface H_{soft} close to the hard surface, the atoms of the S-chain may be weakly attracted to R. Such an attractive region, modeled here as a square well, can account for potential hydrophobic interactions between S and R. To generate the soft surface of R, a sphere of radius $\sigma = 2\,b$ (b is the bond length between atoms of the S-chain) is rolled over the hard surface of R (see Fig. S1 of the Supporting Information). The locus of the center of the rolling sphere defines the soft surface. This procedure is commonly applied to proteins to define the solvent accessible surface area (SASA)[39,40]. A simple rolling-sphere algorithm can be employed if the protein volume is defined by a discrete number of atoms[41]. In the present application we use a different algorithm, storing the computed values of $H_{\text{soft}}(x, y)$ on a rectangular grid (see Appendix C of the Supporting Information for details).

In this study, we consider three different R models for the S-R interaction: (1) the hard R model (R_{hard}) with a hard wall only ($W = 0$) and different topologies of the R surface (variable h), (2) the planar R model (R_{planar}) ($h = 0$) with attractive square well potential (variable W), and (3) the non-planar "soft" R model (R_{soft}) ($h > 0$) with attractive S-R interactions (variable W), eq (9).

3.4. Computation of Effective Concentration C_{eff} and Ensemble Averages η_j.

The quantity of primary interest, EC_{50}, itself depends on $C_{\text{eff}}(R)$ and η_j, whose computation requires knowledge of the probability distribution (density) function of the free S-chain. This distribution is represented by an ensemble of N_{ens} conformations of the S-chain, each of which is generated by a continuous random walk comprising $N_S - 1$ random flights, as detailed in Appendix D of the Supporting Information.

For all models considered here, $U^{(SR)} = \infty$ if $z_i < H(x_i, y_i)$ for any atom i of the S chain (see eq (9)). Therefore, only those free S-chain conformations that do not penetrate the hard surface of R contribute to $\eta_j = \langle \exp(-U^{(SR)}/k_BT) \rangle_j$. The subscript j denotes ensembles for two different situations: j = 1 corresponds to one end of the S-chain fixed

on binding site α of R to form $RD^{(1)}$; j = 2 corresponds to both ends of the S-chain fixed on the sites of R to form $RD^{(2)}$. See Appendix E of the Supporting Information for details.

To control statistical errors, which depend on the model, we employ ensembles of free S-chain conformations ranging from $N_{ens} = 5\times10^6$ to $N_{ens} = 10^{10}$. We estimate statistical errors roughly by using the first and second halves of the generated free S–chain conformations to evaluate η_1 and η_2. Deviations between the two halves larger than the width of the lines in the corresponding figures are indicated by error bars. The statistical errors are negligible in the absence of S-R attraction (Fig. 2) but can become appreciable for long S-chains and large S-R attraction (Fig. 4b).

4. RESULTS AND DISCUSSION

Our principal concern is to explore the impact of the S-R interactions on the stability of the bivalent ligand-receptor complex relative to its monovalent counterpart. For this purpose we focus on EC_{50} (i.e. $[D]_{1/2}$). According to eq (4), $[D]_{1/2}$ depends on ν_M, the effective interaction volume available for a ligating unit in the binding site, $C_{eff}(R)$, the effective concentration of one ligating unit of D when the other unit is bound to one of the two sites of R as well as on η_1 and η_2. The ensemble averages η_1 and η_2 depend, according to eq (B9) of the Supporting Information, on $U^{(SR)}$, which in turn is determined by the model of R (i.e., either R_{hard}, R_{planar} or R_{soft}), as described in Section 3. We estimate a typical value of the effective interaction volume ν_M available for a ligating unit in the binding site of R from experimental data, rather than modeling the binding interaction of a ligating unit with R. In particular, we employ data of Kramer and Karpen on the binding of cGMP to RET[14]. Hence, we set $T = 300$ K, $\rho = 30$ Å, and $\nu_M \approx 0.0335\ \mu M^{-1}$. This estimate of the value of ν_M is based on the measured value of $[M]_{1/2} = 72\ \mu M$[14].

Equation (4) can be rewritten as $[D]_{1/2} = (4\nu_M \eta_1)^{-1}\{2-q+[(2-q)^2+4]^{1/2}\}$, where

$$q = \nu_M C_{eff} \eta_2/\eta_1. \qquad (10)$$

Thus, $[D]_{1/2}$ manifests two dependences on ν_M through the pre-factor ν_M^{-1} and the parameter $q(\nu_M)$ ($q \geq 0$). The inverse proportionality to ν_M reflects monovalent binding in absence of multivalency (i.e., the binding to R of ligating units from different Ds). The enhancement effect of multivalent binding depends crucially on $q(\nu_M)$. It is the key

parameter governing the efficiency of multivalent binding with flexible spacers. If monovalent binding is too weak (i.e. ν_M is small), there is no enhancement effect. Indeed, as indicated in Fig. S3 of the Supporting Information, an appreciable enhancement effect is possible only if q is much greater than 2. Hence, in this regime of q, which corresponds to the minimum of the plot of $[D]_{1/2}$ versus Flory radius R_F (root-mean-square end-to-end distance for free ligand D, eq (D3) of the Supporting Information), an increase in ν_M results in an increase in enhancement effect. Accordingly, the estimate of ν_M that we now use, which is a factor of ~ 1.2 larger than that of ν_M employed in our earlier work[12,13], leads to a substantial increase in the enhancement effect.

The parameter $q(\nu_M)$ [eq (10)] is proportional to ν_M by the factor $C_{eff}\,\eta_2/\eta_1$. It reflects the binding of the second ligating unit of D to R. C_{eff} describes the essence of the enhancement effect (i.e. the enhanced concentration of the second unit of a D whose first unit is already bound to R). The ratio η_2/η_1 describes how the S-R interactions modulate the enhancement effect. It is generally smaller than unity, thus diminishing the enhancement effect appreciably. However, as demonstrated below, the ratio η_2/η_1 grows with increasing concavity of the R surface and attraction between S and R.

We stress that the ensemble averages η_j are the sole parameters of the model that reflect the influence of the S-R interaction on the enhancement effect, as measured by EC_{50}. It is shown in Appendix G of the Supporting Information that the contribution of the S-R interaction to the free energy of binding for $RD^{(j)}$ is $\Delta F_j = -k_B T \ln(\eta_j)$ for j = 1, 2.

4.1. Strictly Repulsive Surface R_{hard}: Variation of Height of Hills. We examine first the dependence of η_j on the height (depth) (h) of the two hills (depressions) of R_{hard} (see eq (7)). Hence, $\Delta F_j = -T\,\Delta S_j$, where $\eta_j = \exp(\Delta S_j/k_B)$ (see Appendix G of Supporting Information) is just the fraction of conformations of the S-chain for which *all* beads lie above the hard surface of R ($z = H_{hard}(x, y)$). Accordingly, ΔS_j is the loss of entropy of the free S-chain of D due to its interaction with R, as the j (= 1, 2) ligating units bind to the sites to form $RD^{(j)}$.

Figs. 2a and 2b respectively display plots for R_{hard} of $\Delta S_1/k_B$ and $\Delta S_2/k_B$ versus S-chain length (N_S) for several heights h of the hills. The horizontal lines ($h = \infty$, $\Delta S_j = 0$) correspond to absence of S-R interactions, as assumed in our prior work[12,13]. As h decreases, the entropy loss increases markedly for both complexes $RD^{(1)}$ and $RD^{(2)}$. Note however, that the rate of entropy loss for $RD^{(2)}$ is nearly twice that for $RD^{(1)}$, because fewer free S-chain conformations survive the requirement that all atoms lie above $H_{hard}(x, y)$ when both ends of the S-chain are bound to the sites of R. We note in passing that the dependences of η_j on N_S approximately obey a power law $\eta_j = \exp(\Delta S_j/k_B) \sim (N_S)^{-j/2}$ as $N_S \to \infty$ (see Fig. S2 of the Supporting Information). For j = 1, the corresponding power law was previously reported for a polymer chain with one end attached to a planar hard wall[42].

We conclude that the hard surface of R engenders strong decreases in the entropy of the S-chain as D binds to R. For positive h, the landscape of R between the two binding sites is concave. With decreasing h the concavity becomes weaker and the entropy loss stronger. The entropy loss becomes dramatically large if R possesses a convex landscape ($h < 0$). Thus, the decrease in entropy as h goes from 0 to -1 is approximately equal to that as h goes from 10 to 0 (see Fig. 2).

This entropy loss gives rise to lower stability of the complexes. Plots of $[D]_{1/2}$ versus the Flory radius R_F (root mean square end-to-end distance of the S-chain, defined by eq (D3) of the Supporting Information) demonstrate this effect (see Fig. 3). As the height h of the hills decreases, the entropy loss of the S-chain increases and $[D]_{1/2}$ rises correspondingly (Fig. 3). For short S-chains (i.e., small R_F), where the second ligating unit is still unable to bind, $[D]_{1/2}$ increases with increasing R_F because of the entropy loss of the S–chain due to its interaction with the R surface. Likewise, for long S-chains (i.e., large R_F) $[D]_{1/2}$ increases again, since C_{eff} decreases. At intermediate R_F, $[D]_{1/2}$ goes through a minimum, and the enhancement effect is maximal here. The minimum in $[D]_{1/2}$ is deeper for the measured than for the computed $[D]_{1/2}$.

In the absence of the S–R interaction ($h = \infty$) the computed minimum of $[D]_{1/2}$ is significantly below the minimum of the measured $[D]_{1/2}$ and located exactly at $R_F = \rho = 30$ Å corresponding to the estimated distance[14] between the two binding sites. However, with increasing S-R interaction (smaller h) the minimum of $[D]_{1/2}$, albeit

shallow, shifts to smaller R_F values (Fig. 3). Although, the curve of $[D]_{1/2}$ in Fig. 3 labeled $h = \infty$ is based on the model with vanishing S-R interaction[12,13] the enhancement effect is now considerably larger, because the present estimated value of v_M, which avoids the all-or-none hypothesis and accounts for molecular symmetry, is greater than the previous one by a factor of 1.2. Thus, room is available for additional modifications of the present model that may diminish the enhancement effect. For the planar R surface (h = 0 Å) the enhancement is negligibly small, while for R with weakly convex surfaces (h = -1 Å), binding of D becomes dramatically hampered compared with binding of M (see Fig. 3).

As Fig. 3 indicates, we need pronounced concavity of the R landscape with hill heights of at least h = 10 Å in order to qualitatively describe the enhancement measured for the activation of RET[14]. If we wanted to reproduce the measured maximum enhancement of $[D(C_{eff} = 0)]_{1/2} / [D]_{1/2}(Min) = 90$, according to Fig. 3, we would need to set $h \sim 30$ Å, which may be unreasonably large. Therefore, we consider other options, namely weak attractive interactions of the PEG chain with the protein R surface.

4.2 Spacer-Receptor Attraction: R_{planar} and R_{soft} models. If the spacer material were strictly hydrophilic, there would be no attraction between S and R. However, there are strong indications that PEG is not perfectly hydrophilic. PEG chains generally repel proteins by loss of conformational entropy, as long as the PEG-protein interface is relatively small. As this interface grows, weak hydrophobic and van der Waals attractions between PEG and protein may become significant[31-35]. Several independent lines of evidence point to the ability of PEG to attract non-polar as well as polar regions of proteins[36]. When PEG is used to foster crystallization of proteins, it may be observed in the crystal structures. For example, structurally ordered PEG chains have been found inside the ion channel of OmpF porin[37]. Such attractive interactions of PEG with protein surfaces were also observed for the cGMP activated ion channels[14]. For RET, it was observed that monomeric cGMP with attached PEG chain binds as efficiently as the bare cGMP monomer, despite the entropy loss that a PEG chain experiences in the neighborhood of the R surface. For the olfactory ion channel the monomeric cGMP with attached PEG chain binds to the receptor even more efficiently than the bare monomeric cGMP[14]. Based on these experimental results, we may conclude that the entropy loss of

the PEG chain upon binding of cGMP is compensated by attractive interactions between the PEG chain and the surface of RET.

Therefore, such strictly repulsive R models as R_{hard} may be unrealistic in that they do not account for the weak attractive forces that come into play as the S-chain approaches the surface of R. To explore the influence of an attractive contribution to the S-R interaction, we employ first a simple R model (R_{planar}) consisting of a hard plane with an attractive square well potential (width σ and depth W) next to it. We vary the well depth W and fix the width to $\sigma = 2\,b = 3.06$ Å, which roughly corresponds to the thickness of a single atomic layer at the R surface. Implicit solvent models of solutes in water mimic the hydrophobic effect[43] usually by a surface energy term whose of strength varies from 0.012 kcal/(mol Å^2) for small molecules[44,45] to 0.030 kcal/(mol Å^2), a value used for proteins[46,47]. The hydrophobic effect acting on a solute (in our case receptor and spacer) is proportional to the SASA of the solutes. Taking the radius of the atoms of the S-chain to be $b = 1.53$ Å, we estimate the effective decrease in SASA for one atom of S in contact with the R surface to be $2\pi\, b^2 = 14.7$ Å^2. Thus, we estimate that when one atom of the hydrophobic S is in contact with the protein R surface, its free energy decreases by about $W = 0.44$ kcal/mol. Since PEG is only weakly hydrophobic, the decrease in free energy should be much smaller. We indeed find that values of W one order of magnitude smaller than this estimate are large enough to explain the experiments.

In Figs. 4a and 4b are plotted $\ln(\eta_1) = -\Delta F_1/(k_B T)$ and $\ln(\eta_2) = -\Delta F_2/(k_B T)$ as functions of the S–chain length (N_S) for the R_{planar} model for several well depths W at fixed well width $\sigma = 2\,b = 3.06$ Å corresponding to two bond lengths of S. The curves labeled $W = 0.000$ in Fig. 4a and 4b correlate with those in Fig. 2a and 2b labeled 0 ($h = 0$). As the well gets deeper, $-\Delta F_j$ falls off more slowly with increasing N_S, eventually becoming almost flat at $W = 0.045$ kcal/mol, where enthalpy gain compensates entropy loss over a large interval of N_S. This condition likely corresponds to the experimental observation that the binding to RET is for monovalent cGMP with an attached PEG chain as efficient as for a bare cGMP[14]. In fact this value of W is about ten times smaller than the above estimate of 0.44 kcal/mol for the free energy of association of an atom of a strongly hydrophobic molecular species with a protein surface.

