Molecular Modelling in Drug Development

Michal Kolář

Doctoral Thesis

Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i.

Univerzita Karlova v Praze
Přírodovědecká fakulta

Modelování chemických vlastností nano- a biostruktur

RNDr. Michal Kolář

Molecular modelling in drug development
Molekulové modelování ve vývoji léčiv

Disertační práce

Vedoucí práce: prof. Ing. Pavel Hobza, Dr.Sc., dr. h. c., FRSC

Praha 2013

Abstract

Molecular modelling has become a well-established tool for studying biological molecules, moreover with the prospect of being useful for drug development. The thesis summarises research on the methodological advances in the treatment of molecular flexibility and intermolecular interactions. Altogether, seven original publications are accompanied by a text which aims to provide a general introduction to the topic as well as to emphasise some consequences of the computer-aided drug design.

The molecular flexibility is tackled by a study of a drug–DNA interaction and also by an investigation of small drug molecules in the context of implicit solvent models. The approaches which neglect the conformational freedom are probed and compared with experiment in order to suggest later, how to cope with such a freedom if inevitable. The noncovalent interactions involving halogen atoms and their importance for drug development are briefly introduced. Finally, a model for a faithful description of halogen bonds in the framework of molecular mechanics is developed and its performance and limits are tested by a comparison with benchmark *ab initio* calculations and experimental data.

Abstrakt

Molekulové modelování představuje etablovaný nástroj vědeckého výzkumu a nachází stále větší uplatnění i při návrhu léčiv. Disertační práce shrnuje výzkum počítačových metod pro popis flexibility molekul a mezimolekulových interakcí. Práce obsahuje sedm původních publikací a doprovodný text, jež si klade za cíl, uvést čtenáře do problematiky molekulových simulací a vysvětlit souvislosti s počítačovým návrhem léčiv.

Část o molekulové flexibilitě zahrnuje studii interakcí malé molekuly s DNA, a dále dvě studie o významu konformačních změn pro solvatační energie malých molekul. Přístup, jež flexibilitu molekul zcela zanedbává a molekuly považuje za rigidní objekty, je detailně zkoumán a srovnávám s experimentálními daty s ambicí navrhnout možnosti, jak konformační volnost molekul do výpočtů zahrnout v případech, kdy je to nevyhnutelné. V druhé části jsou představeny mezimolekulové interakce halogenovaných molekul a je zdůrazněna jejich role v medicinální chemii. Následně je zaveden nový počítačový model, jež umožňuje zjednodušený popis těchto interakcí; jeho kvalita a limity jsou testovány porovnáním výpočtů s referenčními *ab initio* a experimentálními údaji.

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

v Praze 4. dubna 2013

Michal Kolář

Contents

Contents ii

List of Figures iii

List of Abbreviations v

Preface 1

1 Introduction and the State of the Art 3

1.1 Computers in Science 3

1.2 Historical Context 4

1.3 Free Energy 6

1.4 Approximations 7

1.4.1 Conformational Sampling 7

1.4.2 Energy 9

1.5 Free Energy Decomposition 9

1.6 Aims of the Work 12

2 Molecular Dynamics as The Tool 13

2.1 Propagation in Time 14

2.2 Molecular Mechanical Energy Calculations 15

2.3 The Simulation Box 16

2.4 Point-Charge Models 17

2.5 The Solvent 20

2.5.1 Rigid Water Models 20

2.5.2 Implicit Solvent Models 21

3 Conformational Ensembles 25

3.1 Conformational Entropy 25

3.2 Conformations in Implicit Solvation 28
3.2.1 HIV-1 Inhibitors 28
3.2.2 Water–Octanol Transfer Free Energies 31

4 Noncovalent Interactions Involving Halogens 35
4.1 Halogen Bonding 35
4.2 Dihalogen Bonding in Crystals 37
4.3 Halogen Bonding in Molecular Mechanics 39
4.3.1 Drug–Target Interactions 40
4.3.2 Explicit σ-Hole 40
4.3.3 The Models 41
4.3.4 Parametrisation and Performance 42
4.3.5 Improved Docking 45

5 Summary and Outlook 47

Bibliography 49

List of Publications 63

Presentations of the Results 65

A Declaration of the Shared Authorship 67

B Publication 1 – Conformational Entropy 69

C Publication 2 – HIV-1 Inhibitors 81

D Publication 3 – Water–Octanol Transfer Free Energies 95

E Publication 4 – Hexahalogenbenzene Crystals 133

F Publication 5 – Explicit σ-hole 157

G Publication 6 – Docking with Explicit σ-hole 175

H Publication 7 – Halogen Bonds in Drug Development (in Czech) 181

List of Figures

1.1 Lock-and-Key Model of Drug–Target Interactions 6
1.2 Free Energy Contributions 11

2.1 Periodic Boundary Conditions 16
2.2 Explicit vs. Implicit Solvent Model 21

3.1 DNA Intercalation 26
3.2 RESP Charges Variations 29
3.3 Thermodynamic Cycles 31
3.4 Transfer Free Energy Distributions 34

4.1 Halogen Bond 36
4.2 Crystal of Acetone and Bromine Mixture 36
4.3 Halogen Intermolecular Interactions 37
4.4 Electrostatic Potentials of Hexafluoro- and Hexachlorobenzene . 39
4.5 Scheme of Explicit σ-hole 40
4.6 Gromacs Virtual Sites 41
4.7 Dissociation Curves of Bromobenzene–Acetone Complex 43
4.8 Comparison of ESP, RESP Performance 44
4.9 Root-Mean-Square Deviation Matrix 45
4.10 AR, CDK2, CK2 and HIV RT Enzymes 46

List of Abbreviations

AR – Aldose Reductase

BOA – Born-Oppenheimer Approximation

CADD – Computer-Aided Drug Development

CCSD(T) – Coupled-Cluster method with Iterative Single- and Double-Excitations, and Perturbative Triples

CK2 – Casein Kinase 2

COSMO-RS – Conductor-Like Screening Model for Real Solvents

DFT – Density Functional Theory

DNA – Deoxyribonucleic Acid

ESH – Explicit σ-hole

ESP – Electrostatic Potential

GB – Generalised Born

GBSA – Generalised Born, Surface Area

HIV – Human Immunodeficiency Virus

IE – Interaction Energy

LJ – Lennard-Jones

MD – Molecular Dynamics

MM – Molecular Mechanics

MST – Miertus, Scrocco, Tomasi

NMR – Nuclear Magnetic Resonance

PBSA – Poisson-Boltzmann, Surface Area

PM6-DH2 – Parametrised Model 6 - Dispersion and Hydrogen Bonding correction 2

rdf – Radial Distribution Function

RESP – Restricted Electrostatic Potential

RMSD – Root-Mean-Square Deviation

RMSE – Root-Mean-Square Error
RNA – Ribonucleic Acid
RT – Reverse Transcriptase
SAR – Structure–Activity Relationship
SMD – Solvent Model D
SPC – Simple Point-Charge
SQM – Semiempirical Quantum Mechanics
QM – Quantum Mechanics
TIE – Total Interaction Energy
TIP3P – Transferable Intermolecular Potential with Three Sites
vdW – van der Waals
XB – Halogen Bond

Preface

Computers surround our lives everywhere and it has become virtually impossible to avoid them completely, no matter how legitimate this ambition may be. In science, computers play several important roles. They control the action of scientific devices, ranging from small pocket calculators reaching to such complex facilities as Large Hadron Collider. In various scientific fields, however, computers do not serve as an operating component but rather as a source of scientific results. Computer simulations have been shown to provide valuable results which complement well the knowledge gained from other (*e. g.* experimental) sources.

In the past few years which I have spent at the Institute of Organic Chemistry and Biochemistry of the Academy of Sciences of the Czech Republic, I have had a chance to discover the computer simulations of the field, lying on the edge of theoretical chemistry, physical chemistry, biophysics and biology, and sharing the interest in biomolecules. By means of molecular simulations, I have been trying to answer the questions of pharmaceutical chemists, which can lead to more advanced drug development in the (hopefully near) future.

The thesis, which was formally prepared at Charles University in Prague in 2009–2013, aims to summarise seven publications. Six of the publications are either already published or in various stages of the publishing process in world-class peer-reviewed journals, and one presents some of the results in a popular Czech scientific journal. The thesis introduces some consequences which were difficult to present in the publications, while providing some details of particular outcomes of the publications.

The thesis is organised as follows: The first chapter connects the worlds of computer simulations and the pharmaceutical industry and answers the question of "why to do so?". Molecular modelling is briefly introduced with an emphasis on the historical context and drug development. The main advantages and shortcomings of the computer modelling of drugs are provided outlying several problems which I have faced. Next, the aims of the thesis are established. The second chapter explains classical molecular dynamics simulations as they represent the major tool used in this work. Several methods are mentioned in further detail, again to provide the reader with better insight into the publications. The third and fourth chapters present the publications, and finally the fifth chapter summarises the work and brings an outlook for the future.

During my work I spent most of the time in the Prague group of Prof. Pavel Hobza. I express my deep gratitude to him, not only for his supervision of my Ph.D. work but also for his personal approach. I can clearly recollect several situations of desperation of mine, where he used all the means possible to show me the right direction. Thanks to him, I also found the courage to travel across Asia by train, which interestingly affected my scientific career.

I thank the people from “Kaňon” who ha led me through the everyday troubles and shared the space and time with me there. Namely I thank Robo, Adam, Jindra, Martin L., Jirka V., Susanta, Martin B., Filip, Tomáš and many others. Especially, I am grateful to Tom K., who managed to pass a great deal of his knowledge before he left for Germany. Also I thank Jirka P., who was the only person with the Password and the Key all the time.

I thank Tros Pedros: Cigi, Houser and Slavíček; Tom, HV and Zuzka, Lumec and Marodkář, who have been involved in the Chemistry Olympiad and in the extraordinary event held every year in Běstvina for the exceptional motivation and for the unceasing source of ideas. The science in their hands appears to be a joy. I also thank Bronka, Tom and Javier, with whom I have had a great opportunity to collaborate on international scientific projects. I am grateful to Bronka and Tom for valuable comments on the thesis as well as to Mrs. Miller for language proofreading.

Finally with special care, many thanks go to my family and friends who have experienced science from the common man’s side. I thank Romana, Nina Marie, Zuzanka and my parents for the tolerance they have had and for the support they have provided me. Vřelé díky!

1

Introduction and the State of the Art

It would be naive to view the effort of the pharmaceutical companies as the genuine will to save human lives and increase their quality. Sadly, in agreement with the recent trend in the world in general, also in drug development by far the most eminent role is played by economic factors. How to earn quickly a lot of money may be the holy grail of any business; unfortunately, the development of a new drug does not fit these criteria at all. It is very time-consuming and costly!

1.1 Computers in Science

The average time needed to introduce a new drug on the market ranges from 7 to 12 years and the development of a drug may consume as much as $ 2 milliard [1]. Evidently, the efforts are to save the resources by any means. The enormous costs rise from the fact that the company has to know literally *everything* about the drug before it can make the drug available for the market. In short, the structure and physico-chemical properties of the compound must be identified, and both the targeted biomolecule and the way in which the drug reaches the target must be known. Moreover, it is necessary to discover what happens with the drug after *the work is done*. These properties are often known as ADMET (absorption, distribution, metabolism, excretion, toxicity). To emphasise the expenses, one has to realise that as claimed only one of 40,000 drug candidates tested on animals proceeds further [2].

In various pre-clinical phases (earlier first tried on humans) the computers may show their benefits. They are essentially faster, cheaper, and in some sense more "green" (less wasteful) when compared to traditional *in vitro* and *in vivo* screenings; it is therefore no coincidence that computer simulations and theoretical models have become a valuable source of scientific results. The ever-increasing power of computers has so far obeyed the Moore's law [3], which puts stress on scientist and software developers to keep up with hardware progress.

For instance, when I started my Ph.D. studies, a computer called Roadrunner, located in Los Alamos, USA, was the most powerful computer according to the `top500.org` ratings with its peak performance about 1.0 PFlop/s achieved by about 130,000 cores. In the last ratings published in November 2012 by the same organisation, the Roadrunner computer reached the 22nd position, with the winner – the Titan Cray XK7 computer from the Oak Ridge National Laboratory, USA – performing more than 17 times faster (17.6 PFlop/s with 561 thousand cores).[1] Such extended computing systems require innovative algorithms to solve numerical problems.

Actually, as stated by Dirac in 1929 [4] and frequently repeated ever since, "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known. . ." Dirac also admitted that the description of the reality leads us to the equations too complicated to be solved. It must be added – equations too complicated be solved in *a reasonable time*, because it is the question of patience rather than capabilities.[2] Consequently, some approximations are commonly applied to accelerate the calculations into accessible time scales. These approximations shape the research in the fields of computational chemistry or computer-aided drug design (CADD).

It is fair to say that the above-mentioned computers would be too expensive even for pharmaceutical companies; nevertheless, the motivation to use faster and more powerful computers is absolutely apparent. Especially for biological and chemical problems in drug development, a wide range of computational approaches have emerged, differing in the extent of the approximations they adopt and in computational demands. A few of them related to the thesis are discussed in the following sections.

1.2 Historical Context

When the *trial-and-error* approach became inefficient with the increasing number of identified diseases and their possible complexity, drugs started to designed in a way which is now called *knowledge-based* [2, 5, 6]. The new drugs are intentionally designed after their structural features and the mechanism

[1] The performance is measured by the *LINPACK* benchmark suite written by Jack J. Dongarra

[2] In principle, one could obtain highly accurate data on drug–target interactions from *e. g.* quantum mechanics, but this can easily take hundreds of years even on such extensive computers which are available today. And honestly, no one can afford to wait so long.

of their activity have been understood.

It is accepted that most drugs exert their activity by a fairly specific binding to their biological targets. Such a target can be an intracellular protein [7, 8], a piece of nucleic acid [9, 10], a transmembrane receptor [11, 12] or channel [13, 14]. On the other hand, drug-related side effects are caused by off-target binding, and a certain level of drug promiscuity must be expected. In some applications though, the multi-target binding features can be exploited intentionally.[15] In the course of this thesis, I will focus on single-target–ligand applications. This area is typical for CADD although it has contributed also to other areas, such as predictive toxicology [16, 17] or the analysis of drug adsorption [18, 19].

Over time two major lines of CADD have formed. The first one is a *virtual screening* of compounds which is a computational analogy to high-throughput screening. In virtual screening, the existing compounds are computationally tested for their affinity to particular targets [20, 21]. Since the databases contain millions of such compounds, there is a valid expectation that some of them might be active; the role of CADD is to select the best ones. The second is *de novo* design, where new chemical scaffolds are suggested and tested for the activity by computational algorithms [22].