For the R_{planar} model, the influence of the attractive part of the S-R interaction on the enhancement effect is demonstrated in Fig. 5, which shows plots of $[D]_{1/2}$ versus Flory radius R_F that correspond to those of Fig. 3. According to our expectation, $[D]_{1/2}$ decreases as the attractive square well deepens, such that at $W = 0.045$ kcal/mol the enhancement becomes as large as the maximum enhancement of $[D(C_{eff} = 0)]_{1/2} / [D]_{1/2}(Min) = 90$ measured for RET. Furthermore, this value of W coincides with the value for which the monovalent ligand with an attached polymer chain binds as efficiently to R as M alone. See Fig. 4a, which shows that $-\Delta F_1$ nearly vanishes and does not substantially change with the S-chain length. This agrees with measurements on the binding of cGMP to RET[14].

In the R_{soft} model, we combine both concavity and S-R attraction so as to make the enhancement agree with that experimentally observed for RET. In Fig. 6 we show the dependence of $[D]_{1/2}$ on Flory radius R_F for different well depths W with a fixed, moderate concavity of the R surface (h = 10 Å, depicted in Fig. 1a). For this degree of concavity an even smaller S-R attraction of $W = 0.020$ kcal/mol yields an enhancement $\{[D(C_{eff} = 0)]_{1/2} / [D]_{1/2}(Min)\}$ as large as that measured for RET[14] (Fig. 6). We can conclude that even a very small attractive S-R component corresponding to weak hydrophobicity can lead to a large enhancement effect in multivalent ligand binding.

5. CONCLUSIONS

Properly designed multivalent ligands combined with appropriately chosen multivalent receptors bind often by orders of magnitude more efficiently than their monovalent analogs[14-29]. A prime example is the binding of polymer-linked bivalent ligands (cGMP moieties connected by PEG chains) to tetravalent receptors such as RET[14]. The increase in efficiency of multivalent binding is rationalized in terms of the "effective concentration" (C_{eff}) of ligating units. If some units of the multivalent ligand are already bound to R, the effective concentration of the remaining unbound ligating units at the unoccupied binding sites of R can be much greater than that of free ligands in solution. Thus, for a bivalent ligand it is more probable that the second unit binds to an available site of the same R than that another ligands bind from solution. A simple

model[12,13] incorporating this concept was able to explain semi-quantitatively the measured dependence of EC_{50} ($[D]_{1/2}$) on the length of the PEG S–chain.

For the sake of simplicity the S-R interaction was neglected in our earlier studies on bivalent binding[12,13]. The present work includes these interactions. To the best of our knowledge preceding theoretical investigations on multivalent binding ignored S-R interactions, except for one very recent study[30]. Furthermore, we refrain here from invoking Hill's simplifying all-or-none assumption[38], which ignores the semi-saturated ligand-receptor complex $RD^{(1)}$, where one binding site is occupied, while the second site is empty. Avoiding the all-or-none assumption and accounting for molecular symmetry increases the enhancement effect of bivalent binding by more than an order of magnitude, leaving room for additional factors in improved models that may reduce enhancement. In fact, the S-R interaction, absent in previous studies, has a strong influence and can reduce the efficiency of multivalent ligand binding substantially. Furthermore, we have shown that the enhancement effect of bivalent binding with a flexible spacer is governed by a single parameter q [see eq (10)]. The dependence of this parameter on monovalent binding efficiency (ν_M), effective concentration (C_{eff}) and S-R interactions provides guidelines for the design of bivalent ligand-receptor complexes that can exhibit strong enhancement. We believe that such guidelines can also be extended to more general cases of multivalent binding.

The R with a hard surface (R_{hard} model) forces the atoms of S to stay outside of R, thus diminishing the number of allowed S-chain conformations and dramatically lowering the S-chain entropy (or, equivalently, increasing the free energy) as the S-chain approaches the surface of R. As a consequence, the enhancement effect can be reduced by several orders of magnitude, even to the point of rendering the binding of monomeric ligands (M) more efficient than that of bivalent ligands. The reduction of the enhancement effect by S-R interaction is most pronounced for convex R surfaces, very large for planar R surfaces and still significant for concave R surfaces. Only for unreasonably large concavity of the R surface is the computed enhancement sufficiently large to explain that seen for RET[14]. Hence, an attractive component of the S-R interaction is necessary to understand efficient bivalent binding for such complexes.

If an attractive layer next to the hard surface of R is included, the enhancement effect becomes larger again. Already a weak S-R attraction only one tenth the magnitude of that between hydrophobic aliphatic carbon atoms, can restore the enhancement for the planar R surface (R_{planar} model) to the value obtained in absence of S-R interaction and thus reproduce the measured enhancement for the binding of bivalent cGMP to RET[14]. If concavity of R is combined with S-R attraction, a moderate concavity and very small S-R attraction is also sufficient to explain the enhancement effect of bivalent binding. In fact there are a number of indications that PEG has a tendency to attach to protein surfaces and therefore may be slightly hydrophobic. The monomeric cGMP binds to RET with the same efficiency with and without an attached PEG chain[14], which corroborates this behavior. Hence, the bivalent binding of cGMP to RET goes along with conformational entropy loss and gain of binding enthalpy of the spacer due to weak S-R interactions. These two contributions to binding free energy approximately compensate for concave receptor topography. That is why simplified earlier theoretical models of bivalent ligand binding that ignored the S-R interaction were able to provide fortuitous agreement with the measurements for these systems. Such compensation between entropy and binding enthalpy may hide the influence of S-R interactions on multivalent binding also for other systems.

A depression, or valley, between the binding sites of R, as depicted in Fig. 1, results in a concave R surface. This topography is likely not widespread in the universe of protein structures. In fact (based on a preliminary scan, work in progress), bivalent Rs that possess a concave surface between the two ligand binding sites are scarce in the protein data bank[48]. Prominent examples of such receptors with concave topography[49] are the homo- and hetero-dimeric 14-3-3 proteins[50], which are essential for signaling in all eukaryotes and appear also in plants[51]. Figure 7 depicts the concavity of the R-surface between the two sites of a bivalent 14-3-3 protein homodimer (PDB id 3RDH), which is shown binding two drug-like molecules of a recently discovered inhibitor[52]. For artificial tandem peptide ligands involving phosphoserine, it was shown that binding to the 14-3-3 protein can be 30 times more efficient than for the monovalent correlates[53]. Very recently, an artificial bivalent receptor was created by dimerizing carbonic anhydrase (CA)[54]. The binding efficiency of bivalent sulfonamide ligands with this receptor (dime5rCA) was enhanced by up to a factor of 5000.[54] According to the crystal structures

of dCA the two binding sites are face-to-face corresponding to concave receptor topography, which is consistent with our results.

Another homo-dimeric receptor on which several bivalent ligand binding studies have been performed is the estrogen receptor (ER)[55,56]. The topography between the two binding sites of the dimeric ER is neither concave nor planar, but strongly corrugated as can be seen in Fig. 8. As a consequence, the S-chain needs to form a bow to stay away from the ER surface or to adopt a complex non-linear conformation that makes close contact between S and the ER surface. Hence, it is not surprising that several experimental attempts to produce an appreciable enhancement for the binding of bivalent ligands with flexible spacers have so far not been successful for this system.

In summary, the difficulties we encounter in the theoretical description of the binding of D to R with explicit S-R interactions may explain the paucity of experiments that successfully demonstrate an enhancement effect for bivalent complexes in which the spacer is flexible. Use of rigid spacers could be advantageous in cases where the topography of the receptor is not concave, although a major problem with rigid spacers is that they must be adroitly designed so as to be commensurate with the topography of the corresponding receptor. The influence of S-R interactions on multivalent ligand binding was so far mostly overlooked or considered to be negligibly small. In fact, its influence can abolish enhancement of binding or even yield bivalent binding affinities lower than in the monomeric case.

While the spacer-receptor models used in the present study are still based on very simplifying assumptions, they provide general guidelines for more detailed molecular models. For future work, it will be useful to perform simulations in atomic detail in order to enhance the understanding of multivalent binding. Computing the free energy of binding of a multivalent ligand to its receptor for such a detailed model requires molecular dynamics (MD) simulations of more than 10,000 atoms. For the present study, we needed up to 10^{10} independent spacer conformations generated by a Monte Carlo method to obtain precise enough data for the bivalent ligand binding. Assuming that an independent spacer conformation is generated after each picosecond of an MD simulation (actually a very optimistic guess), one would still need a trajectory of 10 ms duration to theoretically model the bivalent binding process with sufficient accuracy. Such MD simulations require an enormous amount of CPU time. To monitor spacer length

dependencies as done in the present study, several such trajectories would be needed. To plan such expensive MD simulations, the present study can pave the way to choose the right conditions.

ACKNOWLEDGEMENT

We thank Tim Meyer for helping to prepare Fig. 7. This work is supported by the collaborative research center SFB 765 project C1 from the Deutsche Forschungsgemeinschaft (DFG). DJD thanks the science administration of the Freie Universität Berlin for travel grants.

Supporting Information available: Figures S1-S3, Appendices A through G and Tables S1-S3 are provided. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

(1) Huskens, J. *Curr. Opin. Chem. Biol.* **2006**, *10*, 537-543.
(2) Mammen, M.; Choi, S.-K.; Whitesides, G. M. *Angew. Chem. Int. Ed.* **1998**, *37*, 2754-2794.
(3) Kiessling, L. L.; Gestwicki, J. E.; Strong, L. E. *Angew. Chem. Int. Ed.* **2006**, *45*, 2348-2368.
(4) Collins, B. E.; Paulson, J. C. *Curr. Opin. Chem. Biol.* **2004**, *8*, 617-625.
(5) Joshi, A.; Vance, D.; Rai, P.; Thiyagarajan, A.; Kane, R. S. *Chem. Eur. J.* **2008**, *14*, 7738-7747.
(6) Rolland, O.; Turrin, C.-O.; Caminade, A.-M.; Majoral, J.-P. *New J. of Chem.* **2009**, *33*, 1809-1824.
(7) Mulder, A.; Huskens, J.; Reinhoudt, D. N. *Org. Biomol. Chem* **2004**, *2*, 3409-3424.
(8) Badjić, J. D.; Nelson, A.; Cantrill, S. J.; Turnbull, W. B.; Stoddart, J. F. *Acc. Chem. Res.* **2005**, *38*, 723-732.
(9) Schalley, C. A.; Lützen, A.; Albrecht, M. *Chem. Eur. J.* **2004**, *5*, 1072-1080.
(10) Baldini, L.; Casnati, A.; Sansone, F.; Ungaro, R. *Chem. Soc. Rev.* **2007**, *36*, 254-266.
(11) Martinez-Veracoechea, F. J.; Frenkel, D. *Proc. Natl. Acad. Sci. U.S.A* **2011**, *108*, 10963-10968.
(12) Diestler, D. J.; Knapp, E. W. *Phys. Rev. Lett.* **2008**, *100*, 178101.
(13) Diestler, D. J.; Knapp, E. W. *J. Phys. Chem. C* **2010**, *114*, 5287-5304.
(14) Kramer, R. H.; Karpen, J. W. *Nature* **1998**, *395*, 710-713.
(15) Blaustein, R. O.; Cole, P. A.; Williams, C.; Miller, C. *Nat. Struct. Biol.* **2000**, *7*, 309-311.

(16) Loidl, G.; Groll, M.; Musiol, H.-J.; Huber, R.; Moroder, L. *Proc. Natl. Acad. Sci. U.S.A* **1999**, ***96***, **5418-5422**.
(17) Kitov, P. I.; Sadowska, J. M.; Mulvey, G.; Armstrong, G. D.; Ling, H.; Pannu, N. S.; Read, R. J.; Bundle, D. R. *Nature* **1999**, *403*, 669-672.
(18) Fan, E.; Zhang, Z.; Minke, W. E.; Hou, Z.; Verlinde, C. L. M. J.; Hol, W. G. J. *J. Am. Chem. Soc.* **2000**, *122*, 2663-2664.
(19) Gargano, J. M.; Ngo, T.; Kim, J. Y.; Acheson, D. W. K.; Lees, W. J. *J. Am. Chem. Soc.* **2001**, *123*, 12909-12910.
(20) Zhang, Z.; Merritt, E. A.; Ahn, M.; Roach, C.; Hou, Z.; Verlinde, C. L. M. J.; Hol, W. G. J.; Fan, E. *J. Am. Chem. Soc.* **2002**, *124*, 12991-12998.
(21) Kitov, P. I.; Shimizu, H.; Homans, S. W.; Bundle, D. R. *J. Am. Chem. Soc.* **2003**, *125*, 3284-3294.
(22) Mulder, A.; Auletta, T.; Sartori, A.; Ciotto, S. D.; Casnati, A.; Ungaro, R.; Huskens, J.; Reinhoudt, D. N. *J. Am. Chem. Soc.* **2004**, *126*, 6627-6636.
(23) Trevitt, C. R.; Craven, C. J.; Milanesi, L.; Syson, K.; Mattinen, M.-L.; Perkins, J.; Annila, A.; Hunter, C. A.; Waltho, J. P. *Chem. and Biol.* **2005**, *12*, 89-97.
(24) Farrera, J.-A.; Hidalgo-Fernández, P.; Hannink, J. M.; Huskens, J.; Rowan, A. E.; Sommerdijk, N. A. J. M.; Nolte, R. J. M. *Org. Biomol. Chem.* **2005**, *3*, 2393-2395.
(25) Krishnamurthy, V. M.; Semetey, V.; Bracher, P. J.; Shen, N.; Whitesides, G. M. *J. Am. Chem. Soc.* **2007**, *129*, 1312-1320.
(26) Shewmake, T. A.; Solis, F. J.; Gillies, R. J.; Caplan, M. R. *Biomacromolecules* **2008**, *9*, 3057-3064.
(27) Kane, R. S. *Langmuir* **2010**, *26*, 8636-8640.
(28) Huskens, J.; Mulder, A.; Auletta, T.; Nijhuis, C. A.; Ludden, M. J. W.; Reinhoudt, D. N. *J. Am. Chem. Soc.* **2004**, *126*, 6784-6797.
(29) Zhou, H.-X. *Biophys. J.*. **2006**, *91*, 3170-3181.
(30) Wang, S.; Dormidontova, E. E. *Biomacromolecules* **2010**, *11*, 1785-1795.
(31) Sheth, S.; Leckband, D. E. *Proc. Natl. Acad. Sci. U.S.A* **1997**, *94*, 8399-8404.
(32) Israelachvili, J. *Proc. Natl. Acad. Sci. U.S.A* **1997**, *94*, 8378-8379.
(33) Vivarès, D.; Belloni, L.; Tardieu, A.; Bonneté, F. *Eur. Phys. J. E* **2002**, *9*, 15-25.
(34) Sheth, S.; Efremova, N.; Leckband, D. E. *J. Phys. Chem. B* **2000**, *104*, 7652-7662.
(35) Leckband, D.; Israelachvili, J. *Q. Rev. Biophys.* **2001**, *34*, 105-267.
(36) Zhou, H.-X.; Rivas, G.; Minton, A. P. *Annu. Rev. Biophys.* **2008**, *37*, 375-397.
(37) Dhakshnamoorthy, B.; Raychaudhury, S.; Blachowicz, L.; Roux, B. *J. Mol. Biol.* **2010**, *396*, 293-300.
(38) Hill, A. V. *J. Physiol.* **1910**, *40 (Suppl)*, iv-vii.
(39) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379-400.
(40) Connolly, M. L. *Science* **1983**, *221*, 709-713.
(41) Shrake, A.; Rupley, J. A. *J. Mol. Biol.* **1973**, *79*, 351-364.
(42) Dacheng, W.; Kang, J. *Science in China B* **1996**, *39*, 608-617.
(43) Chandler, D. *Nature* **2005**, *437*, 640-647.
(44) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243-9244.
(45) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A* **2006**, *103*, 8331-8336.
(46) Hermann, R. B. *Proc. Natl. Acad. Sci. U.S.A* **1977**, *74*, 4144-4145.
(47) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. *Science* **1991**, *252*, 106-109.

(48) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235-242.
(49) Yaffe, M. B. *FEBS Letters* **2002**, *513*, 53-57.
(50) Liu, D.; Bienkowska, J.; Petosa, C.; Collier, R. J.; Fu, H.; Liddington, R. *Nature* **2002**, *376*, 191-194.
(51) Xiao, B.; Smerdon, S. J.; Jones, D. H.; Dodson, G. G.; Soneji, Y.; Aitken, A.; Gamblin, S. J. *Nature* **2002**, *376*, 188-191.
(52) Zhao, J.; Du, Y.; Horton, J. R.; Upadhyay, A. K.; Lou, B.; Bai, Y.; Zhang, X.; Du, L.; Li, M.; Wang, B.; Zhanga, L.; Barbieri, J. T.; Khuri, F. R.; Cheng, X.; Fu, H. *Proc. Natl. Acad. Sci. U.S.A* **2011**, *108*, 16212-16216.
(53) Yaffe, M. B.; Rittinger, K.; Volinia, S.; Caron, P. R.; Aitken, A.; Leffers, H.; Gamblin, S. J.; Smerdon, S. J.; Cantley, L. C. *Cell* **1997**, *91*, 961-971.
(54) Mack, E. T.; Snyder, P. W.; Perez-Castillejos, R.; Bilgiçer, B.; Moustakas, D. T.; Butte, M. J.; Whitesides, G. M. *J. Am. Chem. Soc.* **2012**, *134*, 333-345.
(55) LaFratea, A. L.; Carlsona, K. E.; Katzenellenbogen, J. A. *Bioorg. Med. Chem.* **2009**, *17*, 3528-3535.
(56) Abendroth, F.; Bujotzek, A.; Shan, M.; Haag, R.; Weber, M.; Seitz, O. *Angew. Chem. Int. Ed.* **2011**, *50*, 8592-8596.
(57) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. *Cell* **1998**, *95*, 927-937.

FIGURE CAPTIONS

FIG. 1a: Schematic side view of bivalent receptor (R_{hard}) with binding sites occupied by ligating units of bivalent ligand connected by polymeric spacer. Hard surface of R defined by function $z = H_{hard}(x, y)$, eq (7) ($\rho = 30$, $h = 10$ Å), where h is height and ρ is the distance between two Lorentzian hills (of 0.25 ρ half width at half height) on top of which are binding sites. Function H_{hard} represents distance from hard surface of receptor to basal x-y plane. Second surface (H_{soft}, dashed line) bounds range of distances $H_{hard} < z < H_{soft}$ over which atoms of S may be subject to attractive square well of fixed width $\sigma = 2b = 3.06$ Å and variable depth W.

b: Three-dimensional perspective of R surface concave between binding sites. Displayed are H_{hard} (solid) and H_{soft} (transparent) with bivalent ligand D (including S-chain) bound to binding sites located on tops of two Lorentzian hills.

FIG. 2a: Entropy loss ΔS_1 (equivalent to free energy difference $-\Delta F_1 = T\,\Delta S_1$) due to S-R interaction $U^{(SR)}$ on binding of one ligating unit of D to R, as function of S-chain length (in number of atoms N_S) for hard receptor model (R_{hard}) with different heights h [Å] of Lorentzian hills ($h = -1$, 0, 5, 10, 30, ∞). $h = -1$ Å corresponds to Lorentzian depressions of depth 1 Å; $h = \infty$ corresponds to absence of S-R interaction. Distance between binding sites is $\rho = 30$ Å. $\Delta S_1/k_B = \ln(\eta_1)$ is computed over an ensemble of between 3×10^{10} and 3×10^{11} conformations of free S-chains, constrained so that one end atom is fixed at binding site α of R. Ensemble average $\eta_1 = \langle \exp(-U^{(SR)}/k_BT) \rangle_1$ is equivalent to fraction of S-chain conformations for which all atoms lie above H_{hard}, eq (7) (see Fig. 1).
2b: Entropy loss ΔS_2 (equivalent to free energy difference $-\Delta F_2 = T\,\Delta S_2$) due to S-R interaction on binding of both ligating units of D to R, as function of S-chain length (in number of atoms N_S) for R_{hard} model. Ensemble average $\eta_2 = \langle \exp(-U^{(SR)}/k_BT) \rangle_2$

[yielding $\Delta S_2/k_B = \ln(\eta_2)$] is over conformations of free S-chains constrained so that end atoms are fixed at binding sites of R. Same notation and parameters as in Fig. 2a.

FIG. 3: $[D]_{1/2}$ for binding of D to R for R_{hard} model, as function of Flory radius R_F of S-chain for different heights of Lorentzian hills ($h = -1$, 0, 5, 10, 30, ∞), eq (7). Results correspond to those of Figs. 2a and 2b. Open circles 'o' refer to experimental data on activation of RET[14]. Distance between binding sites fixed at $\rho = 30$ Å, as suggested in ref 14. Symbol 'x' at $R_F = 0$ marks measured value $[M]_{1/2} = 72$ μM for monomeric cGMP[14]. Reference value for limiting concentration (open circle at $R_F = 0$) is $[D(C_{eff} = 0)]_{1/2} = 36$ μM. Curve labelled $h = \infty$ refers to previously proposed model[12,13], where we neglected the S-R interactions and applied the all-or-none hypothesis[38]. Note, however, in contrast to preceding study, we consider here *bivalent* instead of *tetravalent* R.

FIG. 4a: Change in (negative) free energy due to S-R interaction on binding of one ligating unit of bivalent ligand to bivalent R for R_{planar} model ($h = 0$) , as function of S–chain length (in number of atoms N_S). S-R interaction is modelled by attractive square well potential next to R surface. Well depth W [kcal/mol] varies between 0.0 and 0.045. Distance between binding sites is $\rho = 30$ Å.
4b: Change in (negative) free energy due to S-R interaction on binding of both ligating units of bivalent ligand to bivalent R, as function of S-chain length. Error bars shown where the statistical error larger than line width (i.e. for W=0.045 at large N_S). Same R_{planar} model, notation and parameters as in Fig. 4a.

FIG. 5: $[D]_{1/2}$ of binding of D to R, as a function of the Flory radius R_F for the R_{planar} model with different well depths W [kcal/mol]. Results correspond to those of Figs. 4a and 4b. Dotted reference curves labelled $h = 0$ and $h = \infty$ with $W = 0.0$ are reproduced from Fig. 3. Open circles refer to experimental data on RET[14].

FIG. 6: $[D]_{1/2}$ of binding of D to R, as function of the Flory radius R_F for the R_{soft} model with different well depths W [kcal/mol] and fixed height $h = 10$ Å of Lorentzian hills, eq (7). Negative well depths correspond to S-R repulsion. Dotted reference curve labelled $h = \infty$ and $W = 0.0$ reproduced from Fig. 3. Open circles refer to experimental data on activation of RET[14].

FIG. 7: Concave R-surface between the two binding sites of the bivalent 14-3-3 dimeric receptor (PDB 3RDH) pictured binding two FOBISIN101 ligands (magenta)[52]. The backbones of the two polypeptides are traced by rubber bands in green and yellow. The transparent gray shaded area pictures the protein volume including the side chains.

FIG. 8: Corrugated R-surface between the two binding sites of the bivalent homodimeric estrogen receptor (ER), shown binding a bivalent ligand consisting of two diethylstilbestrol (DES) ligating units connected by a PEG spacer of 21 main-chain atoms. The spacer geometry is modelled on the basis of the ER crystal structure (PDB id 3ERD)[57]. The PEG spacer needs to circumvent the two α-helices, which protrude from the ER surface. The same representation is used for the protein as for Fig. 7.

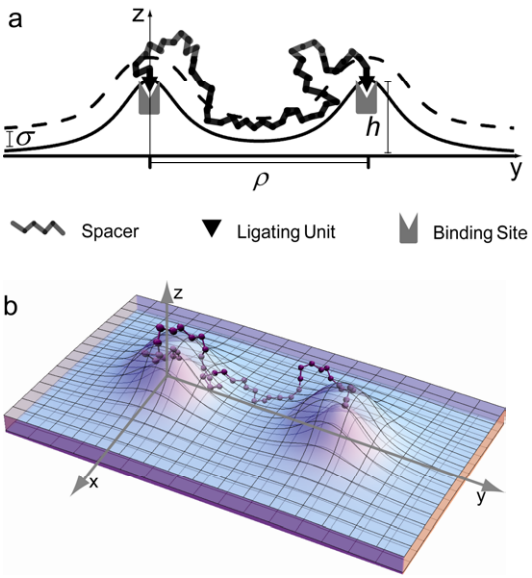

Fig. 1a, b

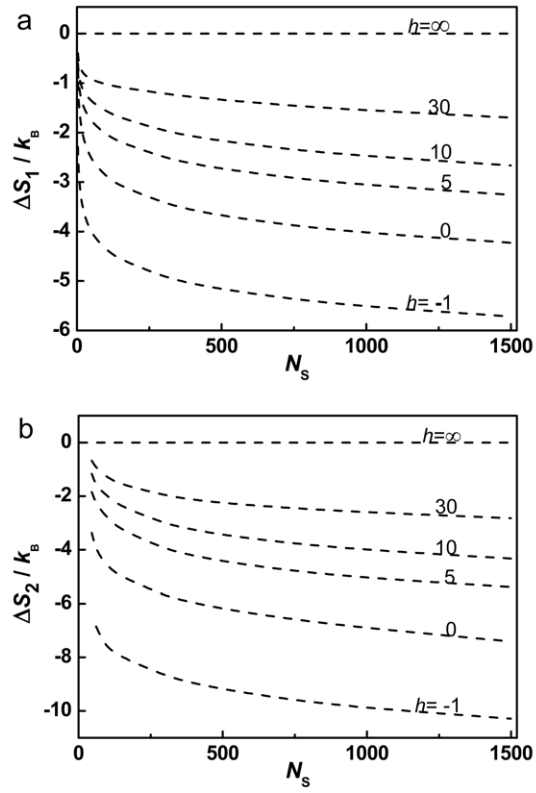

Fig. 2: a, b

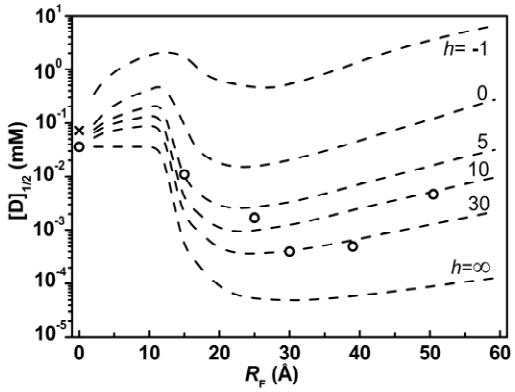

Fig. 3:

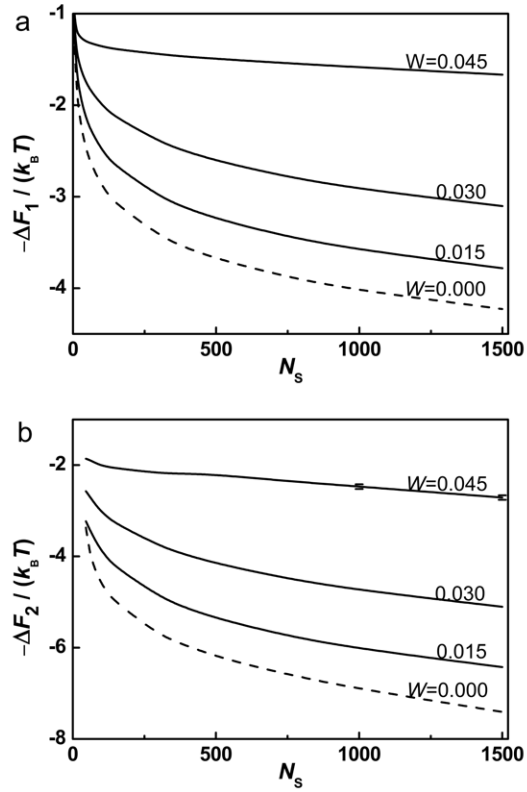

Fig. 4: a, b

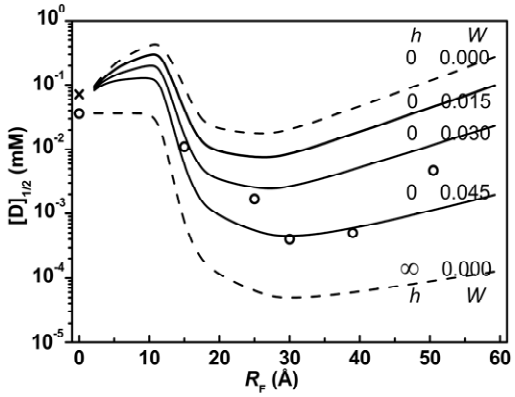

Fig. 5

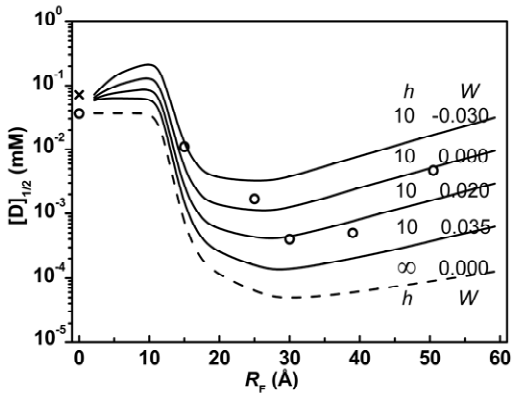

Fig. 6

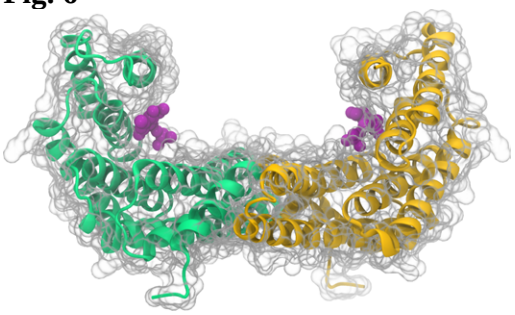

Fig. 7

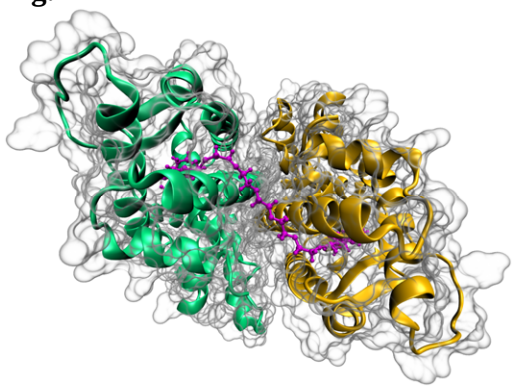

Fig. 8

Supporting Information

Influence of Spacer-Receptor Interactions on the Stability of Bivalent Ligand-Receptor Complexes