Once a promising candidate or a group of candidates (so-called *lead compounds*) is identified, further optimisation of their chemical structures is done either in computers or experimentally. This lead optimisation aims to increase affinity. The exact workflow of the lead generation and optimisation is likely to differ between the companies and academic organisations and it belongs to their confidential property.

Vast majority of drugs exert their activity via noncovalent binding to their targets [23, 24]. The noncovalent interactions play a central and essential role in the living organisms in general. They are weak but numerous, thus effectively strong enough to maintain the structure and function of biomolecules, but their weakness makes it possible to adapt the structure and function to the external stress at a low energy cost. Among others, hydrogen bonding and stacking interactions have been emphasised for the biomolecular functionality. Favourably, a theoretical description of noncovalent interactions seems to be based on a solid ground with the pioneering work done already in the early 1970s [25–27], however as described in the following chapters, there is still some space for improvement.

A typical research case in the context of rational drug development is a noncovalent ligand-enzyme interaction. Ligand binding affects the enzyme function and it is assumed that there exists a direct relation between the drug binding affinity, which is a local microscopic phenomenon, and the observed therapeutic effect, *i. e.* the macroscopic manifestations of the changes appear-

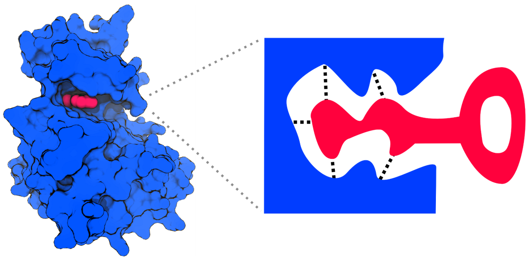


Figure 1.1: Lock-and-key model of drug–target interactions. The enzyme in blue represents the lock. The drug (red key) must match the lock. Each of the dotted black lines stands for a noncovalent interaction (*e. g.* hydrogen bond or stacking contact).

ing on a biomolecular scale.

Drug binding is often compared to a lock-and-key model (Figure 1.2. The target (enzyme) with the unique active site is represented by the lock, and the desired drug is the key which must match the lock to be active. The favourable interaction is ensured by particular kinds of noncovalent interactions, each of which can be viewed as the lock pin. These interactions contribute to the binding free energy.

1.3 Free Energy

The binding free energy is the physical quantity describing binding affinity and it is the appropriate subject of CADD calculations. The statistical thermodynamics [28] provides a formula interrelating the free energy A with the partition function Z (Equation 1.1)

$$A = -k_B T \ln Z \tag{1.1}$$

where the partition function Z for the classical continuum spaces of positions $\vec{x}$ and momenta $\vec{p}$ is defined as

$$Z = \frac{1}{N! h^{3N}} \int e^{-\frac{H}{k_B T}} \, \mathrm{d}\vec{x} \, \mathrm{d}\vec{p} \tag{1.2}$$

where N stands for the number of particles, H is the classical Hamiltonian, T is the temperature, h is the Planck's constant and k_B is the Boltzmann constant. From the above two expressions it is possible to derive an instructive formula for the free energy change, which was first done by Zwanzig almost 60 years ago [29].

$$\Delta A_{I \to F} = -k_B T \ln \left\langle e^{-\frac{E_F - E_I}{k_B T}} \right\rangle \tag{1.3}$$

$\Delta A_{I\rightarrow F}$ stands for the free energy change between two states, denoted as I (initial) and F (final), at the temperature T. E is the total energy of the state, and the angle brackets represent the statistical ensemble average. The Zwanzig formula unravels two aspects of the free energy calculations: the need for an energy calculation and an ensemble sampling. Both tend to follow the undesirable rule in computational chemistry – better accuracy is more computationally intensive. The ensemble sampling consists of many energy calculations, thus more accurate energies we have, less extensive sampling we can afford (see below).

The Zwanzig formula is fundamentally exact for any chemical/physical change between the initial and final states; on the other hand, it requires correct ensemble sampling and accurate energies as the input for it to yield a reliable output. The beauty of the Zwanzig formula is that it becomes clear from it, where to save computational time, either in energy calculations or in the ensemble sampling.

1.4 Approximations

A number of strategies have been developed to estimate drug potency even without explicitly calculating the binding free energy. A large group of approaches completely neglects the structural information on the target, which makes them extremely efficient, but of a very limited value in terms of explaining the prospective drug's mechanism on the molecular level. The quality of the drug candidate is determined on the basis of its physical and chemical properties based on their similarity with the compounds whose activity and properties are known. These ligand-based approaches are collectively called Structure-Activity Relationship (SAR) [30–32]. Conversely, the similarity is difficult to estimate and the comparison with the empirical reference data set does not provide sufficient physical insight.

1.4.1 Conformational Sampling

The structure-based approaches, where the target information is the central aspect in the calculations, maintain the physical description of drug–target interactions. The interacting complex may be described at various levels of detail. In fact, the atomistic structure of the target may be the keystone and the only prerequisite needed for the calculations. More than 75,000 structures of biomolecules have been determined and stored in the Protein Data Bank (pdb.org), mostly by X-ray crystallography or Nuclear Magnetic Resonance (NMR) experiments. The natural form of the X-ray diffraction experiments provides a single geometry of the target – the time average of the dynamical ensemble occurring in the crystal. The dynamical feature of the biomolecules is largely omitted and included only in the form of the B-factors of the atoms.[3] This is the source of the standard approximation which neglects the dynamical behaviour of the drug–target complex. The approximation is called *single-conformation approach*, here. Another option to reduce the ensemble-sampling demands is to consider only some of the most important conformations. Then the questions become what criteria should be used for the selection of the conformations and how to analyse the selected conformations.

It is likely that the single-conformation approach may also arise from the experience acquired in quantum mechanics (QM) frequently relying on the Born-Oppenheimer approximation (BOA) [33]. The BOA claims that the electronic degrees of freedom can be separated well from the nuclei degrees of freedom in a molecule. The electronic energy calculations are notably simplified when BOA is adopted. To be critical, the application of the BOA, *i. e.* the calculation of the electronic energies and derived properties for a single spatial arrangement of atoms, supported in the community of computational chemists the idea of using the single conformation also for other kinds of calculations, where this concept is not directly related to the BOA and is, hence, much less justified. In majority of CADD calculations the BOA holds true, indeed, but the reasons why to consider only single conformations are likely to be different.

The single-conformation approximation seems to be relevant to small molecules, where the ensemble of conformations can be sufficiently replaced by a single representative,[4] but the use of single conformations for biomolecular calculations might be questioned. For virtual screening, the single conformation of the target biomolecule is used almost exclusively [34]. However, as noted by Schneider [35], “Dynamical description of the molecules will still have to replace our predominantly static view of both targets and ligands.”

This has already begun to happen, yet on a small scale. It appears that all the necessary algorithms were developed in the past but the computational power has made it possible to apply them only recently. For instance, rather standard *old-fashioned* molecular dynamics simulations were used to estimate ligand-binding affinities including the conformational sampling of the ligand–target complex [36]. It was demonstrated that the method is particularly useful for the lead optimisation. More recently, Stelzer *et al.* [37] have performed a successful virtual screening against the ensemble of RNA conformations. The conformational ensemble was however prepared by using not only the computational techniques but NMR as well. This thesis is an indirect proof that the number of studies on conformational ensembles increases.

[3] The higher B-factor the atom has, the more mobile it appears.

[4] Since the benzene molecule does not seem to deform dramatically upon temperature fluctuations or phase transitions, one static D6h geometry is sufficient.

1.4.2 Energy

The approximations to the energetics of the studied system and the choice of the way in which the energy is calculated seem to be the main focus in computational chemistry. Generally, it is accepted that the high-level wave function QM calculations can yield highly accurate energies for chemically relevant problems with the advantage that the energies can be improved systematically. The accuracy of the *golden standard* for the noncovalent interactions – CCSD(T)[5] – is claimed to be below 1 kcal/mol (so-called chemical accuracy) [24]. The scaling of the CCSD(T) method with the system size is unfavourable as it is of the order of N^7, with N being approximately the number of orbitals. The CCSD(T) interaction energies are thus not suitable for virtual screening and used more as a reference for lower-level QM methods.

For the energy calculations on biomolecules, in particular in conjunction with the conformational sampling, the QM methods are still too demanding and an additional simplification is needed. The use of semi-empirical QM methods (SQM) and the density functional theory (DFT) in biomolecular calculations has been rapidly growing among others thanks to the accessibility of parallel algorithms run on supercomputers or graphical processor units [38, 39].

An even more simplistic method is the molecular mechanical (MM) approach, where the internal atomic structure is neglected and the atoms are described as classical objects. The internal energy and intermolecular interactions are described by a set of empirical parameters, *i. e.* a force-field, derived either from higher-level calculations or the experimental data. Biomolecular force fields must be tested and verified to reproduce faithfully experimental structural and dynamical data and their quality determines the success or failure of the calculations. Back to the Zwanzig formula, molecular mechanics seem to be the best suited for the energy calculations there, since they are fast enough also for extended ensemble sampling. This compromise between the accuracy of energy calculations and the extent of conformational ensemble sampling has certain limitations, and this thesis aims to uncover some of them.

1.5 Free Energy Decomposition

Gibbs free energy change ΔG has two thermodynamic contributions: the enthalpy change ΔH and the entropy change ΔS (Equation 1.4). If the change is demonstrated by the ligand binding, the quantities are classified as the *binding* free energy, *binding* enthalpy and *binding* entropy.

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

[5] The coupled-cluster method with iterative single- and double-excitations, and perturbative triple-excitations

The enthalpic part covers the changes in noncovalent binding patterns, *e. g.* the number and quality of hydrogen bonds. The favourable noncovalent interactions between the ligand and its cognate target are compensated for by the unfavorable desolvation of the ligand. Once the ligand passes from the aqueous environment into the hydrophobic cavity of the target, the ligand loses the surrounding solvent molecules, which are bound mostly by hydrogen bonds. If the noncovalent interactions established upon ligand–target complexation are strong enough when compared with the noncovalent interactions between the ligand and solvent molecules, then the binding enthalpy is negative and favours binding.

On the contrary, the entropic part comprises two major contributions: i) the loss of the conformational freedom of the ligand upon binding and ii) the release of the solvent molecules bound to the ligand. The former contribution – the conformational entropy [40–42] – is unfavourable. The ligand tends to maximise its conformational freedom, and this is allowed in the solvent rather than in the target active site. On the other hand, the latter contribution favours binding thanks to the higher mobility of the solvent molecules when unbound from the ligand [43, 44]. This is true also for the desolvation of the binding cavity inside the target, though not all enzymes have their binding pockets solvated.

In drug–target binding, the effect of enthalpy-entropy compensation is often observed [45, 46]. The more tightly the ligand is bound into the active site, the more entropy it loses by decreasing its flexibility. As reviewed recently [47], entropy optimisation seems to be easier than the enthalpy optimisation, thanks to the conformational constraining strategy applied on the molecules. The structural scaffolds have been design to minimise the conformational freedom in the aqueous phase. Consequently upon binding, the relative difference in the flexibility between the bound and unbound states is minimised.

In computational chemistry and CADD, the binding free energy is often approximated by a scoring function (*e. g.* Refs. [20, 34, 48–50]). For virtual screening, the function must be simple enough to probe millions of compounds and two classes of scoring functions have been developed. One class includes the scoring functions which have the form of a sum of various energetic terms, often representing particular noncovalent interactions. The weight of such terms might be subject of empirical adjustment to reflect experimental data better. Another class comprises knowledge-based scoring functions [34, 49]. These scoring functions are constructed by an analysis of existing structures of drug–target complexes, from which the distance dependent atomic pair-preferences are extracted.

A particular form of free energy decomposition has become popular in CADD: the ligand–target binding free energy consists of contributions which reflect dif-

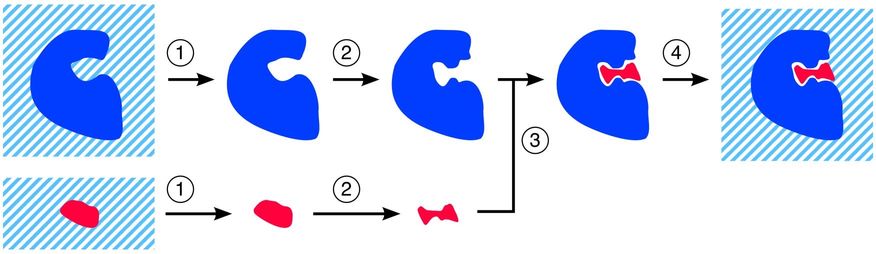


Figure 1.2: Phenomenological decomposition of binding process. First, the enzyme (blue) and the ligand (red) have to be desolvated. Second, they must be deformed to mach each other. Next, they interact to create a complex and finally, the complex is solvated back.

ferent phenomena arising from the binding. In an idealised way, the binding may be viewed as a sequence of several steps. The ligand and the active site need to be desolvated first. Then the ligand and the active site are deformed (*i. e.* conformationally changed) to match, and finally the ligand is inserted into the active site (Figure 1.5) upon creation of new favourable contacts.

Then the binding free energy contains the interaction energy term, reflecting the ligand–target interaction, the solvation/desolvation term and a contribution which stands for the flexibility changes of both the ligand and the target. Optionally, the energy term covering the deformation may be included.

$$\Delta G = \text{interaction} + \text{solvation/desolvation} + \text{flexibility change} \qquad (1.5)$$

Perhaps, the most spread method based on such decomposition is abbreviated as MM/PBSA or MM/GBSA. MM stands for the molecular mechanical treatment of the energies and deformations, and GBSA and PBSA are the solvation free energy methods [51, 52]. The method also includes some conformational entropy change (approximated by a quasiharmonic approach or a normal mode analysis, see bellow), which does not, however, occur in the abbreviated name.

In the group of Prof. Hobza a scoring function based on the single-conformation approach and SQM energies has been developed over a last couple of years [53] The energetics are calculated at the PM6-DH2 level [54, 55] and the conformational sampling is omitted. This is an example of a scoring function employing energetics at a better than MM level of theory [56–60]. Because of the computer demands, this kind of scoring function is not yet suited for virtual screening, although a considerable effort has been made to overcome the issue of the speed of computations.

The PM6-DH2 scoring function contains the following terms: the ligand–target interaction energy, the solvation free energy change upon binding, the conformational entropy change approximated by either a simplified approach using rotatable bonds or the vibrational entropy and the deformation contribution of both the ligand and the target [61].

1.6 Aims of the Work

The thesis focuses on the issues introduced and briefly characterised in the previous paragraphs. Namely, the two of them that are addressed include i) the importance of conformational sampling for the description of biomolecules and biologically relevant small molecules and ii) the theoretical description of noncovalent interactions involving halogen atoms, both in the context of computer-aided drug development. The major questions regarding the conformational ensembles that I have tried to answer are:

- What is the role of flexibility changes upon ligand binding into a DNA double-helix?
- What is the error magnitude brought by the single-conformation approach for the flexible ligand hydration free energy?
- How to go beyond the single-conformation approach in solvation free energy calculations?