Revised Manuscript from 23-jan-2012. This manuscript is the version included in the submitted PhD Thesis. Final version is http://dx.doi.org/10.1021/jp211383s

Jorge Numata[1], Alok Juneja[1,2], Dennis J. Diestler[1,3] and Ernst-Walter Knapp[1]*

[1] Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Fabeckstrasse 36A, D-14195 Berlin, Germany
[2] Department of Biosciences and Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden
[3] University of Nebraska-Lincoln, Lincoln, Nebraska 68583, USA

Table of contents:

Figures S1 – S3

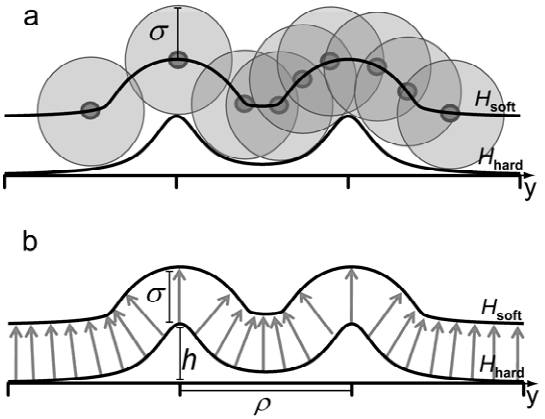

Fig. S1: View along *x*-axis of construction of the surface H_{soft} a) Conceptual procedure: H_{soft} can be regarded as the locus of points generated by the center of a sphere of radius 2*b* rolled over $H_{hard}(x,y)$, where *b* is distance between covalently bound atoms of S. b) Numerical realization: H_{soft} is generated as the locus of tips of vectors proportional to gradient of $[z - H_{hard}(x,y)]$, whose root sweeps over H_{hard} (b). If two gradient tips are assigned to the same point in the *x*-*y* plane, the larger value is used.

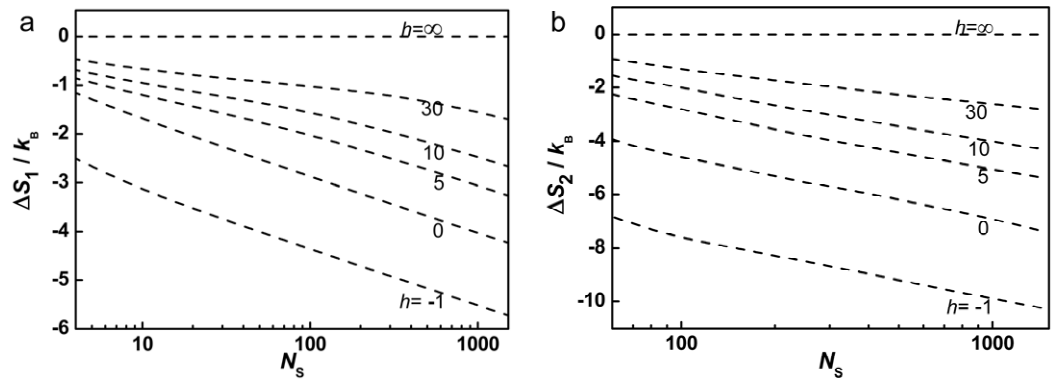

Fig. S2: Same as Fig. 2 in main text, except abscissa (N_S) is on logarithmic scale. Straight lines at large N_S correspond to power law $\eta_j = \exp(\Delta S_j/k_B) \sim (N_S)^{-j/2}$.

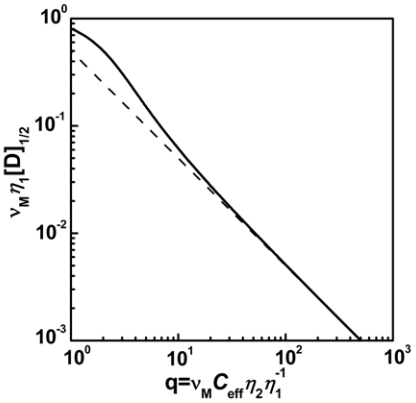

Fig. S3: Plot of $\nu_{\mathrm{M}}\eta_1[\mathrm{D}]_{1/2} = \{2-q+[(2-q)^2+4]^{1/2}\}/4$ vs q (solid line), where $q \equiv \nu_{\mathrm{M}} C_{\mathrm{eff}}(R)\eta_2\eta_1^{-1}$; a version of eq (4) of the main text modified to emphasize key role of parameter q in determining enhancement effect. Asymptotic behavior indicated by dashed line $\nu_{\mathrm{M}}\eta_1[\mathrm{D}]_{1/2} = q^{-1}/2$.

Appendix A: Canonical Partition Functions for Isolated ligand-receptor complexes

Employing a slight modification of our previous notation[1], we write the Hamiltonian of $\mathrm{RD}^{(1)}$ as

$$H_{\mathrm{RD}^{(1)}} = H_{\mathrm{R}} + \sum_{k=1}^{2} P_k^2/2m_{\mathrm{M}} + u(\mathbf{R}_1 - \mathbf{r}_\alpha) + H_{\mathrm{S}} + U^{(\mathrm{SR})}, \tag{A1}$$

In eq (A1) H_{R} refers to the *receptor* R, which is regarded as an extended *rigid* body; k labels the *ligating units*, which are treated as point masses; $u(\mathbf{R}_1 - \mathbf{r}_\alpha)$ is the potential energy of binding of unit 1 to *binding site* α. In case the binding potential can be expressed as a sum of interactions between pairs of atoms, we have explicitly

$$u(\mathbf{R}_1 - \mathbf{r}_\alpha) = \sum_j \phi_{1j}(\mathbf{R}_1 - \mathbf{r}_j), \tag{A2}$$

where the summation on j runs over atoms that make up the "pocket" of site α and $\mathbf{r}_\alpha$ is the minimum of u (i.e., the position occupied by M (or unit 1) at equilibrium at zero temperature). We refer loosely to $\mathbf{r}_\alpha$ as the position of site α. Note that $\mathbf{r}_\alpha$ and $\{\mathbf{r}_j\}$ are implicit functions of the center of mass (c.o.m.) $\mathbf{R}_{\mathrm{R}}$ and orientation $\boldsymbol{\omega}_{\mathrm{R}}$ of R, where

$\boldsymbol{\omega}_{\mathrm{R}}$ collectively represents the three Eulerian angles. In particular, we can express the position of site α explicitly as

$$\mathbf{r}_{\alpha} = \mathbf{R}_{\mathrm{R}} + \boldsymbol{\rho}_{\alpha}(\boldsymbol{\omega}_{R}), \tag{A3}$$

where $\boldsymbol{\rho}_{\alpha}$ is the position of site α with respect to the c.o.m. of R. By analogy the potential energy of binding of unit 1 to the other site (β) can be written

$$\begin{aligned} u(\mathbf{R}_1 - \mathbf{r}_{\beta}) &= u(\mathbf{R}_1 - \mathbf{R}_{\mathrm{R}} - \boldsymbol{\rho}_{\alpha} - \boldsymbol{\rho}_{\beta\alpha}) \\ &= \sum_{l} \phi_{1l}(\mathbf{R}_1 - \mathbf{r}_l), \end{aligned} \tag{A4}$$

where the summation on l is over atoms composing the pocket of site β and $\boldsymbol{\rho}_{\alpha\beta} = \boldsymbol{\rho}_{\beta} - \boldsymbol{\rho}_{\alpha}$ is the distance from site α to site β.

In eq (A1) H_{S} replaces H_{link} of ref 1 as the Hamiltonian of the spacer (S), given by

$$H_{\mathrm{S}} = \sum_{i=1}^{N_{\mathrm{S}}} p_i^2 / 2m_i + U^{(S)}(\mathbf{r}^{N_{\mathrm{S}}}; \mathbf{R}_1, \mathbf{R}_2), \tag{A5}$$

where the summation on index i runs over the N_{S} atoms of S and $U^{(\mathrm{S})}$ represents the *internal* configurational (potential) energy of ligand D. As a concrete example, we instance the bead-spring model[2] of a polymer chain, for which $U^{(\mathrm{S})}$ assumes the form

$$U^{(\mathrm{S})} = \tfrac{1}{2} k_M \left(\left| \mathbf{r}_1 - \mathbf{R}_1 \right| - r_{M1} \right)^2 + \tfrac{1}{2} \sum_{i=1}^{N_{\mathrm{S}}-1} k_0 \left(\left| \mathbf{r}_{i+1} - \mathbf{r}_i \right| - r_0 \right)^2 + \tfrac{1}{2} k_M \left(\left| \mathbf{R}_2 - \mathbf{r}_{N_{\mathrm{S}}} \right| - r_{M1} \right)^2 \tag{A6}$$

The first and last terms of the right member of eq (A6) indicate that the covalent bonds between the end atoms of S (i = 1 and $i = N_{\mathrm{S}}$) and the Ms are accounted for by $U^{(\mathrm{S})}$. If, as is customarily done, we assume that the S-R interaction can be expressed as a sum of interactions between pairs of atoms, then we can write

$$U^{(\mathrm{SR})} = \sum_{i=1}^{N_{\mathrm{S}}} \sum_{j} \phi_{ij} \left(\left| \mathbf{r}_i - \mathbf{r}_j \right| \right), \tag{A7}$$

where the index j runs over atoms of R. Note that the *non-specific* interactions of the ligating units with R are ignored.

In notation employed in ref 1 the partition function of $\mathrm{RD}^{(1)}$ can be expressed

$$q(\mathrm{RD}^{(1)}) = \frac{1}{\Sigma_{\mathrm{RD}^{(1)}}} h^{-12} \int d\mathbf{P}_\mathrm{R} \int d\mathbf{R}_\mathrm{R} \int d\mathbf{J}_\mathrm{R} \int d\boldsymbol{\omega}_\mathrm{R} \int d\mathbf{P}_1 \int_{\Omega_\alpha} d\mathbf{R}_1 \int d\mathbf{P}_2 \int_{\Omega'} d\mathbf{R}_2 \times \exp\{-\beta[H_\mathrm{R} + \sum_{k=1}^{2} P_k^2 / 2m_\mathrm{M} + u(\mathbf{R}_1 - \mathbf{R}_\mathrm{R} - \boldsymbol{\rho}_\alpha) + F^{(1)}]\} \quad \text{(A8)}$$

Here $\Sigma_{\mathrm{RD}^{(1)}}$ is the symmetry number that corrects for multiple counting of indistinguishable orientations; $\beta \equiv (k_B T)^{-1}$, k_B is Boltzmann's constant and T is the absolute temperature;

$$F^{(1)} \equiv -k_B T \ln Q_\mathrm{S}^{(1)} \quad \text{(A9a)}$$

and

$$Q_\mathrm{S}^{(1)} \equiv h^{-3N_\mathrm{S}} \int d\mathbf{p}^{N_\mathrm{S}} \int d\mathbf{r}^{N_\mathrm{S}} e^{-\beta[H_\mathrm{S} + U^{(\mathrm{SR})}]} \quad \text{(A9b)}$$

The quantity $F^{(1)}$ is a generalization of F introduced in ref 1. The superscript (1) emphasizes that the Hamiltonian of S is augmented by the S-R interaction $U^{(\mathrm{SR})}$. $F^{(1)}$, which depends in general on $\mathbf{R}_\mathrm{R}$, $\boldsymbol{\omega}_\mathrm{R}$, $\mathbf{R}_1$, and $\mathbf{R}_2$ (as well as on the parameters N_S and T), is the free energy of S regarded as a thermodynamic subsystem in the field of the ligating units and R in a given (instantaneous) configuration specified by the aforementioned coordinates. It is the thermodynamic state-dependent effective potential energy mediating the interactions between the ligating units and R. In eq (A8) Ω' and Ω_α signify constraints on the regions of integrations on $\mathbf{R}_2$ and $\mathbf{R}_1$ that define the complex. Thus, $\mathbf{R}_1$ must be confined to a small region Ω_α "centered" on $\mathbf{R}_\mathrm{R} + \boldsymbol{\rho}_\alpha$, the position of site α. In contrast, $\mathbf{R}_2$ ranges over the remainder of space Ω' (where the prime denotes the exception of the negligible portions of space corresponding to the pockets).

We now introduce the transformation of variables

$$\begin{aligned}
&\mathbf{R}'_\mathrm{R} = \mathbf{R}_\mathrm{R} && ;\mathbf{P}'_\mathrm{R} = \mathbf{P}_\mathrm{R} \\
&\boldsymbol{\omega}'_\mathrm{R} = \boldsymbol{\omega}_\mathrm{R} && ;\mathbf{J}'_\mathrm{R} = \mathbf{J}_\mathrm{R} \\
&\mathbf{R}'_k = \mathbf{R}_k - \mathbf{R}_\mathrm{R} - \boldsymbol{\rho}_\alpha && ;\mathbf{P}'_k = \mathbf{P}_k, \quad k = 1,2 \\
&\mathbf{r}'_i = \mathbf{r}_i - \mathbf{R}_\mathrm{R} - \boldsymbol{\rho}_\alpha && ;\mathbf{p}'_i = \mathbf{p}_i \ , \quad i = 1,2,\ldots N_\mathrm{S} \\
&\mathbf{r}'_j = \mathbf{r}_j - \mathbf{R}_\mathrm{R} - \boldsymbol{\rho}_\alpha, &&
\end{aligned} \quad \text{(A10)}$$

where the index j labels atoms of R. Noting that the associated Jacobian is unity, we can rewrite eq (A8) as

$$q(\mathrm{RD}^{(1)}) = \frac{1}{\sum_{\mathrm{RD}^{(1)}}} h^{-12} \int d\mathbf{P}'_{\mathrm{R}} \int d\mathbf{R}'_{\mathrm{R}} \int d\mathbf{J}'_{\mathrm{R}} \int d\boldsymbol{\omega}'_{\mathrm{R}} \int d\mathbf{P}'_1 \int_{\Omega_\alpha} d\mathbf{R}'_1 \int d\mathbf{P}'_2 \int_{\Omega'} d\mathbf{R}'_2 \times \exp\{-\beta[H_{\mathrm{R}} + \sum_{k=1}^{2} P_k'^2 / 2m_{\mathrm{M}} + u(\mathbf{R}'_1) + F^{(1)}(\mathbf{R}'_1, \mathbf{R}'_2; N_{\mathrm{S}}, T)]\} \tag{A11}$$

In the new (primed) coordinate system $F^{(1)}$ depends in general on the positions of both ligating units relative to binding site α. But, since the integration over $\mathbf{R}'_1$ is restricted to the relatively small pocket of α, where $\mathbf{R}'_1 \simeq \mathbf{0}$, $F^{(1)}$ depends, to a decent approximation, only on the difference $\mathbf{R}'_2 - \mathbf{R}'_1$. Therefore, integrating on momenta and coordinates $\mathbf{R}'_{\mathrm{R}}$ and $\boldsymbol{\omega}'_{\mathrm{R}}$ and introducing the additional change of variables $\mathbf{R}''_1 = \mathbf{R}'_1; \mathbf{R}'' = \mathbf{R}'_2 - \mathbf{R}'_1$ in eq (A11), we get

$$\begin{aligned} q(\mathrm{RD}^{(1)}) &= \frac{\pi^{1/2}}{\sum_{\mathrm{RD}^{(1)}}} V \Lambda_{\mathrm{R}}^{-3} T_{\mathrm{R}}^{3/2} \Lambda_{\mathrm{M}}^{-6} \int_{\Omega_\alpha} d\mathbf{R}''_1 \int_{\Omega'} d\mathbf{R}'' \exp\{-\beta[u(\mathbf{R}''_1) + F^{(1)}(\mathbf{R}''; N_{\mathrm{S}}, T)]\} \\ &= \frac{\pi^{1/2}}{\sum_{\mathrm{RD}^{(1)}}} \frac{V}{\Lambda_{\mathrm{R}}^3} T_{\mathrm{R}}^{3/2} \frac{v_{\mathrm{M}}(T)}{\Lambda_{\mathrm{M}}^3} \frac{v_{\mathrm{S}}^{(1)}(N_{\mathrm{S}}, T)}{\Lambda_{\mathrm{M}}^3}, \end{aligned} \tag{A12}$$

where

$$v_{\mathrm{M}}(T) \equiv \int_{\Omega_\alpha} d\mathbf{R}''_1 \exp[-\beta u(\mathbf{R}''_1)] \tag{A13a}$$

$$v_{\mathrm{S}}^{(1)}(N_{\mathrm{S}}, T) \equiv \int_{\Omega} d\mathbf{R}'' \exp[-\beta F^{(1)}(\mathbf{R}''; N_{\mathrm{S}}, T)] \tag{A13b}$$

Note that in replacing Ω' by Ω in eq (A13b) we ignore the contribution to the small regions corresponding to the pockets.

Assuming that the spacers do not interact with each other, we can express the Hamiltonian of the isolated complex $\mathrm{RD}^{(1)}{}_2$ as

$$\begin{aligned} H_{\mathrm{RD}_2^{(1)}} &= H_{\mathrm{R}} + \sum_{i=1}^{2} \sum_{k=1}^{2} P_{ik}^2 / 2m_{\mathrm{M}} + u(\mathbf{R}_{11} - \mathbf{R}_{\mathrm{R}} - \boldsymbol{\rho}_\alpha) \\ &+ u(\mathbf{R}_{12} - \mathbf{R}_{\mathrm{R}} - \boldsymbol{\rho}_\alpha - \boldsymbol{\rho}_{\beta\alpha}) + \sum_{k=1}^{2} H_{\mathrm{S},k} + \sum_{k=1}^{2} U^{(\mathrm{SR})}(\mathbf{r}_k^{N_{\mathrm{S}}}; \mathbf{R}_{1k}, \mathbf{R}_{2k}), \end{aligned} \tag{A14}$$

where the indices i (=1, 2) and k (=1, 2) label ligating units and Ss, respectively. Transforming both sets of coordinates according to eq (A10) and invoking the same approximations that were used to reach eq (A12), we obtain

$$
\begin{aligned}
q(\mathrm{RD}_2^{(1)}) = {} & \frac{1}{\sum_{\mathrm{RD}_2^{(1)}}} h^{-18} \int d\mathbf{P}'_{\mathrm{R}} \int d\mathbf{R}'_{\mathrm{R}} \int d\mathbf{J}'_{\mathrm{R}} \int d\boldsymbol{\omega}'_{\mathrm{R}} \int d\mathbf{P}'_{11} \int_{\Omega_\alpha} d\mathbf{R}'_{11} \int d\mathbf{P}'_{21} \int d\mathbf{R}'_{21} \\
& \times \int d\mathbf{P}'_{12} \int_{\Omega_\beta} d\mathbf{R}'_{12} \int d\mathbf{P}'_{22} \int d\mathbf{R}'_{22} \exp\{-\beta[H'_{\mathrm{R}} + \sum_{i=1}^{2} \sum_{k=1}^{2} P'^{2}_{ik} / 2m_{\mathrm{M}}]\} \\
& \times \exp\{-\beta\left[u(\mathbf{R}'_{11}) + u(\mathbf{R}'_{12} - \boldsymbol{\rho}_{\beta\alpha})\right]\} \exp\{-\beta[F^{(1)}(\mathbf{R}'_{21} - \mathbf{R}'_{11}; N_{\mathrm{S}}, T)]\} \\
& \times \exp\{-\beta[F^{(1)}(\mathbf{R}'_{22} - \mathbf{R}'_{12}; N_{\mathrm{S}}, T)]\} = \frac{\pi^{1/2}}{\sum_{\mathrm{RD}_2^{(1)}}} \frac{V}{\Lambda_{\mathrm{R}}^3} T_{\mathrm{R}}^{3/2} \left(\frac{\nu_{\mathrm{M}}}{\Lambda_{\mathrm{M}}^3}\right)^2 \left(\frac{\nu_{\mathrm{S}}^{(1)}}{\Lambda_{\mathrm{M}}^3}\right)^2 .
\end{aligned}
\tag{A15}
$$

For the isolated $\mathrm{RD}^{(2)}$ we have the Hamiltonian

$$
\begin{aligned}
H_{\mathrm{RD}^{(2)}} = {} & H_{\mathrm{R}} + \sum_{k=1}^{2} P_k^2 / 2m_{\mathrm{M}} + u(\mathbf{R}_1 - \mathbf{R}_{\mathrm{R}} - \boldsymbol{\rho}_\alpha) \\
& + u(\mathbf{R}_2 - \mathbf{R}_{\mathrm{R}} - \boldsymbol{\rho}_\alpha - \boldsymbol{\rho}_{\beta\alpha}) + H_{\mathrm{S}} + U^{(\mathrm{SR})}(\mathbf{r}^{N_{\mathrm{S}}}; \mathbf{R}_1, \mathbf{R}_2)
\end{aligned}
\tag{A16}
$$

Again employing the transformation of integration variables eq (A10), we can cast $q(\mathrm{RD}^{(2)})$ as

$$
\begin{aligned}
q(\mathrm{RD}^{(2)}) = {} & \frac{1}{\sum_{\mathrm{RD}^{(2)}}} h^{-12} \int d\mathbf{P}'_{\mathrm{R}} \int d\mathbf{R}'_{\mathrm{R}} \int d\mathbf{J}'_{\mathrm{R}} \int d\boldsymbol{\omega}'_{\mathrm{R}} \int d\mathbf{P}'_1 \int_{\Omega_\alpha} d\mathbf{R}'_1 \int d\mathbf{P}'_2 \int_{\Omega_\beta} d\mathbf{R}'_2 \\
& \times \exp\{-\beta[H'_{\mathrm{R}} + \sum_{k=1}^{2} P'^{2}_k / 2m_{\mathrm{M}} + u(\mathbf{R}'_1) + u(\mathbf{R}'_2 - \boldsymbol{\rho}_{\beta\alpha})]\} \\
& \times \exp\{-\beta F^{(1)}(\mathbf{R}'_2 - \mathbf{R}'_1; N_{\mathrm{S}}, T)\}
\end{aligned}
\tag{A17}
$$

According to the constraints signified by Ω_α and Ω_β, the coordinates $\mathbf{R}'_1$ and $\mathbf{R}'_2$ remain within the pockets of sites α and β, respectively (i.e., $\mathbf{R}'_1 \simeq \mathbf{0}$ and $\mathbf{R}'_2 \simeq \boldsymbol{\rho}_{\beta\alpha}$). Expanding $F^{(1)}$ in Taylor's series about this configuration and neglecting all but the lowest-degree contribution, namely $F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_{\mathrm{S}}, T)$, we deduce from eq (A17)

$$
\begin{aligned}
q(\mathrm{RD}^{(2)}) = {} & \frac{\pi^{1/2}}{\sum_{\mathrm{RD}^{(2)}}} \frac{V}{\Lambda_{\mathrm{R}}^3} T_{\mathrm{R}}^{3/2} \frac{1}{\Lambda_{\mathrm{M}}^6} \int_{\Omega_\alpha} d\mathbf{R}'_1 \exp[-\beta u(\mathbf{R}'_1)] \\
& \times \int_{\Omega_\beta} d\mathbf{R}'_2 \exp[-\beta u(\mathbf{R}'_2 - \boldsymbol{\rho}_{\beta\alpha})] \exp[-\beta F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_{\mathrm{S}}, T)] \\
= {} & \frac{\pi^{1/2}}{\sum_{\mathrm{RD}^{(2)}}} \frac{V}{\Lambda_{\mathrm{R}}^3} T_{\mathrm{R}}^{3/2} \left(\frac{\nu_{\mathrm{M}}}{\Lambda_{\mathrm{M}}^3}\right)^2 \exp(-\beta F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_{\mathrm{S}}, T))
\end{aligned}
\tag{A18}
$$

Appendix B: Alternative Formulas for Binding Constants

In the absence of the S-R interaction ($U^{(\mathrm{SR})} = 0$) eqs (A9) reduce to

$$
F(R; N_{\mathrm{S}}, T) \equiv -k_B T \ln Q_{\mathrm{S}} \tag{B1a}
$$

and

$$Q_{\mathrm{S}}(R;N_{\mathrm{S}},T) \equiv h^{-3N_{\mathrm{S}}} \int d\mathbf{p}^{N_{\mathrm{S}}} \int d\mathbf{r}^{N_{\mathrm{S}}} e^{-\beta H_{\mathrm{S}}} \tag{B1b}$$

Here F, the effective potential energy of interaction between the ligating units, depends by virtue of symmetry only on the distance R between the units.[1] In contrast, $F^{(1)}$, which includes the influence of the receptor ($U^{(\mathrm{SR})} \neq 0$), depends on the (vector) distance $\mathbf{R}$ from unit 1 to unit 2. As pointed out just below eq (A11), $F^{(1)}$ generally depends on both $\mathbf{R}_1$ and $\mathbf{R}_2$. However, in the present application $\mathbf{R}_1$ is *fixed* relative to the receptor, so that $F^{(1)}$ depends only on $\mathbf{R} = \mathbf{R}_2 - \mathbf{R}_1$. Hence, from eqs (A9b) and (B1b) we have

$$\begin{aligned} Q_{\mathrm{S}}^{(1)}(\mathbf{R};N_{\mathrm{S}},T) &= h^{-3N_{\mathrm{S}}} \int d\mathbf{p}^{N_{\mathrm{S}}} \int d\mathbf{r}^{N_{\mathrm{S}}} \exp(-\beta H_{\mathrm{S}}) \frac{h^{-3N_{\mathrm{S}}} \int d\mathbf{p}^{N_{\mathrm{S}}} \int d\mathbf{r}^{N_{\mathrm{S}}} \exp(-\beta[H_{\mathrm{S}} + U^{(\mathrm{SR})}])}{h^{-3N_{\mathrm{S}}} \int d\mathbf{p}^{N_{\mathrm{S}}} \int d\mathbf{r}^{N_{\mathrm{S}}} \exp(-\beta H_{\mathrm{S}})} \\ &= Q_{\mathrm{S}}(\mathbf{R};N_{\mathrm{S}},T) \times \frac{\int d\mathbf{r}^{N_{\mathrm{S}}} \exp(-\beta U^{(S)}(\mathbf{r}^{N_{\mathrm{S}}};\mathbf{R})) \exp(-\beta U^{(\mathrm{SR})})}{\int d\mathbf{r}^{N_{\mathrm{S}}} \exp(-\beta U^{(S)}(\mathbf{r}^{N_{\mathrm{S}}};\mathbf{R}))} \\ &= Q_{\mathrm{S}}(\mathbf{R};N_{\mathrm{S}},T) \left\langle \exp(-\beta U^{(\mathrm{SR})}) \right\rangle_2 (\mathbf{R};N_{\mathrm{S}},T). \end{aligned} \tag{B2}$$

The symbol $\langle X \rangle_2 (\mathbf{R};N_{\mathrm{S}},T)$ signifies the (restricted) canonical ensemble average of the dynamical quantity X over the configuration space ($\mathbf{r}^{N_{\mathrm{S}}}$) of the free spacer, with the end-to-end distance fixed at **R**. Combining eqs (A9a), (B1a) and (B2), we obtain

$$F^{(1)}(\mathbf{R};N_{\mathrm{S}},T) = F(R;N_{\mathrm{S}},T) - k_B T \ln\left[\left\langle \exp(-\beta U^{(\mathrm{SR})}) \right\rangle_2 (\mathbf{R};N_{\mathrm{S}},T) \right] \tag{B3}$$

The effective volume of S in the absence of the S-R interaction is given by

$$v_{\mathrm{S}}(N_{\mathrm{S}},T) \equiv \int d\mathbf{R} \, \exp(-\beta F(R;N_{\mathrm{S}},T)) \tag{B4}$$

Hence, from this relation and that in eq (A13b) we obtain for the ratio of the effective volumes

$$\begin{aligned} \frac{v_{\mathrm{S}}^{(1)}(N_{\mathrm{S}},T)}{v_{\mathrm{S}}(N_{\mathrm{S}},T)} &= \frac{\int d\mathbf{R} \exp[-\beta F^{(1)}(\mathbf{R};N_{\mathrm{S}},T)]}{\int d\mathbf{R} \exp[-\beta F(R;N_{\mathrm{S}},T)]} \\ &= \frac{\int d\mathbf{R} \exp[-\beta F(R;N_{\mathrm{S}},T)] \left\langle \exp(-\beta U^{(\mathrm{SR})}) \right\rangle_2 (\mathbf{R};N_{\mathrm{S}},T)}{\int d\mathbf{R} \exp[-\beta F(R;N_{\mathrm{S}},T)]} \\ &= \int d\mathbf{R} \, C_{\mathrm{eff}}(R;N_{\mathrm{S}},T) \left\langle \exp(-\beta U^{(\mathrm{SR})}) \right\rangle_2 (\mathbf{R};N_{\mathrm{S}},T) \end{aligned} \tag{B5}$$

The last line of eq (B5) invokes the definition of the effective concentration of unit 2 with respect to unit 1 for the free spacer,

$$C_{\text{eff}}(R;N_{\text{S}},T)=\frac{\exp(-\beta F(R;N_{\text{S}},T))}{\int d\mathbf{R}\ \exp(-\beta F(R;N_{\text{S}},T))} \tag{B6}$$

It is equivalent to the probability density $p_{21}(R)$ that unit 2 is located at distance R from unit 1 in the free spacer.