These are thoroughly introduced and discussed in Chapter 3. The major questions that I have attempted to answer with respect to the halogen bonding are:

- What is the nature of the dihalogen bond in the model complexes appearing in the crystal phase?
- How to describe halogen bonding at the molecular mechanical level?
- What are the performance and limitations of the MM description of halogen bonding?
- How to apply the MM description of halogen bonds for the prediction of the ligand–target structures?

The noncovalent interactions involving halogen atoms are introduced and discussed in detail in Chapter 4. From above mentioned, it may be clear that the thesis contributes to the field of computer-aided drug development mostly by methodological advances. Some tools and techniques for CADD are proposed and tested but their direct application to virtual screening or *de novo* design of new drugs remain a task for future.

2

Molecular Dynamics as The Tool

In the following sections the details on the methods used for the computations are provided. Most of them are focused on the classical molecular dynamics simulations, but other associated techniques are briefly introduced as well. Special attention is paid to non-standard approaches, which are further developed in Chapters 3 and 4.

The class of the methods exploring the dynamic behaviour of system is referred to as molecular dynamics (MD); it was introduced by Alder and Wainwright [62]. The form of particle description and the manner in which the time evolution of the system is tackled are among the principal concerns yielding a large number of MD flavours. Molecular dynamics simulations have already been described in detail elsewhere [63–65], so only the parts of the theory relevant for the attached publications are highlighted here. A valuable source of implementation details may be the manuals of the MD simulation packages such as Amber or Gromacs [66, 67].

MD simulations are excellent for free-energy calculations. For CADD the studied systems are often further simplified, but the concept remains the same. Performing MD simulations successfully requires the fine tuning of many simulation parameters. The best practice setups differ from task to task, and the optimisation of the setup is claimed to be the essential skill of computational chemists. Unless it is necessary, the details of the simulations are provided in the attached publications and are not the subject of further explanation in this chapter.

For MD simulations, is worth to mention an important assumption – the ergodic theorem, which states that the statistical time averages are equal to statistical ensemble averages. It is assumed that for a long enough time, the system is able to achieve all possible states, and since MD simulations genuinely provide the time evolution of the system, the ergodic theorem eventually makes MD applicable to real-world problems.

2.1 Propagation in Time

In the framework of classical MD, the time evolution of the system obeys Newton's law. This makes it possible to use such a method only for the problems where the classical (*i. e.* non-quantum) treatment of laws of motions is appropriate, which molecular modelling of biomolecules mostly is. The propagation in time is computationally approached by solving Equation 2.1 numerically.

$$\frac{\partial^2 x}{\partial t^2} = \frac{F}{m} \tag{2.1}$$

where x stands for the position of particle with mass m, and F is the force acting on the particle. The time step Δt has to be chosen as a parameter, and the unknown positions and momenta in the *future*, time of Δt from *now*, are calculated from the known positions and momenta either from *past* (*i. e.* from the previous time steps), or from their actual values. The time propagation proceeds until a desired number of steps is reached.

Upon the numerical propagation the total energy of the system and momenta of atoms may not be conserved precisely. However, this does not seem to be a problem since there are employed other algorithms, such as for temperature control, which disrupts the energy conservation intentionally, providing the desired statistical ensemble.

To be able to propagate the positions x and momenta p, the force acting on each of the particles has to be calculated in each step of the propagation by Equation 2.2

$$F_i = -\frac{\partial V}{\partial x_i}, \tag{2.2}$$

where V stands for the potential energy. As introduced in Section 1.4, there is plenty of choice of the level at which the potential energy is calculated. The choice of the level depends on the kind of the problem, the phenomenon aimed to described and the available computational power. For biomolecular simulations in the aqueous environment, the molecular mechanical treatment of energies is preferred.

2.2 Molecular Mechanical Energy Calculations

At the MM level, the potential energy of a set of particles (atoms and/or molecules) is calculated by Equations 2.3 to 2.7.

$$\begin{aligned}
E_{pot} &= \sum_{bonds} K_r(r-r_{eq})^2 + && (2.3)\\
&+ \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + && (2.4)\\
&+ \sum_{dihedrals} K_\phi \left(1+cos(n\phi-\phi_{eq})\right) + && (2.5)\\
&+ \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + && (2.6)\\
&+ \sum_{i<j} \frac{q_i q_j}{r_{ij}}. && (2.7)
\end{aligned}$$

There are two kinds of contributions: i) the bonded interactions include the terms characterising bond stretching (Eq. 2.3), angle bending (Eq. 2.4) and torsional deformations (Eq.2.5), while ii) the nonbonded part includes Lennard-Jones (LJ) term (Eq. 2.6) and Coulomb electrostatic energy (Eq. 2.7). Each of the bonded contributions is described by two parameters - the force constant of bond, angle and the torsional angle (K_r, K_θ, or K_ϕ) and their equilibrium value (r_{eq}, θ_{eq}, ϕ_{eq}). For torsional contribution, the multiplicity of the energy profile may also be included. The Coulomb interaction depends on interparticle distance r_{ij} and two partial charges, q_i and q_j, and finally, the LJ interaction term has two parameters: ε and σ for each pair type.

The form of MM equations directly points to the situations which are not properly addressed by molecular mechanics. The harmonic potential describing bonding does not allow the bonds to be broken during the MD simulation. As a consequence, any simulations of the changes of chemical bonding[6] must employ a higher level for energy determination, usually the semiempirical QM or density functional theory. Fortunately as mentioned above, the changes on the biomolecular scale, such as drug-target binding, often employ noncovalent bonding, for which Equations **??** to 2.7 provide a suitable description.

The set of the parameters in Equations 2.3 to 2.7 is generally called force field, and the parameters must be determined for each atom or atom type, angle type etc., separately. Many variants of biomolecular force field have been developed so far, interestingly many of them a long time ago [68–71]. More recently, merely (small) correction of existing force files have been appearing [72–74]. The force fields were designed to reproduce experimental macroscopic properties such as densities of simple liquids, vaporisation enthalpies or dielectric constants, or were adjusted to the higher-level QM data (torsion

[6] *i. e.* chemical reactions

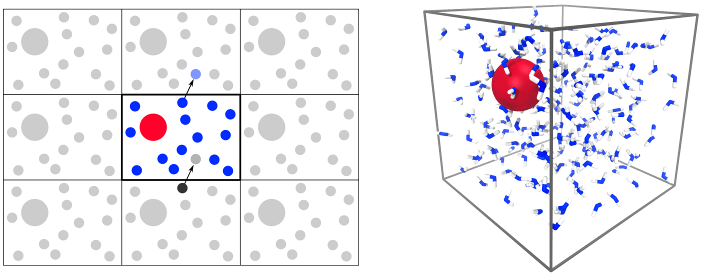


Figure 2.1: Periodic boundary conditions in two dimensions (left) and a 3D simulation box of water (right).

profiles, dissociation curves etc.).

The pair-wise character of the nonbonded interactions makes it possible to design efficient algorithms to accelerate the simulations [75]. On the other hand, the many body effects [76, 77] are either excluded, or included in the implicit manner by the nonbonded parametrisation. Thus, the intermolecular potentials are denominated as to being *effective*.

The functional form of the MM energy (Equation 2.3 to 2.7) does not usually contain any terms representing specific noncovalent interactions, although this is not the case of all biomolecular force fields. For instance, early versions of the biomolecular force fields contained an energetic contribution, the purpose of which was a correct description of hydrogen-bonding patterns [68]. Nevertheless, modern force fields are able to describe hydrogen bonding quite well, only by the combination of LJ and Coulomb interactions. A problem appears when unusual bonding motives are supposed to be described at the MM level. This is the case with halogen bonding illustrated in Chapter 4.

2.3 The Simulation Box

A realistic simulation system of interest in CADD could be composed of a solute (*e. g.* a drug-target complex) and a solvent (*e. g.* ions and water molecules) (Figure 8). The aim is to approximate the macroscopic view,[7] thus the ambition would be to simulate as many particles as possible. The computational expense of calculation of energy scales with the second power of the number of atoms N, which is caused by the nonbonded interactions calculated for each pair (Equations 2.6 and 2.7). For large N,[8] the calculation tends to be unfeasible. Consequently, some algorithms have been developed to save computer time.

A clever way of how to follow the bulk-like reality is to apply the periodic boundary conditions for the computer models (Figure 8), which practically makes them infinitely large.

The former problems arising from the finite size of the system and its unrealis-

[7] One mole of water contains Avogadro's number of molecules which is of the order of 10^{23}, and three times more atoms!

[8] now *large* means already 10^5

tic boundaries are now substituted by the artificial periodicity. The periodicity problem can be partially solved by the sufficient size of the periodic box and by its proper testing, since some effects (*e. g.* solvation patterns, ion interactions) are propagated only to short distances in the condensed phase and, as specified below, they are counteracted anyway.

Still, there may be too many particles in the periodic box to comprehend all nonbonded pair interactions in the energy calculation. Nowadays, the number of the particles in biomolecular simulations may reach several hundreds of thousands. For such short-range noncovalent interactions as LJ interaction, the *cut-off* scheme has been shown to be appealing approximation, where the pair interaction is calculated only for the closest particles, within some predefined distance from the reference particle.

For Coulomb electrostatic interactions between partial atomic charges some more advanced treatments have been developed, such as the Particle Mesh Ewald summation [78, 79] or the reaction field method [80]. The reason for this was the fact that the simple cut-off approach was demonstrated to be inaccurate [81–84].

2.4 Point-Charge Models

One of the first really successful biomolecular force fields was designed by Cornell *et,al.* in 1994 [70], which later gave rise to a whole family of Amber force fields [70, 72, 85–87]. The success was attributed to the adequate determination of the partial atomic charges which in turn resulted in reasonable conformational energies so important for the biomolecular applications.

Partial atomic charges are a widely used concept in chemistry and also for the description of the electrostatic interactions in molecular mechanics, despite many concerns that have been presented [88–90]; most importantly, there are no physical observables for partial charges. The question is: how to describe the delocalised electronic density of a molecule (the reality) by a set of point charges localised on the atoms (the model)?

One way is to use the electronic density obtained at some QM level to determine the electrostatic potential (ESP) generated around the molecule of interest (Equation 2.8, in atomic units)

$$ESP(r) = \sum_{J}^{N_J} \frac{Z_J}{|r - r_J|} + \int \frac{\rho\left(r'\right)}{|r - r'|} \mathrm{d}r' \tag{2.8}$$

where the sum goes over all nuclei of atomic number Z and the integral covers the electron density $\rho(r)$ contribution over the entire space. This is usually done for a discrete grid of points surrounding the molecule with the shortest distance of the atomic centre being its van der Waals (vdW) atomic radius.

Then the atomic charges are the subject of least-square fitting to reproduce the ESP grid values [91, 92]. The fit is performed in such a way that the net charge of the molecule is preserved. The quality which is minimised is the square difference between QM ESP $V(QM)$ and ESP generated by the trial MM charges $V(MM)$ (Equation 2.9)

$$l^2 = \sum_{i}^{N} \left[V_i\,(\mathrm{QM}) - V_i\,(\mathrm{MM}) \right]^2 \tag{2.9}$$

where the summation is taken over all grid points i and the trial ESP for each grid point i is calculated from the point charges j according to Equation 2.10.

$$V_i\,(\mathrm{MM}) = \sum_j \frac{q_j}{r_{ij}} \tag{2.10}$$

where r_{ij} is the distance between a point charge and an ESP grid point. The minimisation is carried out until the convergence criteria (*e. g.* the energy difference between l^2 of two consecutive steps) are reached. As the initial *guess* charges, the Mulliken charges [93] based on QM Hatree-Fock (HF) [94–97] or some SQM method are used.

In the early work of Momany [91], the experimental molecular dipole moment was included into the fitting scheme. Without such correction, the resulting charges tended to represent rather the QM dipole moments, which differ from the experimental ones notably, at the QM level used at that time.[9] This was, however, not the only problem with the ESP fitting. It was found that the ESP charges were too conformationally dependent, which was more pronounced especially for the buried atoms such as sp^3 carbon, causing the low transferability of the ESP charges between the same fragments on different molecules. Although the intermolecular interactions seemed to be well reflected by the ESP charges, they were not well suited for the intramolecular interactions. Luckily, this deficiency was overcome by Bayly *et al.* [98] who introduced *restricted* electrostatic potential (RESP) fitting. In their work they admitted "Time will tell whether this approach is the best for deriving effective two-body potentials, but the consistent use of ESP charges for any molecule or fragment... offers a most promising approach to biomolecular simulations which is easily generalisable and aesthetically pleasing and consistent."

The essence of RESP is the penalty function aiming at restraining the non-hydrogen charges to a targeted set of charges. The subject of minimisation now has two contributions (Equation 2.11):

$$l^2 = l^2_{ESP} + l^2_{PEN} \tag{2.11}$$

The first term equals Equation 2.9 and the second term is the penalty function itself, preferably in the hyperbolic form (Equation 2.12)

[9] The QM dipole moments were calculated from the Mulliken charges derived at the HF level with the minimal basis set.

$$l_{PEN}^2 = a \sum_j \left(\sqrt{q_j^2 + b^2} - b \right) \tag{2.12}$$

where a and b are the parameters defining the *strength* of the penalty. As the target charges, the zero charges are preferred to the Mulliken ones. Additionally, the constraints are included to make sure that *e. g.* methyl hydrogens have identical charges (as required by the local symmetry). Undesirably, the hyperbolic form of the penalty function leads to the iterative solution of the equations.

Bayly *et al.* have also provided the best practice setup for determination of the RESP charges [98, 99]: The HF/6-31G* level of theory used for the reference ESP grid determination is claimed to yield overpolarised charges, which are, however, compatible with the popular models of water, TIP3P and SPC (see below). Well balanced solute-solute and solute-solvent interactions are thus ensured, despite the lack of explicit polarisation in MM.

The standard RESP fitting is performed in two stages: in first stage all of the charges are optimised with weak penalty on non-hydrogen atoms, and in the second stage all the charges are fixed except those in methyl and methylene groups, which are re-optimised with strong pentalty on non-hydrogen atoms (*e. g.* carbons) [99].

The grid of the reference ESP points has to be defined carefully as well. As stated by Singh and Kollman [92], the grid points located within the vdW radius of the atoms cause very high variations in the charges. The recommended grid starts at a distance of 1.4 times the vdW surface of the molecule and reaches up to the double vdW radius from the atomic centres. The surface density is normally 1 ESP point per Å^2.