An alternative expression for this ratio can be reached by the following sequence:

$$\begin{aligned}\frac{v_{\text{S}}^{(1)}}{v_{\text{S}}}&=\frac{\int d\mathbf{R}\, e^{-\beta F^{(1)}(\mathbf{R};N_{\text{S}},T)}}{\int d\mathbf{R}\, e^{-\beta F(R;N_{\text{S}},T)}}\\&=\frac{\int d\mathbf{R}\, Q_{\text{S}}^{(1)}}{\int d\mathbf{R}\, Q_{\text{S}}}=\frac{\int d\mathbf{R}\int d\mathbf{p}^{N_{\text{S}}}\int d\mathbf{r}^{N_{\text{S}}}e^{-\beta[H_{\text{S}}+U^{(\text{SR})}]}}{\int d\mathbf{R}\int d\mathbf{p}^{N_{\text{S}}}\int d\mathbf{r}^{N_{\text{S}}}e^{-\beta H_{\text{S}}}}\\&=\frac{\int d\mathbf{R}\int d\mathbf{r}^{N_{\text{S}}}e^{-\beta U^{(S)}(\mathbf{r}^{N_{\text{S}}};\mathbf{R})}e^{-\beta U^{(\text{SR})}}}{\int d\mathbf{R}\int d\mathbf{r}^{N_{\text{S}}}e^{-\beta U^{(S)}(\mathbf{r}^{N_{\text{S}}};\mathbf{R})}}\equiv\left\langle\exp(-\beta U^{(\text{SR})})\right\rangle_1,\end{aligned} \tag{B7}$$

where the second line depends on eqs (B1) and (A9) and the third line on eq (A5). The symbol $\langle X\rangle_1$ stands for the average of X over the ensemble of configurations of the free spacer specified by the $(3N_{\text{S}}+3)$-dimensional vector $(\mathbf{R},\mathbf{r}^{N_{\text{S}}})$. Using eq (B7), we can rewrite $K(\text{RD}^{(1)})$ as shown in the fourth column of Table S1.

From eqs (B3) and (B4) we have

$$\begin{aligned}\frac{\exp(-\beta F^{(1)})}{v_{\text{S}}}&=\frac{\exp(-\beta F(\mathbf{R};N_{\text{S}},T))\left\langle\exp(-\beta U^{(\text{SR})})\right\rangle_2(\mathbf{R};N_{\text{S}},T)}{\int_{\Omega'}d\mathbf{R}\ \exp(-\beta F(R;N_{\text{S}},T))}\\&=C_{\text{eff}}(\mathbf{R};N_{\text{S}},T)\left\langle\exp(-\beta U^{(\text{SR})})\right\rangle_2(\mathbf{R};N_{\text{S}},T)\end{aligned} \tag{B8}$$

Hence, we can recast $K(\text{RD}^{(2)})$ as indicated in the fourth column of Table S1.

To simplify the notation in Table S1 and the main text, we define

$$\eta_1=\eta_1(N_{\text{S}},T)\equiv\left\langle\exp(-\beta U^{(\text{SR})})\right\rangle_1 \tag{B9a}$$

$$\eta_2=\eta_2(\mathbf{R};N_{\text{S}},T)\equiv\left\langle\exp(-\beta U^{(\text{SR})})\right\rangle_2(\mathbf{R};N_{\text{S}},T) \tag{B9b}$$

Appendix C: Generation of Surface H_{soft}

The "soft" surface of R can be conceptualized as the locus of points generated either by the center of a sphere of radius $\sigma = 2b$ rolling over H_{hard} (Fig. S1a) or by the tip of the vector, proportional to the gradient of $[z–H_{hard}(x, y)]$, whose root roams over H_{hard} (Fig. S1b). In practice we implement the latter view numerically as follows: We first construct a low-resolution (coarse: 0.1Å) and a high-resolution (fine: 0.02Å) grid in the basal *xy*-plane. We consider a point (x_0, y_0) on the fine grid, which corresponds to the point $\mathbf{r}_0 = (x_0, y_0, H_{hard}(x_0, y_0))$ on H_{hard}. We wish to find the point on H_{soft} that lies a distance $2b$ from $\mathbf{r}_0$ in the direction of the normal to H_{hard} at $\mathbf{r}_0$. The locus of points that define H_{hard} is described by

$$\mathbf{r} = x\mathbf{e}_x + y\mathbf{e}_y + H_{hard}(x, y)\mathbf{e}_z . \tag{C1}$$

Hence, the normal to H_{hard} at $\mathbf{r} = \mathbf{r}_0$ can be expressed as

$$\begin{aligned}
\mathbf{n}(\mathbf{r}_0) &= \left[\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} \right]_{\mathbf{r}=\mathbf{r}_0} \\
&= \left[(\mathbf{e}_x + \frac{\partial H_{hard}}{\partial x}\mathbf{e}_z) \times (\mathbf{e}_y + \frac{\partial H_{hard}}{\partial y}\mathbf{e}_z) \right]_{\mathbf{r}=\mathbf{r}_0} \\
&= \left(\mathbf{e}_z - \frac{\partial H_{hard}}{\partial x}\mathbf{e}_x - \frac{\partial H_{hard}}{\partial y}\mathbf{e}_y \right)_{\mathbf{r}=\mathbf{r}_0} \\
&= \{\nabla_{\mathbf{r}}[z - H_{hard}(x, y)]\}_{\mathbf{r}=\mathbf{r}_0} .
\end{aligned} \tag{C2}$$

The point on H_{soft} corresponding to $(x_0, y_0, H_{hard}(x_0, y_0))$ is then given by

$$(x_1, y_1, H_{soft}(x_1, y_1)) = (x_0, y_0, H_{hard}(x_0, y_0)) + 2b \frac{\{\nabla_{\mathbf{r}}[z - H_{hard}(x, y)]\}_{\mathbf{r}=\mathbf{r}}}{\left|\{\nabla_{\mathbf{r}}[z - H_{hard}(x, y)]\}_{\mathbf{r}=\mathbf{r}}\right|} . \tag{C3}$$

We assign $z_1 = H_{soft}(x_1, y_1)$ to the point (x_1', y_1') on the coarse grid that is nearest (x_1, y_1). Often several values of $z_1 = H_{soft}(x_1, y_1)$ resulting from points on the fine grid can be assigned to the same point (x_1', y_1') of the coarse grid, but only the largest value is retained. In this way a discrete numerical representation of H_{soft} is generated prior to the

computation of ensemble averages. During the simulation $H_{\text{soft}}(x, y)$ corresponding to the position (x, y) of a bead of S is set equal to H_{soft} on the nearest point of the coarse grid.

Appendix D: Generation of Spacer Ensembles, Evaluation of C_{eff} and R_F

In order to generate the S–chain conformations and analyze them, a program in C++ with around 5000 lines of original code plus numerical libraries was written to achieve the necessary efficiency in the simulations.

The conformations of the S–chain are specified in the "laboratory" reference frame by the set of position vectors of the beads fa, where *j* labels the N_{ens} conformations that make up the ensemble and *i* labels the atoms of the S-chain.

The following continuous random walk (CRW) algorithm generates a single conformation of a chain of N_S beads:

1. Set the center of bead 1 at the origin **0**.
2. Set $n = 1$.
3. Set $n = n + 1$.
4. Generate a vector $\mathbf{b}_n$, whose tip is uniformly randomly distributed (URD) on a sphere of radius *b*, where *b* is the bond length. This is accomplished through an algorithm due to Marsaglia[3], which is an optimized version of von Neumann's algorithm[4]. The independent, identically distributed pseudorandom numbers required by Marsaglia's algorithm are generated by the algorithm Ran088 (or Taus088) due to L'Ecuyer[5].
5. Place the center of bead n at the position $\mathbf{b}_n$ with respect to the center of bead $n-1$.
6. If $n < N_S - 1$, go to step 3; if $n \geq N_S$, go to step 7.
7. Stop.

We generate between 5×10^6 and 1×10^{10} free S-chain conformations.

We demonstrated previously[1] that the effective concentration, C_{eff}, defined in eq (B6) of Appendix B (above) is equivalent to the probability distribution of the end-to-end distance $R_{ee}^{(j)} = |\mathbf{r}_{N_S}^{(j)} - \mathbf{r}_1^{(j)}|$ of the S-chain. The latter can be expressed discretely in terms of the N_{ens} numerically generated S-chain conformations as

$$C_{\text{eff}}(R_l) \simeq \frac{n(R_l)}{V(R_l)\,N_{\text{ens}}}, \tag{D1}$$

where $n(R_l)$ is number of conformations having end-to-end distances $R_{ee}^{(j)}$ that satisfy the constraint $|R_{ee}^{(j)} - R_l| < \Delta$ (regardless of the chain orientation) and the volume of the corresponding spherical shell is $V(R_l) = 4\pi[(R_l + \Delta)^3 - (R_l - \Delta)^3]/3$. Note that the discrete distribution, eq (D1), is normalized as

$$\sum_l V(R_l)\,C_{\text{eff}}(R_l) = N_{\text{ens}}^{-1} \sum_l n(R_l) \simeq \int d\mathbf{R}_{ee} C_{\text{eff}}(R_{ee}) = 1, \tag{D2}$$

where the summation on l runs over the discrete intervals of *fixed* length 2 Δ. The Flory radius R_F (*i.e.*, the root mean square end-to-end distance) of the free S-chain is defined as

$$R_F^2 \equiv < R_{ee}^2 >_{free} = \int d\mathbf{R}_{ee}\, R_{ee}^2 C_{\text{eff}}(R_{ee}) \simeq N_{\text{ens}}^{-1} \sum_{j=1}^{N_{\text{ens}}} [R_{ee}^{(j)}]^2. \tag{D3}$$

The parameter Δ, depending on the length of the chain, varies between 0.2 Å for the shortest ($N_S \leq 121$) and 2.0 Å for the longest S-chains ($N_S = 1500$) considered.

Appendix E: Ensemble Averages η_1 (N_S, T) and η_2 (ρ;N_S,T)

η_1 (N_S, T) = <exp(–$U^{(SR)}$/k_BT)>$_1$: One Ligating Unit of D Bound to R . The symbol <…>$_1$ signifies an average over the configurations of the S-chain of N_S beads with bead 1 confined to site α at $\mathbf{r}_\alpha$, as described in the Section 3 of the main text. By translating the beads (i) of each member (j) of the ensemble of free chains by $-\mathbf{r}_1^{(j)} + \mathbf{r}_\alpha$, such that bead 1 is at $\mathbf{r}_\alpha = h\,\mathbf{e}_z$ (binding site α), we can use all N_{ens} conformations of the free chain to compute η_1. Note that we can as well place the last bead (N_S) at $\mathbf{r}_\alpha$ for each of the N_{ens} conformations, thus doubling the size of the ensemble. Since the orientation of R (as specified by the two-dimensional unit normal to the basal plane of R) is not unique, the size of the ensemble used to evaluate η_1 can be effectively enhanced by averaging over random orientations of R. Consequently, the *orientation* of R (initially specified by the unit vector $\mathbf{n} = \mathbf{e}_z$ normal to the basal plane ($z = 0$) can be arbitrary. For each orientation of R we can use the complete ensemble of free S-chain conformations. The two extra degrees of freedom associated with **n** give rise to an additional substantial increase in the effective size of the ensemble and, consequently, in the accuracy of ensemble averages.

We generate $N_{1\mathrm{or}}^{\mathrm{R}} = 2500$ vectors **n** whose tips are uniformly randomly distributed (URD) on the unit sphere (see Appendix C) and use each of them as a new *z*-axis (i.e., $\mathbf{e}_z = \mathbf{n}$) for the reoriented R. Hence, the total number of configurations in the extended ensemble is $N_{\mathrm{config}} = 2 \times N_{1\mathrm{or}}^{\mathrm{R}} \times N_{\mathrm{ens}}$. Table S3 lists the actual numbers used, which were chosen to minimize the statistical error in the plots in Figs. 2a and 4a. For the $\mathrm{R}_{\mathrm{planar}}$ model, where ψ depends only on the distance of the atom from the basal plane, we need only $\mathbf{e}_z = \mathbf{n}$. However, in the case of the $\mathrm{R}_{\mathrm{hard}}$ and $\mathrm{R}_{\mathrm{soft}}$ models, where ψ depends on the vector position $\mathbf{r}_i$ of the atom, new *x*- and *y*-axes must also be determined. We take the new unit vector along the *y*-axis to be $\mathbf{e}_y = \mathbf{e}_z \times \mathbf{n}' / |\mathbf{e}_z \times \mathbf{n}'|$, where $\mathbf{n}'$ is a second URD unit vector. The unit vector in the new *x*-direction is $\mathbf{e}_x = \mathbf{e}_y \times \mathbf{e}_z$.

Thus, we can express the ensemble average as

$$\eta_1 = \tfrac{1}{2} \sum_{l=1}^{2} \left(N_{1\mathrm{or}}^{\mathrm{R}} \right)^{-1} \sum_{k=1}^{N_{1\mathrm{or}}^{\mathrm{R}}} N_{\mathrm{ens}}^{-1} \sum_{j=1}^{N_{\mathrm{ens}}} \exp[-\sum_{i=1}^{N_{\mathrm{S}}} \psi(\mathbf{r}_i^{(j)}(l, \mathbf{n}_k)) / k_B T] , \qquad \text{(E1)}$$

where $\mathbf{r}_i^{(j)}(l,\mathbf{n}_k)$ stands for the position of the *i*th bead of the *j*th conformation of the free chain for the *k*th orientation of R; the index *l* refers to the two possible placements of the end beads at $\mathbf{r}_\alpha = h\, \mathbf{e}_z$.

In the case of the $\mathrm{R}_{\mathrm{soft}}$ model, as well as the $\mathrm{R}_{\mathrm{planar}}$ model with the attractive square well, the formal expression in eq (E1) for the ensemble average is susceptible to appreciable round-off error, when evaluated straightforwardly. The essential problem is that a configuration with a large number of beads in the attractive region could occur early in the accumulation of the sum. If so, its Boltzmann factor would be far larger than those of the following configurations, which would therefore be lost because of the limited number of significant digits available for floating point computations. This problem is avoided by recasting eq (E1) as

$$\eta_1 = \sum_{n_{\mathrm{b}}=0}^{N_{\mathrm{S}}} g(n_{\mathrm{b}}) \exp\left(\frac{n_{\mathrm{b}} W}{k_B T} \right) \Bigg/ \sum_{n_{\mathrm{b}}=0}^{N_{\mathrm{S}}} g(n_{\mathrm{b}}) \ , \qquad \text{(E2)}$$

where $g(n_{\mathrm{b}})$ is the number of configurations for which all beads are above H_{hard}, or above the basal plane, and n_{b} beads are in the square well.

$\eta_2(\rho;N_S,T) = \langle \exp(-U^{(SR)}/k_BT) \rangle_2$: Both Ligating Units of D Bound to R. The symbol $\langle \ldots \rangle_2$ signifies the average over configurations of the free chain under the stipulation that bead 1 and bead N_S are locked in binding sites α and β respectively, at distance ρ, such that $R_{ee}^{(j)}$ must be close to ρ. To evaluate η_2, we use a subset of the free S-chain ensemble, namely the $n(\rho)$ members for which the condition holds. In the discrete representation of C_{eff} the end-to-end distance $R_{ee}^{(j)}$ of these S-chain conformations lies in the spherical shell of radius ρ. The complex $RD^{(2)}$ is formed as follows: (i) the positions $\mathbf{r}_i^{(j)}$ of the beads (i) of the $n(\rho)$ S-chain conformations (j) are translated by $-\mathbf{r}_1^{(j)} + \mathbf{r}_\alpha$; (ii) the basal plane of R is oriented so that it passes through the origin **0** and contains the end-to-end distance vector $\mathbf{R}_{ee}^{(j)}$ such that the unit normal vector **n** is orthogonal to $\mathbf{R}_{ee}^{(j)}$. Note that the orientation of the basal plane of R is not unique in that the tip of **n** can lie at any point on the unit circle contained in a plane perpendicular to $\mathbf{R}_{ee}^{(j)}$. This ambiguity in the orientation of **n** introduces an additional degree of freedom that can be used to enlarge the ensemble and consequently improve the ensemble average. We generate $N_{2\text{or}}^{\text{R}} = 360$ vectors **n** distributed uniformly on the unit circle in the plane normal to $\mathbf{R}_{ee}^{(j)}$, as described in Appendix F. For given **n** we take $\mathbf{e}_z = \mathbf{n}$, $\mathbf{e}_y = \mathbf{R}_{ee}^{(j)} / |\mathbf{R}_{ee}^{(j)}|$ and $\mathbf{e}_x = \mathbf{e}_y \times \mathbf{e}_z$. The ensemble of S-chains with end-to-end distances of about ρ contains $n(\rho)$ members and is a subset of the "free-chain" ensemble of size N_{ens}. Hence, the number of configurations used in the ensemble average for η_2 is $N_{\text{config}} = N_{2\text{or}}^{\text{R}} \times n(\rho)$. Table S4 gives actual numbers used, which were chosen to minimize the statistical error in the plots in Figs. 2b and 4b.

By analogy with eq (E1) we have

$$\eta_2 = \left(N_{2\,\text{or}}^{\text{R}}\right)^{-1} \sum_{k=1}^{N_{2\,\text{or}}^{\text{R}}} \frac{1}{n(\rho)} \sum_{j=1}^{n(\rho)} \exp[-\sum_{i=1}^{N_S} \psi(z_i^{(j)}(\mathbf{n}_k)) / k_B T)] \qquad \text{(E3)}$$

A procedure analogous to the one employed in computing η_1 for the R_{soft} model, and the "attractive" R_{planar} model, is also invoked here to circumvent the problem of round-off error.

Appendix F: Generation of Unit Normal Vectors Perpendicular to End-to-end Vector of Spacer

Let $\mathbf{R}_{ee} = \mathbf{r}_N - \mathbf{r}_1$ be the (vector) distance from end atom (bead) 1 to end atom (bead) N_S and $\mathbf{n}_0$ be a *reference* unit vector normal to $\mathbf{R}_{ee}$. Then $\mathbf{n}_0$ and $\mathbf{R}_{ee}$ must satisfy the requirements

$$\mathbf{n}_0 \cdot \mathbf{n}_0 = n_{0x}^2 + n_{0y}^2 + n_{0z}^2 = 1 \tag{F1}$$

$$\mathbf{n}_0 \cdot \mathbf{R}_{ee} = n_{0x}X + n_{0y}Y + n_{0z}Z = 0 \tag{F2}$$

These constraints yield two equations in three unknowns and are insufficient to determine uniquely the reference vector. Thus, we introduce the additional, arbitrary condition:

$$n_{0z} = 0 \tag{F3}$$

That is, we require the reference normal to be parallel to the *x*-*y* plane. Plugging eq (F3) into eqs (F1) and (F2), we get

$$n_{0x}^2 + n_{0y}^2 = 1 \tag{F4a}$$

$$n_{0x}X + n_{0y}Y = 0\,, \tag{F4b}$$

the general solution of which is

$$n_{0x} = \pm(1+\alpha^2)^{-1/2} \tag{F5a}$$

$$n_{0y} = \mp\alpha(1+\alpha^2)^{-1/2}, \tag{F5b}$$

where

$$\alpha \equiv X / Y\,. \tag{F6}$$

We now take the particular solution

$$\begin{aligned} n_{0x} &= (1+\alpha^2)^{-1/2} \\ n_{0y} &= -\alpha(1+\alpha^2)^{-1/2} \\ n_{0z} &= 0 \end{aligned} \tag{F7}$$

We wish to construct another unit vector $\mathbf{n}$ normal to $\mathbf{R}_{ee}$ that is rotated in the plane normal to $\mathbf{R}_{ee}$ that contains $\mathbf{n}_0$ so that the angle between $\mathbf{n}$ and $\mathbf{n}_0$ is ψ. Then the constraints on $\mathbf{n}$ are

$$\mathbf{n}\cdot\mathbf{n} \;=\; n_x^2 \;+\; n_y^2 \;+\; n_z^2 = 1 \tag{F8a}$$

$$\mathbf{n}_0 \cdot \mathbf{n} = n_{0x}n_x + n_{0y}n_y + n_{0z}n_z = \cos\psi \tag{F8b}$$

$$\mathbf{R}_{ee} \cdot \mathbf{n} = Xn_x + Yn_y + Zn_z = 0 \tag{F8c}$$

Rearranging eqs (F8b) and (F8c), we have

$$\begin{pmatrix} n_{0x} & n_{0y} \\ X & Y \end{pmatrix}\begin{pmatrix} n_x \\ n_y \end{pmatrix} = \begin{pmatrix} \cos\psi \\ -Zn_z \end{pmatrix} \tag{F9}$$

The solution of eq (F9) is

$$n_x = \frac{1}{\gamma}\cos\psi - \frac{\alpha\beta}{\gamma}n_z \tag{F10a}$$

$$n_y = -\frac{\alpha}{\gamma}\cos\psi - \frac{\beta}{\gamma}n_z \tag{F10b}$$

where

$$\gamma = (1+\alpha^2)^{1/2}$$
$$\beta = \frac{Z}{\gamma Y} \tag{F11}$$

By combining eqs (F8a) and (F10) we obtain the following quadratic equation for n_z:

$$\left[1+\beta^2\right]n_z^2 + \left[\left(\frac{\beta\cos\psi}{\gamma}\right)\left(n_{0y}+\alpha n_{0x}\right)\right]n_z - \left[1-\cos^2\psi\right] = 0 \tag{F12}$$

But from eq (F7) we deduce that the quantity $\left(n_{0y}+\alpha n_{0x}\right)$ vanishes and hence that the linear term in eq (F12) also vanishes, yielding the solution

$$n_z = \pm\left(\frac{1-\cos^2\psi}{1+\beta^2}\right)^{1/2} \tag{F13}$$

Collecting the results, we have finally the following formulas for the components of the sought vector **n:**

$$
\begin{aligned}
n_z &= \pm\left(\frac{1-\cos^2\psi}{1+\beta^2}\right)^{1/2} \\
n_x &= \frac{1}{\gamma}\cos\psi - \frac{\alpha\beta}{\gamma}n_z \\
n_y &= -\frac{\alpha}{\gamma}\cos\psi - \frac{\beta}{\gamma}n_z
\end{aligned}
\tag{F14}
$$

Appendix G: Relation of η_j to Free Energy of Binding

We consider the reaction

$$\mathrm{R} + \mathrm{D} \rightleftarrows \mathrm{RD}^{(1)} \tag{G1}$$

The standard Gibbs free energy of binding for $\mathrm{RD}^{(1)}$ is

$$
\begin{aligned}
\Delta F^0_{\mathrm{RD}^{(1)}} &= -k_\mathrm{B}T\ln[K(\mathrm{RD}^{(1)}] \\
&= -k_\mathrm{B}T\ln[4v_\mathrm{M}\eta_1] \\
&= -k_\mathrm{B}T\ln 4 - k_\mathrm{B}T\ln v_\mathrm{M} - k_\mathrm{B}T\ln\eta_1 \\
&\equiv -k_\mathrm{B}T\ln 4 - k_\mathrm{B}T\ln v_\mathrm{M} + \Delta F_1
\end{aligned}
\tag{G2}
$$

(see Table S1). The first term is due to the loss of orientational symmetries of R and D, the second to the confinement of the ligating unit (M) to one site of R (in the absence of the S-R interaction), and the third (ΔF_1) to the S-R interaction itself. It is the last contribution that is of primary interest.

The ensemble average can be cast as

$$\eta_1 = X/Y\,, \tag{G3}$$

where

$$X \equiv \int d\mathbf{r}^N \exp[-\beta U^{(\mathrm{S})}(\mathbf{r}^N)/(k_\mathrm{B}T)]\exp[-U^{(\mathrm{SR})}/(k_\mathrm{B}T)] \tag{G4a}$$

$$Y \equiv \int d\mathbf{r}^N \exp(-U^{(\mathrm{S})}(\mathbf{r}^N)/(k_\mathrm{B}T)) \tag{G4b}$$

Thus, from eqs (G2) and (G3) we get

$$\Delta F_1 = -k_\mathrm{B}T\ln(X/Y) \tag{G5}$$

The corresponding difference in standard entropy of binding is

$$\begin{aligned}\Delta S_1 &= -\frac{\partial \Delta F_1}{\partial T} \\ &= k_{\mathrm{B}} \ln(XY^{-1}) + \frac{k_{\mathrm{B}}T}{XY^{-1}} \left\{ \frac{\partial X}{\partial \beta} Y^{-1} - XY^{-2} \frac{\partial Y}{\partial \beta} \right\} \frac{\partial \beta}{\partial T} \\ &= k_{\mathrm{B}} \ln(XY^{-1}) + k_B T \left\{ \frac{1}{X} \frac{\partial X}{\partial \beta} - \frac{1}{Y} \frac{\partial Y}{\partial \beta} \right\} \frac{\partial \beta}{\partial T},\end{aligned} \tag{G6}$$

where $\beta = 1/k_{\mathrm{B}}\mathrm{T}$. From eq (G4) we deduce the required partial derivatives:

$$\frac{\partial X}{\partial \beta} = -\int d\mathbf{r}^N \exp[-\beta U^{(\mathrm{S})}(\mathbf{r}^N)] \exp(-\beta U^{(\mathrm{SR})})[U^{(\mathrm{S})} + U^{(\mathrm{SR})}] \tag{G7a}$$

$$\frac{\partial Y}{\partial \beta} = -\int d\mathbf{r}^N \exp[-\beta U^{(\mathrm{S})}(\mathbf{r}^N)]\, U^{(\mathrm{S})} \tag{G7b}$$

Combining eqs (G6) and (G7) and invoking the relation $\partial\beta / \partial T = -1/k_{\mathrm{B}}T^2$, we obtain

$$\begin{aligned}\Delta S_1 = k_{\mathrm{B}} \ln &\left[\left\langle \exp(-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)) \right\rangle_1 \right] \\ &+ \frac{1}{T} \left\{ \frac{\int d\mathbf{r}^N \exp[-U^{(\mathrm{S})}(\mathbf{r}^N)/(k_{\mathrm{B}}T)] \exp(-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T))[U^{(\mathrm{S})} + U^{(\mathrm{SR})}]}{\int d\mathbf{r}^N \exp[-U^{(\mathrm{S})}(\mathbf{r}^N)/(k_{\mathrm{B}}T)] \exp(-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T))} \right\} \\ &- \frac{1}{T} \left\{ \frac{\int d\mathbf{r}^N \exp[-U^{(\mathrm{S})}(\mathbf{r}^N)/(k_{\mathrm{B}}T)]\, U^{(\mathrm{S})}}{\int d\mathbf{r}^N \exp[-U^{(\mathrm{S})}(\mathbf{r}^N)/(k_{\mathrm{B}}T)]} \right\}\end{aligned} \tag{G8}$$

Equation (G8) can be rewritten

$$\begin{aligned}\Delta S_1 = k_{\mathrm{B}} \ln &\left[\left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)] \right\rangle_1 \right] \\ &+ \frac{1}{T} \left\{ \frac{\left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)][U^{(\mathrm{S})} + U^{(\mathrm{SR})}] \right\rangle_1}{\left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)] \right\rangle_1} \right\} - \frac{\left\langle U^{(\mathrm{S})} \right\rangle_1}{T}\end{aligned} \tag{G9}$$

In the special case of the freely-jointed chain, the internal configurational energy $U^{(\mathrm{S})}$ is independent of the conformation. Hence, eq (G9) reduces to

$$\begin{aligned}\Delta S_1 = k_{\mathrm{B}} \ln &\left[\left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)] \right\rangle_1 \right] \\ &+ \frac{1}{T} \left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)]\, U^{(\mathrm{SR})} \right\rangle_1 \Big/ \left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)] \right\rangle_1\end{aligned} \tag{G10}$$

The difference in the standard enthalpy of binding is given by

$$\Delta H_1 = \Delta F_1 + T\Delta S_1 = \left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)]\, U^{(\mathrm{SR})} \right\rangle_1 \Big/ \left\langle \exp[-U^{(\mathrm{SR})}/(k_{\mathrm{B}}T)] \right\rangle_1 \tag{G11}$$

One can carry out a similar analysis for η_2, starting from

$$\begin{aligned}\Delta F^{0}_{\mathrm{RD}^{(2)}} &= -k_{\mathrm{B}}T\ln\{K(\mathrm{RD}^{(2)})\} \\ &= -k_{\mathrm{B}}T\ln 2 - k_{\mathrm{B}}T\ln(v_M) - k_{\mathrm{B}}T\ln(v_M C_{\mathrm{eff}}) - k_{\mathrm{B}}T\ln(\eta_2)\end{aligned} \tag{G12}$$