Momany [91] also proposed a modification of the weight of those ESP points which surround the more important part of the molecule. Later, it was proposed [100] that a denser grid may provide better results; the density 1 point/Å^2 reflects the best demands/accuracy ratio in the early 1990s. This is however not valid nowadays. Some further criticism appeared while, for instance, the non-linearity of the equations and the linear dependencies of the ESP grid points was questioned. Nevertheless finally, it is worth mentioning Bayly *et al.* [98] again: "Although the solution we propose may not be the final answer, we feel the work here is the major step… in making ESP derived charges a general and useful way to generate atomic charges for simulations of complex systems."[10]

The RESP charges have found a place in the modern biomolecular force fields of Amber family, which were systematically used also in the studies presented by the thesis. The force fields of protein and nucleic acids employ these charges [70, 72, 85–87]. Since biomolecules consist of a small number

[10] It turned out they were actually right! March 2013, the Web of Science's number of citations of the RESP papers: Bayly *et al.* [98] > 2,050, Cornell *et al.* [99] > 530 and Cornell *et al.* [70] > 6,180.

of building-block types (aminoacids, nucleotides), their charges were earlier derived for the building blocks and have been used ever since.

The situation around the small drug-like molecules is different: the molecules may consist of many distinct chemical fragments, and it is advantageous to calculated RESP charges for each particular molecule alone. The General Amber Force Field (GAFF) [101] was designed exactly for this purpose – to provide intra- and intermolecular interactions compatible with the biomolecular Amber force fields. This is a truly CADD direction of the RESP application. Currently, it is possible to calculate RESP charges in automated *black-box-like* fashion using *e. g.* Antechamber [102] or the online R. E. DD. B. tool [103].

2.5 The Solvent

An important part of the simulation box is the solvent. One has to bear in mind that the solvent often represents the majority of particles for which the force has to be calculated during the MD. This section presents the ways in which the computational time attributed to the solvent can be reduced.

2.5.1 Rigid Water Models

The flexibility of the water molecule is usually the first to be neglected in CADD MD simulations. There have been many atomistic models of water proposed; the most used include the Simple Point Charge (SPC) model [104] and the Transferable Intermolecular Potential with three interacting sites (TIP3P) [105].

The water intermolecular interactions are the same as described in Equations 2.6 and 2.7. The oxygen carries both the negative charge and the Lennard-Jones attraction and repulsion, whereas both hydrogens carry only the positive charges of a half magnitude as compared to the oxygen. Thus in both models, the water appears as a soft sphere with a rigid triangle inside carrying dipole.[11]

The charges and the LJ parameters were adjusted to represent liquid water properties such as density and vaporisation enthalpy. Further, the quality of the oxygen-oxygen radial distribution function (rdf) was taken as the criterion. Unlike the TIP3P, the SPC water correctly shows the second solvation shell peak in the rdf [106].

The water molecule geometry differs for the models as well. While TIP3P adopts the experimental bond angle observed in the liquid phase (104.52°), SPC water has the ideal tetrahedral shape (109.47°). The computational time is saved by the fact that the HO bond lengths as well as the HOH angle are kept rigid; this allows using large time steps in the MD integration which in turn makes it possible to reach longer simulation times. If the water model

[11] SPC water model with the HOH angle 109.47°.

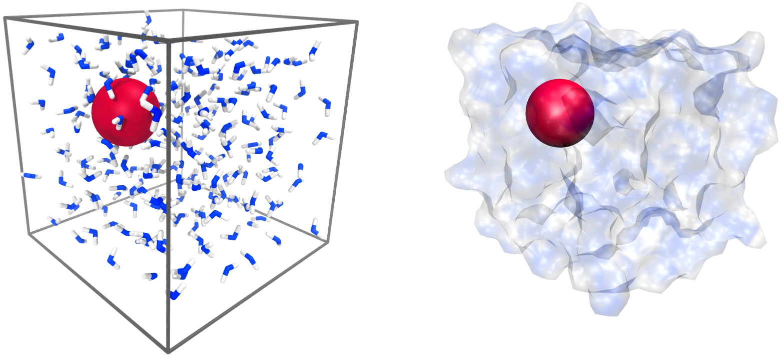


Figure 2.2: Explicit solvent model (left) describes each water molecule (in blue-white) separately, typically employing periodic boundary conditions. On the other hand, implicit solvent models treats water as structure-less continuum, periodic boundary conditions are not used.

is in a special format[12], a very efficient implementation of the nonbonded interaction of three-site water models can be used in the Gromacs program package [75].

The polarisation contribution is missing in the models completely. Thus the parametrisation to the experimental observables leads to the effective parameters inherently including the many-body effects. The demonstration of this, for example, is the magnitude of the dipole moment μ, which is typically too large for both three-site models ($\mu(\text{SPC}) = 2.27\ D$, $\mu(\text{TIP3P}) = 2.35\ D$) when compared with the gas-phase experimental value ($\mu(\text{EXP}) = 1.85\ D$ [107]).[13] Nonetheless, the rigid three-site models have been shown [109–111] to provide sufficiently accurate solvation features of biomolecular complexes, in spite of their extreme simplicity and the crude parametrisation. For other purposes though, other water models have been developed [112–114].

2.5.2 Implicit Solvent Models

The macroscopic view on water as the medium of high permittivity could have been the inspiration for another class of approximations, which neglects the internal microscopic structure of water at all. The simulation system then contains only the solute surrounded by a continuum of certain permittivity. It is clear that the number of particles is dramatically reduced, because only the solute atoms are present in the simulation, which leads to a decrease of computational demands once the continuum solvent contribution is calculated efficiently (Figure 2.5.2).

The fundamental quantity is the solvation free energy, which is the free en-

[12] It works for three-site models only, and the atoms of water must have certain properties and be in certain order.

[13] It is fair to say that the experimental liquid-state values is higher than all of the above mentioned ($\mu(\text{EXP}) = 2.95\ D$ [108]).

ergy needed for the transfer of a particle (atom, ion or molecule) from vacuum to the solvent. For aqueous environment, the solvation free energy is denoted as *hydration* free energy and can be decomposed into two terms (Equation 2.13)

$$\Delta G_{sol} = \Delta G_{pol} + \Delta G_{nonpol} \tag{2.13}$$

where the former term stands for the electrostatic/polar contribution and the latter term is the nonpolar contribution. The pioneering work goes back to Born and Onsager, who derived the solvation free energy of an ion and a dipole in a spherical cavity in water [115, 116].

The free energy of moving the spherical cavity with a charge q located in its centre from vacuum to the solvent of relative permittivity ε can be in atomic units expressed by Equation 2.14.

$$\Delta G = \left(\frac{1}{\varepsilon} - 1\right)\frac{q^2}{2a} \tag{2.14}$$

where a stands for the cavity radius. The Born model is well suited for molecular mechanics, where it is called the *generalised Born* (GB) model. The generalisation lies in modelling a set of charges rather than only a single one. Indeed, when a set of particles (*e. g.* a molecule) carrying the partial charge q_i and having the radius a_i is inserted into water, the polar part of the solvation free energy can be calculated by Equation 2.15

$$\Delta G_{pol} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_i\sum_j \frac{q_i q_j}{f\left(r_{ij}, a_i, a_j\right)} \tag{2.15}$$

where r_{ij} is the interparticle distance and a_i and a_j are *effective* Born radii. The function f has usually the form of Equation 2.16.

$$f\left(r_{ij}, a_i, a_j\right) = \sqrt{r_{ij}^2 + a_i a_j \mathrm{exp}\left\{-\frac{r_{ij}^2}{4b_i b_j}\right\}} \tag{2.16}$$

The function f is well behaved in the sense that it yields the Born formula (Equation 2.14) in the limit of $i = j$. The effective Born radii stand for the effective distance of the charge from the boundary of the molecule with the continuum solvent and their calculations represent the major issue in the implementation. It is not far from the truth to say that *many* of the implementations of GB model were proposed for all, MM, SQM and QM, methods [117–120]. They differ mostly in subtle details in the functional form for the effective Born radii determination and in the training set used for the parametrisation.

The effective Born radii apparently depend on the conformation of the molecule. Once the conformation of the molecule changes upon MD simulation, the effective Born radii must be recalculated. The GB model was implemented into the Gromacs package rather late, in version 4.5, which appeared in 2010. The reason for this lay in the very efficient explicit water implementation available, thus there was only a small motivation to proceed to GB. Also it was necessary to find a parallelisable algorithm to keep the entire Gromacs package fast, free and flexible [121].

3

Conformational Ensembles

This chapter focuses on the first of the two CADD issues particularly studied in this thesis – the conformational behaviour of molecules. In the next sections, three publications are briefly presented, highlighting some of their most important results. One of the publications is under review, two others are already published; their full texts are available as Appendices B, C and D.

3.1 Conformational Entropy

Biomolecules are large particles composed of many (thousands of) atoms, commonly dissolved in the aqueous solution inside cells; it is not surprising that they undergo conformational changes while thermally fluctuating and/or interacting with each other. The more a biomolecule fluctuates, the more entropy it exhibits, and it may be of interest to know how the entropy is changed upon a generalised chemical reaction.

Here the attribute *generalised*, means any reaction where no chemical bonds are broken or created. As an example it is possible to mention protein folding or noncovalent binding. As introduced in Section 1.5, conformational freedom is reflected by conformational (sometimes also referred to as configurational) entropy. The importance of conformational entropy could be seen in the fact that such a term completes the phenomenological free energy decomposition, so popular in the description of drug–target interactions.

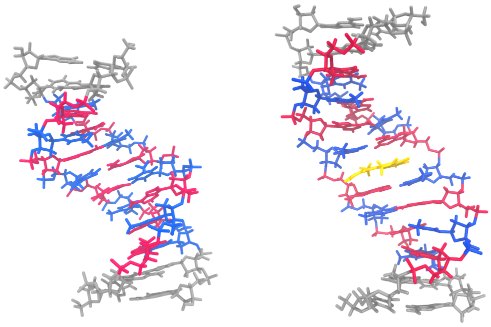


Figure 3.1: The structure of DNA without (left) and with (right) the intercalator (yellow). Guanine-cytosine steps are in grey, adenines in blue and thymines in red.

Karplus and Kushik suggested a method for the estimation of the conformational entropy of biomolecules from their covariance matrix in internal coordinates [40], which was further modified to allow using also the Cartesian coordinates [122, 123]. Using so-called quasi-harmonic approximation, one is able to calculate the absolute entropies of the biomolecules. Moreover, it seems to be possible, although not completely rigorously, to trace the origin of the entropic changes coming from various biomolecular fragments.

We investigated a drug–DNA complex by means of classical MD. The mode of binding of the drug – the anticancer agent ellipticine – is intercalation (Figure 3.1) [124, 125]. This binding motif is characterised by an increase of the distance between two consecutive base-pair steps and a possible local distortion of the sugar-phosphate backbone. There has been an abundance of medically active intercalators identified mostly in connection with cancer [126]. An important role of stacking interactions facilitated by dispersion energy have been emphasised for intercalators, computationally [127].

We investigated four adenine-thymine rich DNA sequences (Table 3.1), since there was evidence that an AT step may be slightly preferred as an intercalation site. It should be noted that the binding into another DNA-binding site – the minor groove – is also given some sequence preference [128]. However, the small molecule has much easier work when recognising the sequence, because the molecule usually spans more than 3 base-pair steps in the minor groove [129, 130].

Our interest was focused on the manifestation of DNA dynamics in terms of energy/entropy. The MD trajectory was analysed in such a manner that the atomic coordinates of the DNA were used for the construction of a mass-weighted covariance matrix, the elements of which are defined by Equation 3.1.

A	`5’-CGATAT(int)ATATCG-3’`
B	`5’-CGTTAT(int)ATAACG-3’`
C	`5’-CGTAAT(int)ATTACG-3’`
D	`5’-CGTATT(int)AATACG-3’`

Table 3.1: Four DNA AT-rich sequences with the position of intercalator labeled (int).

$$C_{ij} = \left\langle M_{ii}^{\frac{1}{2}} \left(x_i - \langle x_i \rangle\right) M_{jj}^{\frac{1}{2}} \left(x_j - \left\langle x_j \right\rangle\right) \right\rangle , \tag{3.1}$$

where x stands for the Cartesian coordinate, M is the element of mass matrix and the angle brackets stand for the ensemble average. The covariance matrix can be transformed into a set of frequencies which stand for the effective fluctuations of the solute atoms in water and which can be further processed to provide the absolute entropy (Equations 1 to 5, Appendix B). This analysis was performed separately for the apoDNA, free ellipticine and ellipticine–DNA complex to obtain the entropy change upon binding. In other words, it was possible to express the changes in the flexibility of the DNA double helix in terms of a thermodynamical contribution to the binding free energy.

When the binding free energy of ellipticine into DNA is about 6 kcal/mol (*i. e.* the complex is the preferred state over separated molecules) it has been shown that the configurational entropy contribution can be in the absolute magnitude several times larger and sequence-dependent. Previously, it was observed that the minor groove binding causes a decrease of DNA flexibility [131]; DNA becomes stiffer, which disfavours binding according to the entropy–enthalpy compensation mentioned in Section 1.5. Contrary to this result, it was found that the intercalation actually makes the DNA more flexible, which stabilises the drug–DNA complex.

The major entropic change was located in the DNA backbone, because the set of nucleobases was rather unaffected by the noncovalent binding. Since the sequence D disagreed with the general trends, the MD trajectories were further analysed and the BI/BII conformational change of the sugar-phosphate backbone was uncovered. The magnitude of this change was different for the sequence D when compared to the other sequences, which also produced a difference in conformational entropy changes.

Controversially, the accuracy of the MM description of the DNA backbone has been addressed [132, 133], and it appears that intensive research will have to be performed in this direction. Thus, the BI/BII conformational transition might be updated in future in the context of conformational entropies.

Finally, it became possible to estimate the binding orientation of ellipticine in the DNA. There had been some concerns about the orientation of the drug

[125, 134]. Under the assumption that the solvation thermodynamics are comparable between the DNA sequences as well as between the two binding orientations, we proposed the one with the pyrrole nitrogen oriented into the major groove (Figure 2, Appendix B) as the more probable.

As the final remark here, it must be pointed out that conformational entropy is often approximated by vibrational entropy, which reflects the changes in vibrational characteristics upon binding. The advantage of such an approximation is that no ensemble sampling is needed, since the vibrational frequencies can be calculated from the second derivatives of the energy. Consequently, the energy can be calculated for a single conformation, which puts us back to Section X and allows to continue fluently to the next section.

3.2 Conformations in Implicit Solvation

As introduced in Section 1.4.1, the single-conformation approximation is a widely used method to reduce computer demands. For implicit solvation, it is the only way as the implicit solvation free energy is the function of atomic coordinates. That means that there is a direct association between the molecular geometry and the value of solvation free energy. This arises from the manner in which the effective Born radii are calculated, when the shape of the molecular cavity is needed and anytime the shape is changed, the effective Born radii need to be recalculated.

3.2.1 HIV-1 Inhibitors

In the first study, we investigated a series of nine human immunodeficiency virus type 1 (HIV-1) protease inhibitors. The HIV is a retrovirus which infects the human T-lymphocytes, eventually causing the failure of the human immune system [135, 136]. The virus cycle requires the produced protein to be processed through a protease enzyme, which thus draws attention as a possible target for anti-HIV cure.

The HIV-1 protease is an enzyme composed of two monomeric units [137, 138], each of which contains a flexible flap, creating an active site for the natural binder – the immature protein. The approved anti-HIV-1 protease agents are rather large molecules (the smallest one has 70 atoms) trying to mimic the peptidic character, which makes them rather flexible. The flexibility is in fact a natural feature of both the protease and its inhibitors.

A new SQM based scoring function has been recently introduced and tested in our laboratory [53]. The method heavily employs single-conformation approximation, which is necessary due to the use of the SQM level of energy calculations. It has been revealed that the binding free energy calculated as the

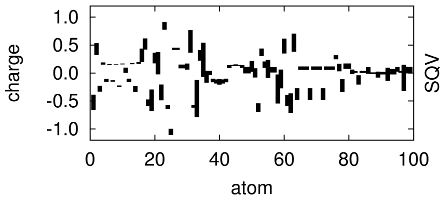


Figure 3.2: Variations of the RESP charges for 10 conformations of the HIV-1 protease inhibitor saquinavir.

SQM score has two important contributions large in magnitude yet with opposite signs; the interaction energy favours binding, but it is from a large part compensated for by the energetic penalty arising from ligand desolvation.

Our goal was to estimate the error brought by single-conformation approximation with respect to solvation free energies. For such large molecules as HIV-1 inhibitors, no experimental solvation free energies were available; therefore, we focused on the description of the distributions of the calculated values.

The conformational spaces of the inhibitors were extensively sampled by classical MD (see Section 2.1). For such calculations, the atomic partial charges were calculated employing the RESP technique (see Section 2.4). The RESP charges are conformation-dependent but it is not completely clear which conformation should be used for their determination. Thus we performed our investigations in several steps: i) pre-sampling, ii) sampling and iii) solvation free energy calculations themselves.

The pre-sampling aimed at generating a small number of conformations for which the charges were supposed to be determined. The simulations were performed at 700 K to ensure that the energetic barriers would be overcome. Ten conformations from the pre-sampling were used for the re-evaluation of the charges with which the production trajectories were generated. The simulations yielded 1600 conformers per inhibitor, and these were the subject of implicit solvation free energy calculations employing the SMD model by Marenich *et al.* [139] with the PM6-DH2 SQM level of electronic energy calculations [54, 55].

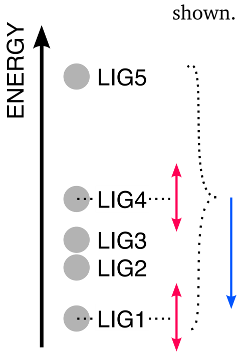
The resulting distributions of the hydration free energies were compared with those calculated on a single conformation and with the mean values of a limited number of conformations (*i. e.* 50 conformers arising from SQM dynamics). The distributions provided by the simulations of RESP charges, were compared leading to a surprising result. Despite the fact that the charges varied notably for the conformations (Figure 3.2.1), the mean values and standard deviations of the resulting distributions of the solvation free energies were rather similar.

This could suggest that the conformational dependence of implicit solvation energies is ambiguous; the values are definitely dependent on the conformation, but the subtle differences between the conformational ensembles of different RESP charges are somewhat hindered by the implicit solvation energy calculations.

For the set of nine HIV-1 inhibitors, the conformational energies E_{conf} were calculated as the sum of SMD solvation free energy and PM6-DH2 electronic energy. Complete set of the distributions is shown in Figures S2, S3 and S4 in Appendix C.

The mean deviation between the single-conformation E_{conf} and the distribution averages (*i. e.* the multi-conformation E_{conf}) was 1.5 ± 3.5 kcal/mol. These two numbers need a detailed explanation. The mean deviation of 1.5 kcal/mol says that the values of the single-conformation E_{conf} were on average 1.5 kcal/mol more positive than the mean values of the 1600 conformers. The error estimate of 3.5 kcal/mol stands for the standard deviation between the single-conformation E_{conf} and the multi-conformation E_{conf}. Therefore, once we are interested in the relative order of the inhibitors, the latter number is of higher importance, because it shows what error in the relative order of the compounds can be expected upon the single-conformation approximation.

Finally, we have studied the deviations between a small ensemble of conformations and a large truly multi-conformation approach. When the average over 50 conformers of the SQM sampling was taken instead of the single-conformation value, the error in the relative order decreased from 3.5 to 2.7 kcal/mol. The values were, however, shifted by −5.9 kcal/mol from the multi-conformational averages as compared to +1.5 kcal/mol for the single-conformation approach.[14]



If one considers the typical experimental binding free energies of such inhibitors to HIV-1 protease, which are between −15 and −8 kcal/mol, the error brought by the single-conformation approach, namely of about 3 kcal/mol, is critical. However, the range of the *calculated* single-conformation binding free energies of the HIV-1 protease complexes studied by Fanfrlík *et al.* [53] was between −15 and +30 kcal/mol. According to these authors, the deviation of about 3 kcal/mol presented by our results is rather convenient.

The largest deficiency in this first study on implicit solvents was the lack of relevant experimental data which motivated us for the extension of our calculations towards water-octanol partition coefficients.

[14] The scheme shows the situation of five ligands. In blue there is absolute shift of the entire set (*i. e.* −5.9 kcal/mol) whereas in red, the relative shifts of each of the ligand (*i. e.* 2.7 kcal/mol) are shown.

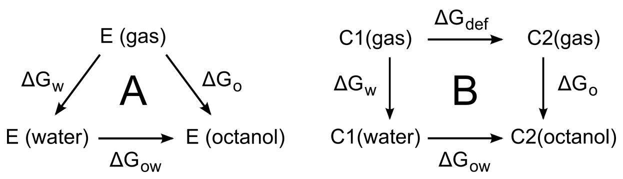


Figure 3.3: Thermodynamics cycles used for derivation of the transfer free energy from solvation free energies.

3.2.2 Water–Octanol Transfer Free Energies

The second study on implicit solvation extends our previous efforts by explicitly considering experimental data. We compiled a set of 20 approved drugs with known water-octanol partition coefficients which were available also for some of the HIV-1 protease inhibitors. Using water-octanol partition coefficients instead of hydration free energies, in which we would be interested much more, was a compromise of our part. For such large molecules as HIV-1 inhibitors, the hydration free energies are not available.

The water-octanol partition coefficient is a measure of the hydrophilic/hydrophobic properties of a compound. In drug development, it is highly appreciated to know the behaviour of a drug on the membrane or in protein environment, both quite hydrophobic. The water-octanol partition coefficient is directly related to water-octanol transfer free energy, *i. e.* the free energy change associated with the transfer of a compound from water to water-saturated octan-1-ol. According to the thermodynamic cycle depicted in Figure 3.2.2, the transfer free energy ΔG_{ow} can be calculated by Equation 3.2.

$$\Delta G_{ow} = \Delta G_o - \Delta G_w \tag{3.2}$$

When, however, two distinct conformations C1 and C2 are present in the phases, the thermodynamic cycle should be modified as shown in Figure 3.2.2. This produces another contribution, which stands for the free energy of the deformation of the two conformations. This seems to be a way to explore not only the conformational treatment in conjunction with implicit solvation, but, since the experiment data provide a well suited reference, also the performance of various implicit solvent models.

In a similar way as described in the previous section, we generated a set of conformers for which we subsequently calculated the implicit solvation free energies. First, the classical MD simulations in water and water-saturated octanol were performed, yielding 100 snapshots. Next, the implicit solvation free energy was calculated for each of the snapshots. Several popular implicit solvent models were tested, both based on MM charges [119, 120, 140, 141] and those employing QM electronic densities: the Conductor-like Screening Model for Real Solvents (COSMO-RS) [142], the SMD [139] and Miertus, Scrocco and Tomasi (MST) [143] models. Unlike in the previous study of free energy distributions, here we calculated only the mean values and standard deviations of the series.

Since it is not straightforward to process the ensemble averages of the solvation free energies, we suggested several *estimators* of the transfer free energies defined by Equations 3.3 to 3.7 and compared them with the experimental data directly.

$$G_0 = G_o\left(\mathrm{Xray}\right) - G_w\left(\mathrm{Xray}\right) \tag{3.3}$$

$$G_1 = \left\langle G_o \right\rangle_o - \left\langle G_w \right\rangle_w \tag{3.4}$$

$$G_2 = \left\langle G_o \right\rangle_o - \left\langle G_w \right\rangle_w + \left\langle E_{def} \right\rangle \tag{3.5}$$

$$G_3 = \left\langle G_o - G_w \right\rangle_w \tag{3.6}$$

$$G_4 = \left\langle G_o - G_w \right\rangle_o \tag{3.7}$$

Using Equation 3.3, the single-conformation transfer free energies were calculated (G_0) based mostly on experimental X-ray structures. The simple difference of the ensemble averages illustrated by the angle brackets $\langle\ \rangle$ was calculated by Equation 3.4 (G_1). Next, the thermodynamic cycle in Figure **??** was exploited by Equation 3.5. The deformation was approximated by Equation 3.8

$$\left\langle E_{def} \right\rangle = \langle E \rangle_o - \langle E \rangle_w \tag{3.8}$$

where E is the vacuum electronic energy of the compounds. In fact, the G_2 estimator is the difference between the *conformational* energies as defined in Section 3.2.1. Equations 3.6 and 3.7 describe the typical approach employed in many CADD studies including the Ref. [53], where it is assumed that the conformational ensembles are identical in both phases, here water (G_3) or octanol (G_4).

Some interesting results have been obtained from an analysis of MD simulations. For each of the molecules, the number of conformational families was evaluated[15] and put into relation with the number of rotatable bonds. The rotatable-bond concept is a way to estimate molecular flexibility from the structural formula of a compound [144]. Usually, the higher the number of sp^3–sp^3 or sp^3–sp^2 hybridised atom bonds is, the more flexible a compound is claimed to be. Our simulations revealed that not the number but also the kind of rotatable bonds should be considered. For example, we showed that although two molecules have six rotatable bonds each, they notably differ in the flexibility reflected by the number of conformational families.

[15] A conformation belongs to the family if its root-mean-square deviation with respect to any member of the family is lower than 1 Å

The set of 20 drugs was divided into two subsets according to the number of conformational families in water. The *rigid* compounds were those with only one conformational family, with the others being *flexible*.

The performance on the *rigid* compounds was satisfactory across all of the implicit solvent models. There was good agreement even for the single-conformation-based estimates in terms of the correlation coefficient (R^2) and root-mean-square error ($RMSE$). Surprisingly, no improvement was observed when conformational sampling was involved. COSMO-RS was identified as the overall best performing implicit solvent model for the *rigid* subset. COSMO-RS is an exception among the implicit solvent models tested: unlike the other solvent models, COSMO-RS already includes some technique to cover conformational flexibility. Hence, the transfer free energies were not calculated according to Equation 3.3 to 3.7, but rather as described in Appendix D and Refs [142, 145].

The flexible molecules were tackled much worse; when the single-conformation approach was utilised, hardly any implicit solvent model used was able to reach a higher correlation than 0.2 except for SMD ($R^2 = 0.42$). The conformational ensembles improved the results only slightly. COSMO-RS ($R^2 = 0.42$) and SMD with G_1 estimator ($R^2 = 0.66$) were presented as the most promising (see Appendix D, Figures S2 to S6).

Concerning the estimators (Equations 3.3 to 3.7), there were only minor differences found between them. It was also surprising that the G_2 estimator, virtually the most physical one, provided worse results than the others in terms of R^2. The G_2 estimator covers the deformation contribution, which seems to be the source of problems. Figure 15 depicts the Gaussian probability density functions corresponding to the mean values and standard deviations of the series of transfer free energies.

It appears that when the deformation is included, the distribution is much wider, which can lead to a blurred relative order of the molecules. The G_3 and G_4 estimators have the narrowest distributions, which in turn leads to a rather good correlation coefficient. It must be noted that G_2 showed lower a $RMSE$ than G_1 for many implicit solvation models. Most likely, the level at which the deformation energy (or internal electronic energy) is calculated, is too sensitive to the conformation. If there is any error cancellation between the solvation free energy in octanol and hydration free energy, it can be disturbed by the deformation contribution.

The error cancellation was further exploited by a separate comparison of the hydration free energies. They were found to be quite similar for different implicit models. A typical correlation coefficient between a solvent model and the SMD model (chosen as a reference arbitrarily) was higher than 0.8 (Appendix D, Figure 5). It was concluded that the differences between the implicit solvation methods used for the transfer free energies lie in the various accu-

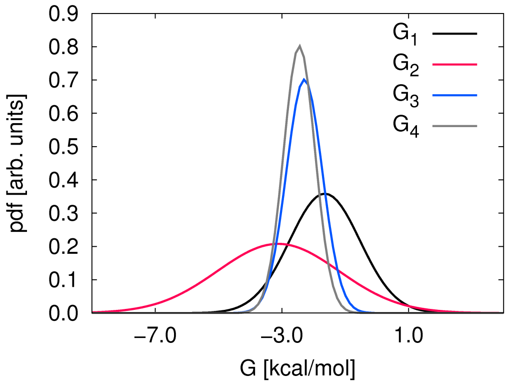


Figure 3.4: Gaussian probability density functions (pdfs) representing the mean values and standard deviations of transfer free energies as calculated according to Equations 3.4 to 3.7. The results of atropine are presented.

racy of the octanol solvent and/or in the various extent of error cancellation between the octanol and water phases.

In summary, the situation with the ligand flexibility is quite optimistic. The error estimate brought by the single-conformation approach of 2.7 kcal/mol, presented in the previous paragraphs, was related to HIV-1 inhibitors which are among the large ones. The error can be expected to be lower for less flexible ligands. The second study also showed that for rigid ligands the single-conformation approach is not a critical approximation; at least there exist some conformations which provide good agreement with experiment. In the second study, this role was played by the conformations observed in crystals. It remains a question what conformations should be used for *de novo* ligands, for instance. For flexible ligands, it was shown that we may expect problems. Our approaches do not provide the final answers but rather point to the direction in which it could be prospective to go.