(see Table S1). The first term on the right side of eq (G12) is due to the orientational symmetries of the species involved, the second term to the binding of the M moiety (i.e., the binding of the first ligating unit of D in the absence of the S-R interaction), the third term to the binding of the second unit of D in the absence of the S-R interaction, and the fourth term to the S-R interaction. This last contribution, which is of principal interest in the main text, is defined by

$$\Delta F_2 \equiv -k_B T\ln\eta_2 \tag{G13}$$

In the special case of the $\mathrm{R}_{\mathrm{soft}}$ and "attractive" $\mathrm{R}_{\mathrm{planar}}$ models, we can use the analogue of eq (E2) of Appendix E to express the difference in the standard enthalpy of binding as

$$\Delta H_{\mathrm{j}} = \frac{\sum_{n_{\mathrm{b}}=0}^{N_{\mathrm{S}}}(-n_{\mathrm{b}}W)g(n_{\mathrm{b}})\exp(n_{\mathrm{b}}W/k_{\mathrm{B}}T)}{\sum_{n_{\mathrm{b}}=0}^{N_{\mathrm{S}}}g(n_{\mathrm{b}})\exp(n_{\mathrm{b}}W/k_{\mathrm{B}}T)} \tag{G14}$$

The corresponding difference in the entropy of binding is

$$\Delta S_{\mathrm{j}} = (\Delta H_{\mathrm{j}} - \Delta F_{\mathrm{j}})/T \tag{G15}$$

For the $\mathrm{R}_{\mathrm{soft}}$ model with well depth $W = 0.045$ kcal/mol, we observe that $-\Delta F_{\mathrm{j}}/(k_{\mathrm{B}}T)$ is very small and remains nearly constant as a function of the S-chain length (Fig. S4), where j =1 or 2 refer respectively to the binding of one or both ligating units to R. The small value of $-\Delta F_{\mathrm{j}} = T\Delta S_{\mathrm{j}} - \Delta H_{\mathrm{j}}$ is a consequence of the near cancellation of the entropic contribution $T\Delta S_{\mathrm{j}}$ by the enthalpic contribution ΔH_{j}. As the well gets deeper ($W > 0.045$ kcal/mol), the enthalpic contribution grows faster than the entropic contribution decreases, such that $-\Delta F_{\mathrm{j}}$ becomes positive with increasing S-chain length (N_{S}). On the other hand, as the well becomes shallower ($W < 0.045$ kcal/mol), the entropic term shrinks faster than enthalpic term grows, so that $-\Delta F_{\mathrm{j}}$ becomes negative, as indicated in Fig. S4.

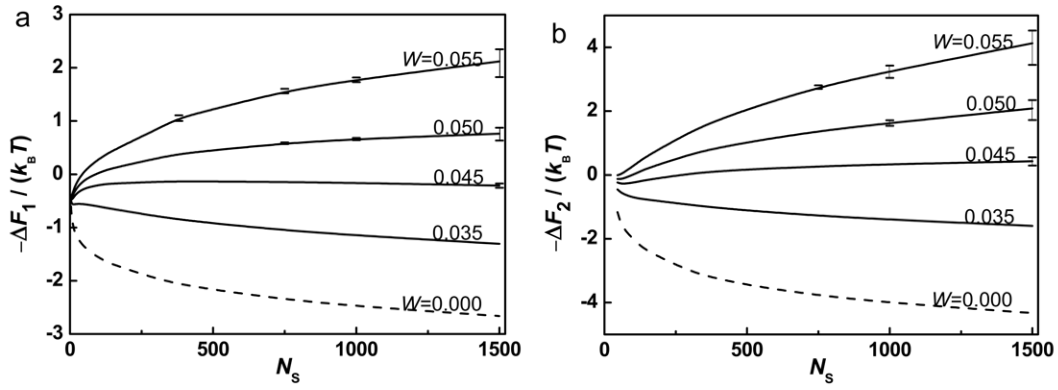

Fig. S4: a: Change in (negative) free energy due to S-R interaction on binding of one ligating unit of D, as a function of S-chain length N_S (number of atoms (beads) in S-chain) for R_{soft} model. Distance between binding sites is $\rho = 30$ Å; height of Lorentzian hills $h = 10$ Å. Numbers labelling curves refer to square well depth W [kcal/mol]. Error bars shown only if deviations larger than line widths. **b:** Change in (negative) free energy due to S-R interaction on binding of both ligating units, as function of S-chain length for R_{soft} model. Same notation and parameters as in Fig. S4a.

It is noteworthy that the entropy loss inherent in $-\Delta F_j$ grows very rapidly with increasing well depth W . Indeed, the entropy loss is so large that for convenience we plot the entropy change (ordinate) on a logarithmic scale in Fig. S5. A logarithmic scale with respect to the entropy is double-logarithmic with respect to the number of states. With increasing S-R attraction (W), the S-chain conformations preferentially accommodate to the 2D surface of R_{soft} instead of extending out into the 3D space, thereby severely decreasing the number of states (expressed as S–chain entropy).

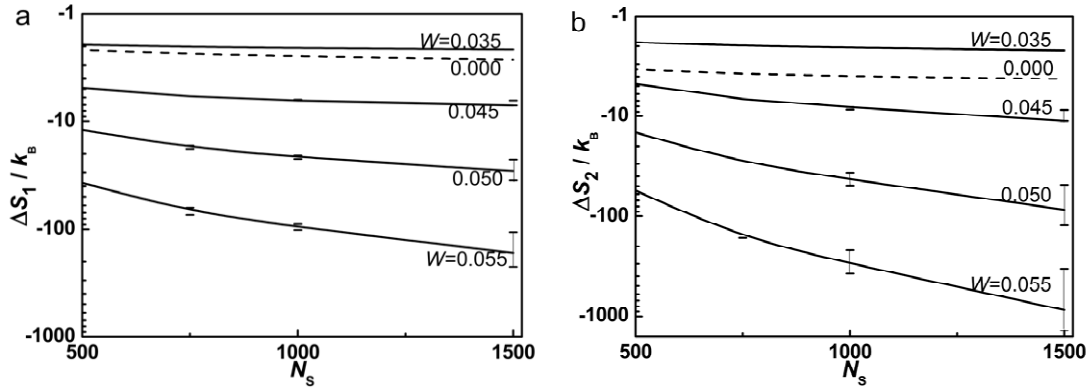

Fig. S5: Entropy loss due to attractive S-R interaction, as function of S-chain length (N_S) for the R_{soft} model with same parameters as in Fig. 5 of the main text. Error bars are only shown if larger than the line widths.

Table S1. Canonical molecular partition functions $q(\mathrm{X})$ and binding constants $K(\mathrm{X})$ for prototypal bivalent system. V is volume of solution; $T_R \equiv T/[\prod_{n=1}^{3}\Theta_n^R]^{1/3}$ is dimensionless absolute temperature, where $\Theta_n^R = h^2/8\pi^2 I_n^R k_B$ are characteristic rotational temperatures, given in terms of principal moments of inertia I_n^R, T is absolute temperature, k_B is Boltzmann's constant, h is Planck's constant; $\Lambda_X \equiv (h^2/2\pi m_X k_B T)^{1/2}$, where m_X is mass of species X. $F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_S, T)$, v_M, $v_S^{(1)}$, v_S, $\eta_2(\rho; N_S, T) = \langle \exp(-U^{(SR)}/k_B T)\rangle_2$, $\eta_1 = \langle \exp(-U^{(SR)}/k_B T)\rangle_1$ (subscripts 1 and 2 denote respectively that one ligating unit of D and both units of D bind to R), and $C_{\mathrm{eff}}(\rho_{\beta\alpha}; N_S, T)$ are defined respectively by eqs (A9a), (A13a), (A13b), (B4), (B2, B9b), (B7, B9a) and (B6). $\boldsymbol{\rho}_{\beta\alpha}$ is distance between binding sites of R. Symmetry numbers for R and D are $\sigma_R = 2$, $\sigma_D = 2$; since M is point mass, it has no symmetry number; symmetry numbers of complexes are $\Sigma_{RM} = 1$, $\Sigma_{RD^{(1)}} = 1$, $\Sigma_{RM_2} = 2$, $\Sigma_{RD_2^{(1)}} = 2$, $\Sigma_{RD^{(2)}} = 2$.

species X	$q(\mathrm{X})/V$	$K(\mathrm{X})$	$K(\mathrm{X})$
R	$\sqrt{\pi}\, T_R^{3/2}\Lambda_R^{-3}/\sigma_R$	----	----
M	Λ_M^{-3}	----	----
D	$v_S\Lambda_M^{-6}/\sigma_D$	----	----
RM	$\sqrt{\pi}\, T_R^{3/2} v_M \Lambda_R^{-3}\Lambda_M^{-3}/\Sigma_{RM}$	$\sigma_R v_M/\Sigma_{RM}$	$2v_M$
$RD^{(1)}$	$\sqrt{\pi} T_R^{3/2} v_M v_S^{(1)} \Lambda_R^{-3}\Lambda_M^{-6}/\Sigma_{RD^{(1)}}$	$\sigma_R\sigma_D v_M v_S^{(1)}/\Sigma_{RD^{(1)}} v_S$	$4v_M\eta_1$
RM_2	$\sqrt{\pi} T_R^{3/2} v_M^2 \Lambda_R^{-3}\Lambda_M^{-6}/\Sigma_{RM_2}$	$\sigma_R v_M^2/\Sigma_{RM_2}$	v_M^2
$RD^{(1)}{}_2$	$\sqrt{\pi}\, T_R^{3/2}(v_M v_S^{(1)})^2 \Lambda_R^{-3}\Lambda_M^{-12}/\Sigma_{RD^{(1)}}$	$\sigma_R\sigma_D^2(v_M v_S^{(1)})^2/\Sigma_{RD_2^{(1)}} v_S^2$	$4v_M^2\eta_1^2$
$RD^{(2)}$	$\sqrt{\pi} T_R^{3/2} v_M^2 \Lambda_R^{-3}\Lambda_M^{-6}/\Sigma_{RD^{(2)}} \times e^{-\beta F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_S, T)}$	$\sigma_R\sigma_D v_M^2 v_S^{-1} \times e^{-\beta F^{(1)}(\boldsymbol{\rho}_{\beta\alpha}; N_S, T)}/\Sigma_{RD^{(2)}}$	$2v_M^2\eta_2 \times C_{\mathrm{eff}}(\rho_{\beta\alpha}; N_S, T)$

Table S2. The value of $[D(C_{eff} = 0)]_{1/2}$ according to assumptions made. Throughout this work, we have discarded all-or-none assumption and included symmetry numbers. See discussion of eq. 5 in main text.

	All-or-none assumed	All-or-none discarded
Symmetry numbers ignored	$\frac{1}{\nu_M}$	$\frac{1+\sqrt{5}}{2\nu_M}$
Symmetry factors included	$\frac{1}{2\nu_M}$	$\frac{1+\sqrt{2}}{2\nu_M}$

Table S3. Number of S-chain conformations N_{ens} and number of configurations $N_{config} = 2 \times N_{1or}^{R} \times N_{ens}$ (N_{1or}^{R} is number of receptor orientations relative to S-chain) used in calculation of $\eta_1 = \langle \exp(-U^{(SR)}/k_B T) \rangle_1$ and thus of $\Delta S_1/k_B = \ln(\eta_1)$.

	R_{hard}, convex, W=0, h=-1		R_{planar}, R_{hard} concave, W=0, h>0		R_{planar} $W \neq 0$, h=0		R_{soft}, $W \neq 0$, h=10	
N_S	N_{ens}	N_{config}	N_{ens}	N_{config}	N_{ens}	N_{config}	N_{ens}	N_{config}
3 to 300	5.E+06	3.E+10	5.E+06	3.E+10	1.E+07	5.E+10	5.E+06	3.E+10
380 to1500	5.E+06	3.E+10	5.E+06	3.E+10	5.E+07	3.E+11	5.E+06	3.E+10

Table S4. Number of S-chain conformations N_{ens} and number of configurations N_{config} used in calculation of $\eta_2 = \langle \exp(-U^{(SR)}/k_B T) \rangle_2$ and thus of $\Delta S_2/k_B = \ln(\eta_2)$. Columns showing N/A have no conformations with an end-to-end distance R_{ee} of around $\rho = 30$ Å, so that $n(\rho) = 0$. This means that such S-chains are too short (Ns = 3 to 31) to bridge the gap between binding sites, and are thus incapable of bivalent binding.

	R_{hard}, convex, W=0, h=-1		R_{planar}, R_{hard} concave, W=0, h>0		R_{planar} $W \neq 0$, h=0		R_{soft}, $W \neq 0$, h=10	
N_S	N_{ens}	N_{config}	N_{ens}	N_{config}	N_{ens}	N_{config}	N_{ens}	N_{config}
3 to 31	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
46	5.E+08	4.E+05	5.E+06	4.E+03	1.E+09	8.E+05	1.E+09	8.E+05
61	5.E+08	9.E+06	5.E+06	9.E+04	1.E+09	2.E+07	1.E+09	2.E+07
91	5.E+08	1.E+08	5.E+06	1.E+06	1.E+09	3.E+08	1.E+09	3.E+08
121	5.E+08	5.E+08	5.E+06	5.E+06	1.E+09	9.E+08	1.E+09	9.E+08
151	5.E+08	1.E+09	5.E+06	1.E+07	1.E+09	3.E+09	1.E+09	3.E+09
300	5.E+08	5.E+09	5.E+06	5.E+07	1.E+09	1.E+10	1.E+09	1.E+10
380	5.E+08	6.E+09	5.E+06	6.E+07	5.E+09	6.E+10	5.E+09	6.E+10
500	5.E+08	5.E+09	5.E+06	5.E+07	5.E+09	5.E+10	5.E+09	5.E+10
750	5.E+08	8.E+09	5.E+06	8.E+07	5.E+09	8.E+10	1.E+10	2.E+11
1000	5.E+08	1.E+10	5.E+06	1.E+08	5.E+09	1.E+11	1.E+10	2.E+11
1500	5.E+08	9.E+09	5.E+06	9.E+07	5.E+09	9.E+10	1.E+10	2.E+11

References

(1) Diestler, D. J.; Knapp, E. W. *J. Phys. Chem. C* **2010**, *114*, 5287-5304.
(2) Teraoka, I. *Polymer solutions: An Introduction to Physical Properties*, 2002.
(3) Marsaglia, G. *Ann. Math. Statist.* **1972**, *43*, 645-646.
(4) Neumann, J. v. *NBS Appl. Math. Ser.* **1951**, *12*, 36-38.
(5) L'Ecuyer, P. *Mathematics of computation* **1996**, *65*, 203-213.