4

Noncovalent Interactions Involving Halogens

This chapter presents the second part of the results related to CADD – the importance of well-balanced energy calculations. The chapter summarises four publications on noncovalent interactions involving halogen atoms, three research articles and one popular scientific contribution. In two of the publications dealing with the molecular mechanical description of halogen bonds, I am the first author. In the publication by Trnka *et al.*, I participated in the discussion and interpretation of the calculations as well as in the manuscript preparation.

With the project based on the publication in the Journal of Chemical Theory and Computation I was awarded the Jean-Marie Lehn award in Chemistry by the French Embassy in Prague and the Sanofi company.

4.1 Halogen Bonding

Besides the ensemble sampling, the role of accurate energy calculations for free energy estimations was emphasised in Section 1.3. Generally, we can seek problems among non-standard molecules and phenomena, and it remains to define what the standard means. The following paragraphs aim at the description of halogen bonding – a noncovalent interaction which has attracted much attention only recently.

A typical halogen bond (XB) is depicted in Figure 4.1. It is an attractive directional force between a halogen atom and a Lewis base, *i. e.* a chemical

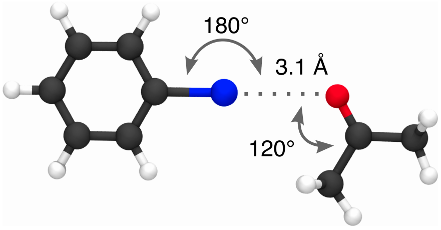


Figure 4.1: A typical halogen bond between bromobenzene and acetone with geometrical features is depicted.

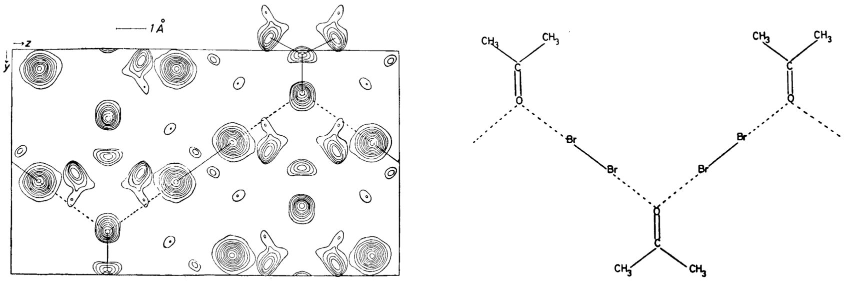


Figure 4.2: Crystal of acetone and bromine mixture with the O–Br distance of 2.82 Å. Left: electron density, right: proposed structure. Reproduced from [148].

group with a lone-electron-pair.[16] The early evidence about noncovalent complexes of halogens dates back to the 19th century [146] and then to the 1950s, where the halogen bonding motif was discovered in X-ray crystals (Figure 16) [147–149]. Its importance was soon recognised by Hassel [147]: "The O–Br distance is only 2.71 Å. This is the most striking feature of the whole structure as it indicates a very strong interaction between the bromine and oxygen atoms."

At that time, the nature of XB was slightly attributed to charge transfer [150, 151], and the label *halogen bond* had not begun to be used. Later on, it was accepted that XB is of electrostatic nature [152], which turned into a concept of so-called σ-hole, explaining many features of halogen bonding [153, 154]. Using QM calculations, it was revealed that there is a region of positive electrostatic potential (ESP) located on top of the halogen atom. Such a region (the σ-hole) facilitates the electrostatic attraction with the Lewis base (a negative charge) and also helps to assure the directionality of the XB. Finally, highly accurate QM calculations have shown that the dispersion contribution is also important [155], which is not surprising regarding the high polarisabilities of larger halogens (bromine, iodine).

Typical geometric features are described in Figure 4.1. An important pa-

[16] In 2012, the International Union of Pure and Applied Chemistry provided A Provisional Recommendation for the Definition of the Halogen Bond.

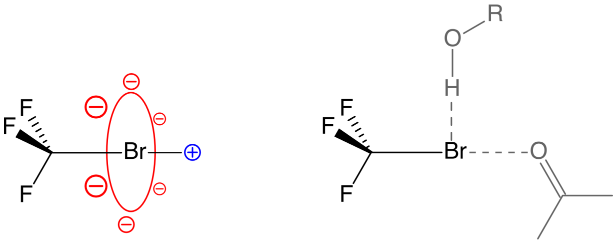


Figure 4.3: Left: scheme of charge distribution around bromine in bromotrifluoromethane. Right: two kinds of interactions may appear – with the negative ring around the bromine atom, with the positive σ-hole or both together.

rameter is the distance between the halogen and the Lewis base, which is shorter than the sum of the respective van der Waals radii. The *size* of the σ-hole increases with the increasing atomic number of the halogen and so does the strength of the XB. The stabilisation energy of a halogen-bonded complex can reach several kcal/mol. For instance, the stabilisation energy for trifluoroiodomethane–formaldehyde is 4.1 kcal/mol [156] which is comparable with the water dimer (stabilisation of 4.9 kcal/mol [157]). Some controversy has appeared around the halogen bonding of fluorines. It has been concluded that fluorine in fact does not create halogen bonds unless bound to a very electronegative atom or chemical groups, such as another fluorine or cyano group [158–160]. Hence, this is out of relevance for biological applications, where fluorines are usually bound to an aromatic cycle or sp^3-hybridised carbon, neither of which allows the fluorine to have the σ-hole.

Recently, the σ-hole concept has been extended also to the atoms of groups V and VI, which means there are so-called *pnicogen* and *chalcogen* bonds, respectively [161]. In the following sections, some applications of halogen bonding for both the crystalline phase and biological drug–target complexes are discussed.

4.2 Dihalogen Bonding in Crystals

Halogen bonding has found a distinguished place in crystal engineering [162–165], with which it has been connected since its discovery [147].

The electrostatic potential around halogen atoms exhibits two features which predetermine the halogen to participate in different noncovalent interactions. First, there is a positive σ-hole in the direction of the elongated C–X bond (where X stands for the halogen) (see Figure 4.2). And second, the region of positive ESP is surrounded by a negative ESP forming a ring shape.

It has already been mentioned that halogen bond is, from a large part, an electrostatic interaction with the σ-hole. A valency is still available to create a noncovalent interaction by the electrostatic attraction of the negative ring with a Lewis acid, *i. e.* an atom or chemical group with a positive ESP. An example of such an interacting partner might be hydrogen, which can create, and actually does, a hydrogen bond with a well-oriented halogen atom (Figure 4.2). Such a kind of interaction has been observed in both protein-ligand complexes [166] and crystalline materials [162], and it may be interesting that the hydrogen and halogen bonds can be created simultaneously by the same halogen atom.

A close inspection of the charge distribution around halogen atoms may suggest an existence of a halogen–halogen noncovalent interaction, where the σ-hole on one atom interacts with the negative ring located on the other atom.[161] This interaction, which is called a *dihalogen bond*, has been found in crystal phase [164], and its stability has been proved also by QM calculations [167]. Moreover, the same study showed that the strength of halogen and dihalogen bonds in model complexes is comparable [167].

By means of Symmetry-Adapted Perturbation Theory (SAPT) [168] and density functional theory (DFT) augmented by an empirical dispersion correction term (-D3) [169], we investigated the importance of dihalogen bonding in crystals of hexahalogenbenzenes, namely hexafluorobenzene (C_6F_6) [170], hexachlorobenzene (C_6Cl_6) [164] and hexabromobenzene (C_6Br_6) [164]. For comparison, the crystal of benzene (C_6H_6)[171] was included in our considerations as well. The study was conducted through an analysis of pair interactions within the crystals. A central molecule was chosen arbitrarily and its interactions with the nearest neighbours were studied.

The structural motifs observed in the crystals were characterised in terms of interaction energy. In the case of the benzene crystal,[17] there are mostly T-shape pairs, which are those competing in the gas phase along with a parallel-displaced (PD) arrangement; there is still some debate about which of them is more energetically stable [172, 173]. The hexachloro- and hexabromobenzene crystals are very similar to each other[18] but different from the benzene crystal. The most stable pairs are in the PD arrangement which are much more stable than the most stable benzene pair. Further, an important interaction motif found in their crystals is the dihalogen bond. For some pairs though, even two dihalogen bonds are created between the monomers (Appendix E, Figure 3). This motif has not been observed in the crystal of hexafluorobenzene[19] which is in accordance with the lack of the σ-hole (Figure 19).

The interaction energies (IEs) were calculated for each pair and also for the complex of the reference molecule and the group of all its neighbours (referred to as the *total interaction energy* (TIE)). An attempt to correlate the experimental values of sublimation energies with the QM calculations was made. It was

[17] Orthorhombic symmetry, space group *Pbca*.

[18] Monoclinic symmetry, space group $P2_1/n$.

[19] Monoclinic symmetry, space group $P2_1/n$.

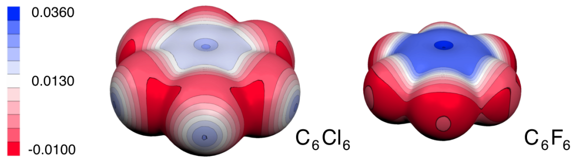


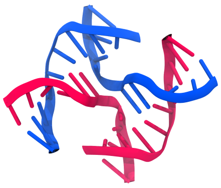
Figure 4.4: Electrostatic potentials of hexafluoro- and hexachlorobenzene mapped on a surface of electron isodensity of 0.001 e/Bohr^3. The scale is in a. u. There is only negative ESP on the halogens of hexafluorobenzene contrary to the positive σ-hole on hexachlorobenzene.

justified that the TIE increases in the order of $C_6H_6 < C_6F_6 < C_6Cl_6 < C_6Br_6$ and this increase is proportional to the increase in the sublimation energy.

A valuable conclusion was that the dihalogen bonding is not essential for the sublimation energies observed. The IE of about 2 kcal/mol for C_6Cl_6 dihalogen-bonded complexes was of a similar magnitude like the electrostatic (non-dihalogen bonding) interactions in C_6F_6, therefore, this interaction cannot be responsible for the increase of the sublimation energies. On the other hand, a SAPT analysis proved that the difference in the dispersion interaction can explain the observed sublimation data. Further, there were only small differences identified in the IE for single dihalogen-bonded pairs and doubly dihalogen-bonded pairs.

Finally, it should be mentioned that the dihalogen bond may be more important for crystal design (in a sense of solid-phase or liquid-crystal chemistry) than for biological applications. If one realises that the XB helps to specify the drug–target contact better, then the dihalogen bond would have to have an interacting halogen atom located on the target, which is not very *nature-like*. What can be a problem for a living cell, may be overcome in *in vitro* experiments, which are however only rarely utilised yet. An example is the work of Hays *et al.*, who incorporated 5-bromouridine into Holliday junction DNA [174][20] the effect of which was a conformational change facilitated by an XB [176]. The utilisation of a dihalogen bond definitely suffers from the lack of halogenated targets, which can be changed, for instance, by brominated nucleobases.

[20] Peculiar four stranded piece of DNA. Figure based on X-ray structure [175].



4.3 Halogen Bonding in Molecular Mechanics

The most significant stimuli to incorporate a halogen-bonding correction into molecular mechanics were two: i) the ever-increasing importance of halogen bonding in drug development and ii) the complete failure of classical force fields to describe halogen bonding. Both are exploited below.

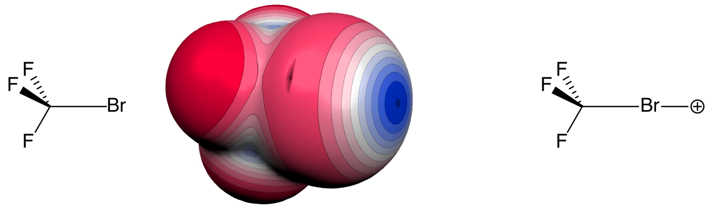


Figure 4.5: Left and middle: Structural formula of bromo-trifluoromethane with the ESP mapped on a surface of electron isodensity of 0.001 e/Bohr^3. Positive ESP is in blue, negative in red. Right: The scheme of explicit σ-hole.

4.3.1 Drug–Target Interactions

A large portion of drugs, either on the market or in the development, contain some halogens. Medicinal chemists have utilised halogenation as a modification of a chemical structure for many years, mostly for non-specific effects it could have had. Halogen atoms facilitate the oral drug absorption, improve the crossing of the blood-brain barrier,[177] which is all related to their higher hydrophobicity. Further, halogens became used to fill hydrophobic cavities once the geometry of binding site was known.

In the last few decades however, the use of halogens was accelerated thanks to their recognised ability to create specific, directional noncovalent interactions, which can serve as the lock pins as depicted in Figure 1.2 and thus improve the affinity of a drug. The halogen bonding in drug–target complexes has been thoroughly reviewed [166, 178–181] and it really seems to be beneficial. For example, a halogen bond was found in the complex of aldose reductase (the enzyme related to *diabetes mellitus*) [182], transthyretin (the transporter of iodinated thyroid hormones) [183], or in protein kinase CK2, related to cancer [184].

4.3.2 Explicit σ-Hole

In MM, the atom-centred point-charge treatment of electrostatic interactions cannot describe the anisotropy of the charge distribution around halogens. The lack of σ-hole in MM precludes a faithful description of halogen bonding, and it will be shown below how dramatic such a failure is.

Our attempt was to design a model for halogen bonding, simple enough to be used in MM with prospects for CADD. In the model, which is called *explicit σ-hole* (ESH) here, the region of positive electrostatic potential was approximated by a massless pseudo-atom (ghost atom) (Figure 4.3.2). Such an approach was previously used by Dixon and Kollman [185] for modelling electron lone-pairs on *e. g.* an oxygen atom, where they developed their initial

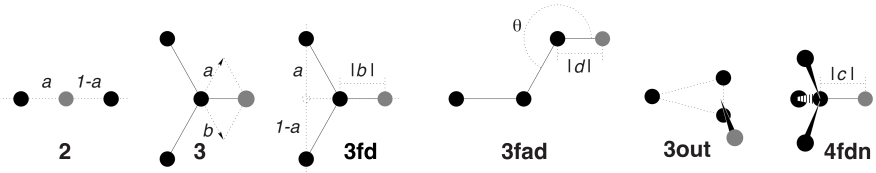


Figure 4.6: Virtual sites as defined in Gromacs [75]. Black circles are real atoms from which the virtual sites (grey circles) are constructed. Reproduced from [67].

efforts [92]. Actually, using off-centred point charges goes back to explicit water models, where the charge is shifted side of the oxygen [106, 186].

In fact, the ESH model is very similar to other σ-holemodels developed simultaneously in other laboratories [187–189]. This further emphasises the importance of a MM description of halogen bonds. The models differ in details: in the mass of the sigma-hole-mimicking atom and in the number and type of parameters.

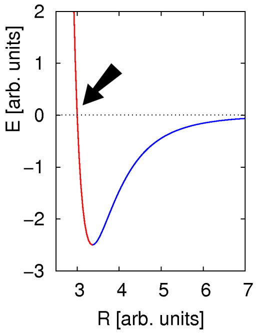
The implementation of ESH makes use of a *virtual-site* concept (Figure 4.3.2) available in the Gromacs program package [75]. The massless virtual site is constructed from the real atomic positions (*e. g.* from the particles which have mass) and it can carry both charge and LJ parameters. Since the zero mass is difficult to tackle by the equations of motion, the force acting on the virtual site must be redistributed back to the real atoms [67]. Now the questions regarding the ESH were: i) where to place it and ii) what charge to use.

4.3.3 The Models

Bearing in mind the popularity of the GAFF force field for the description of small molecules in simulations of drug–target complexes, we tried to design a σ-hole model GAFF-compatible in terms of partial charges and LJ parameters. Another condition was to use as few parameters as possible. It became clear that the ESH must be located within the van der Waals radius or, more precisely, in such a region which is well covered by the repulsive part of the LJ potential on the halogen.[21] Otherwise, some (large) instabilities should be expected in the simulations employing ESH.

Three ESH-charge models were suggested differing in their complexity and computer demands. The most simple one, called nF (*no fit*), has two parameters – the ESH charge and its distance from the respective halogen – and does not employ any QM calculations. The prerequisite for its use is the previous knowledge of the partial charges of all atoms within a molecule. Then the ESH charge is subtracted from the respective halogen atom to conserve the net molecular charge. The other atomic charges are kept intact.

[21] Lennard-Jones potential with the repulsion part in red. The arrow points to the value of σ LJ parameter.



		nF	rF	aF
range	charge	0.05–0.50 e	0.05–0.50 e	–
	distance	0.8–1.6 Å	0.8–1.6 Å	0.8 – 2.6 Å
step	charge	0.05 e	0.05 e	–
	distance	0.1 Å	0.1 Å	0.2 Å
charges needed		yes	no	no
ESP grid needed		no	yes	yes

Table 4.1: Summary of the three ESH charge models and the one- or two-dimensional parameter scan properties. It is to be noted that ESP grid has to be generated by an *ab initio* calculation if needed.

The second model, called rF (*rest fit*), employs two-stage RESP fitting as introduced in Section 2.4, which is fully applicable in conjunction with GAFF. It also has two parameters – again the ESH charge and the ESH–X distance. An additional fitting position with a fixed charge is included in both RESP-fitting stages. A denser ESP grid than the standard one is used [98], which is to improve the statistics of the complicated ESH shape around the halogen atom.

Finally the third model, called aF (*all fit*), is identical with rF but contains only one parameter – the ESH–X distance. The ESH charge is the subject of the two-stage RESP fitting, hence such a model should be viewed as the most physical one. Clearly, both rF and aF models need the ESP grid points to be generated by a QM calculation. This makes them perhaps too complicated for large-scale virtual screenings.

All three models are summarised in Table 4.1. So far across all models, the position of ESH has been fixed with respect to the halogen and the next bound carbon atom, but this can be changed in future to allow the ESH to move *e. g.* on a spherical cap.

4.3.4 Parametrisation and Performance

A one-dimensional scan for the aF model and two-dimensional scans for the rF and nF models were performed, consequently covering the entire parameter spaces of the respective models. For each point of the parameter space, the dissociation curves of three XB complexes were calculated (see Figure 1, Appendix F) and compared with highly-accurate QM data.[22] Typical curves are shown in Figure 22.

Although empirical force fields are not designed for gas-phase calculations, a few conclusions could be drawn. Most importantly, the effect of ESH is dramatic. When the ESH is shifted by 0.5 Å from the halogen, the interac-

[22] CCSD(T) dissociation curves calculated at the complete basis set limit extrapolated from augmented double- and triple-ζ basis sets.

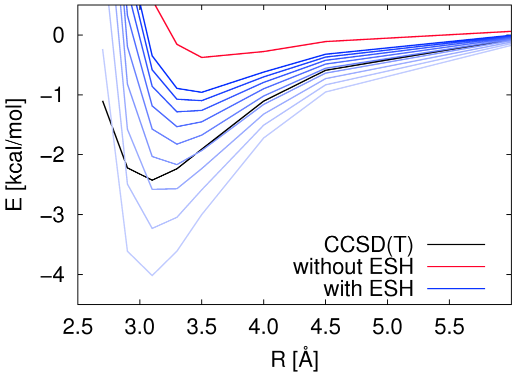


Figure 4.7: Dissociation curves of bromobenzene–acetone complex. CCSD(T) black curves stands for highly accurate QM calculations. The lighter the blue line is, the higher was the ESH-halogen distance. Calculated employing the rF charge model with the charge of 0.2 e.

tion energy at the optimal intermolecular distance can be changed by about 0.5 kcal/mol, which is as much as 20 % of the total interaction energy, depending on the complex. Furthermore, as mentioned before, the RESP charges are overpolarised; this makes them compatible with TIP3P and SPC water models, which both suffer from overestimated dipole moments. Thus one would expect an overstabilisation of gas-phase complexes described at the MM level with the RESP HF/6-31g* charges as compared to more accurate QM data. In the case of the halogen-bonded complex of bromobenzene and acetone, there is hardly any stabilisation (0.2 kcal/mol) when the standard GAFF is employed. Desirably, there exist some points in the parameter space which can lead to well-stabilised complexes.

The key quality of interest is, of course, the electrostatic potential. It has been shown that the standard GAFF ESP was qualitatively wrong while the inclusion of ESH made it qualitatively correct; if the ESH parameters were adjusted accordingly, the generated ESP was accurate even quantitatively (Figure 22 and Figure 4 in Appendix F). The deviation from the QM ESP was plotted to identify what ESH position is preferred (Figure **??**).

The size of the σ-hole depends on the chemical environment and is in some sense tunable [190]. For instance, the introduction of fluorines into the meta-positions on the aromatic ring increases the ESP on the reference halogen, which makes the halogen bond eventually stronger. This feature is actually well reflected in ESH models, when the ESPs of bromobenzene and 3,5-diflurobromobenzene are compared (Figure 4, Appendix F). The performance of the aF and rF models is slightly better than that of the nF model. Although not tested,

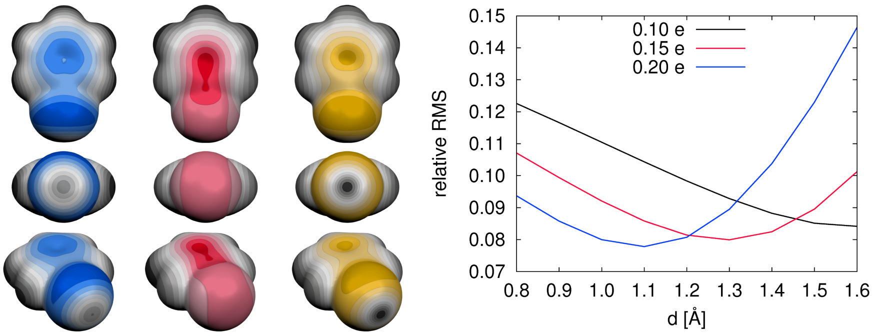


Figure 4.8: Right surfaces: Bromobenzene ESP projections on the electron isodensity surfaces of 1 e/Bohr^3; left blue: the QM ESP; middle red: ESP without explicit σ-holeand right yellow: ESP with explicit σ-hole. Left plot: RESP fitting performance expressed as relative root-mean-square (RMS) error, depending on the ESH–X distance (d). Three charges of ESH are shown for the bromobenzene case. For comparison, the relative RMS for the fitting without ESH was 0.16.

it is likely that because of the shape of the electrostatic potential, the ESH model may be suitable also for the description of dihalogen bonds.

Further, realistic systems of CADD – drug–target complexes – were investigated in closer detail. A series of known X-ray structures of protein kinase CK2 with tetrabrominated inhibitors [191–193], which were known to create one or two halogen bonds, was gathered. With the series, we performed geometry optimisations employing an implicit solvent model to determine the effect of ESH on the drug–target geometry. The problem complexity was decreased by making irrelevant parts of the protein rigid.

It was demonstrated that the ESH approach can ensure the halogen-bonding pattern, unlike the standard GAFF without ESH. The parameter spaces were scanned in the same manner as in the case of dissociation curves. The quality of the MM-optimised geometry was estimated on the basis of the root-mean-square deviation (RMSD) of the heavy atoms of such parts of the complex which were free to move. As the reference the X-ray experimental structures were taken.

The geometries obtained with ESH were considerably better than those without ESH. In accordance with previous results, it was shown that the Amber force field fails in the description of halogen bonding in CK2 [194]. To recognise the best parameters for ESH, their RMSD dependence was plotted (Figure 22). Generally, there were low-RMSD areas in the two dimensional plots which pointed to the suitable parameters. In the case of the one-parameter aF model, the best performing distance was too high (more than the van der Waals radius of the halogen).

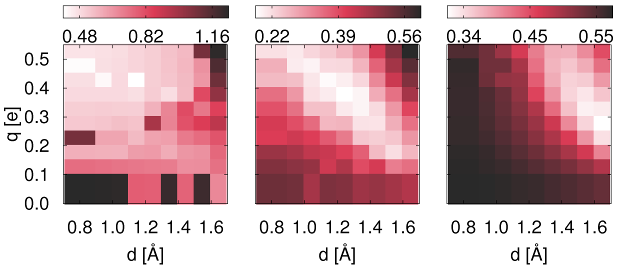


Figure 4.9: Dependence of root-mean-square deviation (RMSD) on the ESH charge (q) and EXH–X distance (d). The RMSD matrices of three CK2 complexes were calculated for the heavy atoms of ligands with respect to the experimental geometries.

An important result concerned the RMSD sensitivity to the ESH parameters. It was demonstrated that the performance on CK2 is rather insensitive, which means that once the ESH is included in the calculations, it is highly probable that the results (in terms of RMSD) will be better than without ESH. Finally, the recommendations of the ESH parameters were provided: a charge of about 0.2 e located at a distance of 1.5 Å from bromine seems to be a suitable choice. However, the choice is ambiguous, as presented in Figure 22.

4.3.5 Improved Docking

In CADD, a wide area of research focuses on the prediction of the drug–target geometry, *i. e.* what the position of the drug in the target active site looks like. The prediction is done by so-called *docking* algorithms, which aim to place the drug into a pre-defined active site in the best-matching way. Numerous such algorithms have been developed [49, 195, 196] because of the need for the drug–target geometry for more advanced calculations.

The ESH nF model was used with the docking program suite DOCK6 [196] to predict the binding poses of 92 drug-enzyme complexes containing some halogens (chlorine, bromine or iodine). The set comprises 7 aldose reductase complexes, 32 cyclin-dependent kinase complexes, 16 protein kinase CK2 complexes and 37 HIV reverse transcriptase complexes complexes (Figure 4.3.5); 55 % of the ligands contained more than one possible halogen bond-donor. The docked poses were compared with the known X-ray structures and analysed.

The analysis revealed that the ESH can improve the ability to predict the native (*i. e.* experimentally observed) pose and that the number of halogen bonds, their length and the fidelity of the XB acceptors are also improved. Unlike without ESH, the predicted poses with ESH had shorter XBs with more relevant XB acceptors, both in better agreement with experiment. This was especially ap-

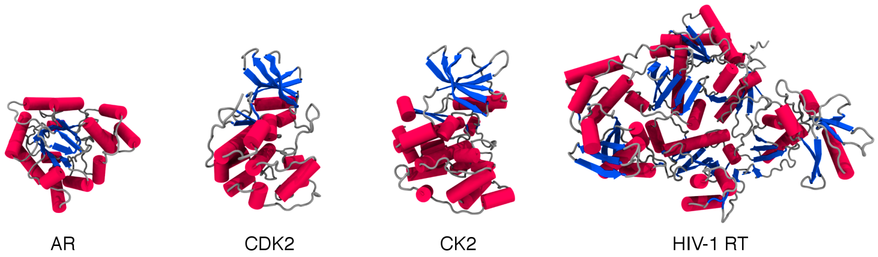


Figure 4.10: The protein targets used for the docking study: human aldose reductase (AR) [197], human cyclin-dependent kinase 2 (CDK2) [198], catalytic subunit of protein kinase 2 (CK2) [192] and HIV-1 reverse trasncriptase (HIV-1 RT) [199].

parent in the case of CK2, where more than one XB was established. It must be noted that not all of the complexes were halogen-bonded; large portion of the complexes (about 43 %) contained some halogen which was not involved in any XB. Not to create a XB where it should not be seems to be an appreciable feature of ESH.

5

Summary and Outlook

The thesis has presented some methodological advances of molecular modelling which can be utilised in computer-aided drug development. The introduction for seven original articles was given to provide the reader with a better insight and to highlight some consequences which are difficult to bring by rather narrowly focused research publications.

The advances have been divided into two groups: the investigation of the role of (bio)molecular conformations and the description of a specific kind of noncovalent interactions involving halogen atoms. Each group covers three research publication, the latter one also one popular scientific article. It has been emphasised that both aspects, the conformational treatment and accurate energy calculations, are essential for free energy calculations as a computational way of estimating the drug efficacy. In order to highlight the bases on which the publications were built, some details on the general molecular modelling techniques have been presented.

Two chapters focus on the results. First, the role of conformational entropy as a measure of flexibility changes was elucidated for the drug–DNA complex. Upon drug binding, the DNA was shown to become more flexible and the flexibility change was expressed as a thermodynamic quantity, being as much as 38 kcal/mol at 300 K. Next, the conformational sampling was investigated in the context of implicit solvent models. It was concluded that the rigid molecules are suitable for the single-conformation approach by the implicit solvent models; the flexible ones, such as HIV-1 protease inhibitors, should involve some conformational sampling. It was merely pointed out that even bigger problems may be expected for more flexible molecules, such as proteins, when the single-conformation approach is applied to them.

In the second chapter of the results, the halogen and dihalogen bonds were introduced as noncovalent interactions with a certain importance in the design of new materials and also in drug development. The molecular mechanical model for σ-hole, considerably improving the MM description of halogen bonds, was proposed. Such a model was presented in detail, and the overlap with drug development was emphasised. This was perhaps the most apparent in case of the improvement of a docking platform.

This is also one of the promising directions for the further expansion of the results. The methodological advances need to be used to show their real prospects. The projects which are being implemented in the laboratory of Prof. Hobza in cooperation with the Heidelberg Institute for Theoretical Studies and institutions in the Czech Republic involve the ESH model for the description of halogenated ligands. Particular points to answer could be the advanced free energy techniques (such as those based on the Zwanzig formula), where the model should be improved. The preliminary results show that it may not be so straightforward, even though a solid foundation has been established.

In the case of conformational sampling, the results are mostly utilised with the SQM scoring function used in our laboratory, and for a more careful interpretation of the outcoming results. As mentioned before, the treatment of flexible ligands has not found its ultimate solution and it remains a question, for instance, what (a small number of) conformations should represent the entire ensemble sufficiently.

Despite the fact that there is still a long way to new drugs, the results of the thesis may shorten the way at least to some of them.

Bibliography

[1] C. P. Adams and V. V. Brantner, “Estimating The Cost Of New Drug Development: Is It Really $802 Million?,” *Health Affairs*, vol. 25, pp. 420–428, Mar. 2006.

[2] I. M. Kapetanovic, “Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach,” *Chemico-Biological Interactions*, vol. 171, pp. 165–176, Jan. 2008.

[3] G. E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics*, vol. 38, pp. 114–117, Apr. 1965.

[4] P. A. M. Dirac, “Quantum Mechanics of Many-Electron Systems,” *Proceedings of the Royal Society of London. Series A*, vol. 123, pp. 714–733, Apr. 1929.

[5] D. C. Rees, M. Congreve, C. W. Murray, and R. Carr, “Fragment-based lead discovery,” *Nature Reviews Drug Discovery*, vol. 3, pp. 660–672, Aug. 2004.

[6] M. Congreve, C. W. Murray, and T. L. Blundell, “Keynote review: Structural biology and drug discovery,” *Drug Discovery Today*, vol. 10, pp. 895–907, July 2005.

[7] M. E. M. Noble, J. A. Endicott, and L. N. Johnson, “Protein Kinase Inhibitors: Insights into Drug Design from Structure,” *Science*, vol. 303, pp. 1800–1805, Mar. 2004.

[8] J. Adams, “The proteasome: a suitable antineoplastic target,” *Nature Reviews Cancer*, vol. 4, pp. 349–360, May 2004.

[9] L. H. Hurley, “DNA and its associated processes as targets for cancer therapy,” *Nature Reviews Cancer*, vol. 2, pp. 188–200, Mar. 2002.

[10] J. R. Thomas and P. J. Hergenrother, “Targeting RNA with Small Molecules,” *Chem. Rev.*, vol. 108, pp. 1171–1224, Mar. 2008.

[11] S. R. George, B. F. O’Dowd, and S. P. Lee, “G-Protein-coupled receptor oligomerization and its potential for drug discovery,” *Nature Reviews Drug Discovery*, vol. 1, pp. 808–820, Oct. 2002.

[12] M. C. Lagerström and H. B. Schiöth, “Structural diversity of G protein-coupled receptors and significance for drug discovery,” *Nature Reviews Drug Discovery*, vol. 7, pp. 339–357, Apr. 2008.

[13] G. Bkaily, L. Avedanian, and D. Jacques, “Nuclear membrane receptors and channels as targets for drug development in cardiovascular diseasesThis article is one of a selection of papers from the NATO Advanced Research Workshop on Translational Knowledge for Heart Health (published in part 1 of a 2-part Special Issue).,” *Canadian Journal of Physiology and Pharmacology*, vol. 87, pp. 108–119, Feb. 2009.

[14] G. J. Kaczorowski, O. B. McManus, B. T. Priest, and M. L. Garcia, “Ion Channels as Drug Targets: The Next GPCRs,” *The Journal of General Physiology*, vol. 131, pp. 399–405, May 2008.

[15] A. Merino, A. K. Bronowska, D. B. Jackson, and D. J. Cahill, “Drug profiling: knowing where it hits,” *Drug Discovery Today*, vol. 15, pp. 749–756, Sept. 2010.

[16] R. Bertz and Granneman, “Use of In Vitro and In Vivo Data to Estimate the Likelihood of Metabolic Pharmacokinetic Interactions,” *Clinical Pharmacokinetics*, vol. 32, no. 3, pp. 210–258, 1997.

[17] A. Rostami-Hodjegan and G. T. Tucker, “Simulation and prediction of in vivo drug metabolism in human populations from in vitro data,” *Nature Reviews Drug Discovery*, vol. 6, pp. 140–148, Feb. 2007.

[18] K. Ito, H. Kusuhara, and Y. Sugiyama, “Effects of Intestinal CYP3A4 and P-Glycoprotein on Oral Drug AbsorptionTheoretical Approach,” *Pharmaceutical Research*, vol. 16, no. 2, pp. 225–231, 1999.

[19] H. E. Selick, A. P. Beresford, and M. H. Tarbit, “The emerging importance of predictive ADME simulation in drug discovery,” *Drug Discovery Today*, vol. 7, pp. 109–116, Jan. 2002.

[20] M. Stahl and M. Rarey, “Detailed Analysis of Scoring Functions for Virtual Screening,” *J. Med. Chem.*, vol. 44, pp. 1035–1042, Mar. 2001.

[21] W. L. Jorgensen, “Efficient Drug Lead Discovery and Optimization,” *Acc. Chem. Res.*, vol. 42, pp. 724–733, Mar. 2009.

[22] R. S. DeWitte and E. I. Shakhnovich, “SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence,” *Journal of the American Chemical Society*, vol. 118, pp. 11733–11744, Jan. 1996.

[23] K. Müller-Dethlefs and P. Hobza, “Noncovalent Interactions: A Challenge for Experiment and Theory,” *Chem. Rev.*, vol. 100, pp. 143–168, Dec. 1999.

[24] K. E. Riley, M. Pitoňák, P. Jurečka, and P. Hobza, “Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories,” *Chem. Rev.*, vol. 110, pp. 5023–5063, May 2010.

[25] K. Morokuma, “Molecular Orbital Studies of Hydrogen Bonds. III. C=O···H–O Hydrogen Bond in $H_2CO \cdots H_2O$ and $H_2CO \cdots 2H_2O$,” *The Journal of Chemical Physics*, vol. 55, no. 3, pp. 1236–1244, 1971.

[26] P. A. Kollman and L. C. Allen, “Theory of the hydrogen bond,” *Chem. Rev.*, vol. 72, pp. 283–303, June 1972.

[27] M. Urban and P. Hobza, “Weak intermolecular interaction,” *Theoretica chimica acta*, vol. 36, no. 3, pp. 207–214, 1975.

[28] D. A. McQuarrie, *Statistical Mechanics*. University Science Books, June 2000.

[29] R. W. Zwanzig, "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases," *The Journal of Chemical Physics*, vol. 22, no. 8, pp. 1420–1426, 1954.

[30] C. Hansch and T. Fujita, "P–Analysis. A Method for the Correlation of Biological Activity and Chemical Structure," *J. Am. Chem. Soc.*, vol. 86, pp. 1616–1626, Apr. 1964.

[31] D. Rogers and A. J. Hopfinger, "Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships," *J. Chem. Inf. Comput. Sci.*, vol. 34, pp. 854–866, July 1994.

[32] M. Karelson, V. S. Lobanov, and A. R. Katritzky, "Quantum-Chemical Descriptors in QSAR/QSPR Studies," *Chem. Rev.*, vol. 96, pp. 1027–1044, Jan. 1996.

[33] M. Born and R. Oppenheimer, "Zur Quantentheorie der Molekeln," *Ann. Phys.*, vol. 389, no. 20, pp. 457–484, 1927.

[34] I. Muegge and Y. C. Martin, "A General and Fast Scoring Function for ProteinLigand Interactions: A Simplified Potential Approach," *J. Med. Chem.*, vol. 42, pp. 791–804, Feb. 1999.

[35] G. Schneider, "Virtual screening: an endless staircase?," *Nat Rev Drug Discov*, vol. 9, pp. 273–276, Apr. 2010.

[36] J. Åqvist, V. B. Luzhkov, and B. O. Brandsdal, "Ligand Binding Affinities from MD Simulations," *Acc. Chem. Res.*, vol. 35, pp. 358–365, Feb. 2002.

[37] A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz, and H. M. Al-Hashimi, "Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble," *Nature Chemical Biology*, vol. 7, pp. 553–559, June 2011.

[38] I. S. Ufimtsev and T. J. Martinez, "Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation," *J. Chem. Theory Comput.*, vol. 4, pp. 222–231, Jan. 2008.

[39] J. C. Phillips and J. E. Stone, "Probing biomolecular machines with graphics processors," *Commun. ACM*, vol. 52, pp. 34–41, Oct. 2009.

[40] M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," *Macromolecules*, vol. 14, pp. 325–332, Mar. 1981.

[41] H.-X. X. Zhou and M. K. Gilson, "Theory of free energy and entropy in noncovalent binding.," *Chemical reviews*, vol. 109, pp. 4092–4107, Sept. 2009.

[42] S.-R. Tzeng and C. G. Kalodimos, "Protein activity regulation by conformational entropy," *Nature*, vol. 488, pp. 236–240, Aug. 2012.

[43] D. Chandler, "Interfaces and the driving force of hydrophobic assembly," *Nature*, vol. 437, pp. 640–647, Sept. 2005.

[44] A. Godec and F. Merzel, "Physical Origin Underlying the Entropy Loss upon Hydrophobic Hydration," *J. Am. Chem. Soc.*, vol. 134, pp. 17574–17581, Sept. 2012.

[45] K. Sharp, "Entropy–enthalpy compensation: Fact or artifact?," *Protein Science*, vol. 10, pp. 661–667, Mar. 2001.

[46] E. B. Starikov and B. Nordén, "EnthalpyEntropy Compensation: A Phantom or Something Useful?," *J. Phys. Chem. B*, vol. 111, pp. 14431–14435, Nov. 2007.

[47] E. Freire, "Do enthalpy and entropy distinguish first in class from best in class?," *Drug Discovery Today*, vol. 13, pp. 869–874, Oct. 2008.

[48] H.-J. Böhm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure," *Journal of Computer-Aided Molecular Design*, vol. 8, no. 3, pp. 243–256, 1994.

[49] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," *Journal of Molecular Biology*, vol. 295, pp. 337–356, Jan. 2000.

[50] J. L. Paulsen and A. C. Anderson, "Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization," *J. Chem. Inf. Model.*, vol. 49, pp. 2813–2819, Dec. 2009.

[51] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case, "Continuum Solvent Studies of the Stability of DNA, RNA, and PhosphoramidateDNA Helices," *J. Am. Chem. Soc.*, vol. 120, pp. 9401–9409, Aug. 1998.

[52] M. Feig and C. L. Brooks, "Recent advances in the development and application of implicit solvent models in biomolecule simulations," *Current Opinion in Structural Biology*, vol. 14, pp. 217–224, Apr. 2004.

[53] J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka, and P. Hobza, "A Reliable Docking/Scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands," *J. Phys. Chem. B*, vol. 114, pp. 12666–12678, Sept. 2010.

[54] J. Řezáč, J. Fanfrlík, D. Salahub, and P. Hobza, "Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes," *J. Chem. Theory Comput.*, vol. 5, pp. 1749–1760, May 2009.

[55] M. Korth, M. Pitoňák, J. Řezáč, and P. Hobza, "A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods," *J. Chem. Theory Comput.*, vol. 6, pp. 344–352, Dec. 2009.

[56] K. Raha and K. M. Merz, "A Quantum Mechanics-Based Scoring Function: Study of Zinc Ion-Mediated Ligand Binding," *J. Am. Chem. Soc.*, vol. 126, pp. 1020–1021, Jan. 2004.

[57] K. Raha and K. M. Merz, "Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein-Ligand Complexes," *J. Med. Chem.*, vol. 48, pp. 4558–4575, June 2005.

[58] K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wollacott, L. M. Westerhoff, and K. M. Merz, "The role of quantum mechanics in structure-based drug design," *Drug Discovery Today*, vol. 12, pp. 725–731, Sept. 2007.

[59] P. Mikulskis, S. Genheden, K. Wichmann, and U. Ryde, "A semiempirical approach to ligand-binding affinities: Dependence on the Hamiltonian and corrections," *J. Comput. Chem.*, vol. 33, pp. 1179–1189, May 2012.

[60] H. S. Muddana and M. K. Gilson, "Calculation of HostGuest Binding Affinities Using a Quantum-Mechanical Energy Model," *J. Chem. Theory Comput.*, vol. 8, pp. 2023–2033, May 2012.

[61] P. S. Brahmkshatriya, P. Dobeš, J. Fanfrlík, J. Řezáč, K. Paruch, A. K. Bronowska, M. Lepšík, and P. Hobza, "Quantum Mechanical Scoring: Structural and Energetic Insights into Cyclin-Dependent Kinase 2 Inhibition by Pyrazolo[1,5-a]pyrimidines," *Current Computer-Aided Drug Design*, vol. 9, no. 1, pp. 118–129, 2013.

[62] B. J. Alder and T. E. Wainwright, “Studies in Molecular Dynamics. I. General Method,” *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 459–466, 1959.

[63] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*. Oxford science publications, Clarendon Press, Oxford, June 1989.

[64] A. Leach, *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall, 2 ed., Apr. 2001.

[65] C. Chipot and A. Pohorille, eds., *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer Verlag, 2007.

[66] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M. J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman, “Amber 11,” 2010.

[67] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen, “Gromacs User Manual version 4.5,” 2010.

[68] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, “A new force field for molecular mechanical simulation of nucleic acids and proteins,” *J. Am. Chem. Soc.*, vol. 106, pp. 765–784, Feb. 1984.

[69] W. L. Jorgensen and J. Tirado-Rives, “The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin,” *J. Am. Chem. Soc.*, vol. 110, pp. 1657–1666, Mar. 1988.

[70] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules,” *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, May 1995.

[71] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, “All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†,” *J. Phys. Chem. B*, vol. 102, pp. 3586–3616, Apr. 1998.

[72] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, “A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations,” *J. Comput. Chem.*, vol. 24, pp. 1999–2012, Dec. 2003.

[73] D.-W. Li and R. Brüschweiler, “NMR-Based Protein Potentials,” *Angewandte Chemie International Edition*, vol. 49, pp. 6778–6780, Sept. 2010.

[74] M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. Cheatham, and P. Jurečka, “Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles.,” *Journal of chemical theory and computation*, vol. 7, pp. 2886–2902, Sept. 2011.

[75] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *J. Chem. Theory Comput.*, vol. 4, pp. 435–447, Feb. 2008.

[76] L. X. Dang, J. E. Rice, J. Caldwell, and P. A. Kollman, "Ion solvation in polarizable water: molecular dynamics simulations," *J. Am. Chem. Soc.*, vol. 113, pp. 2481–2486, Mar. 1991.

[77] M. J. Elrod and R. J. Saykally, "Many-Body Effects in Intermolecular Forces," *Chem. Rev.*, vol. 94, pp. 1975–1997, Nov. 1994.

[78] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, pp. 10089–10092, June 1993.

[79] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *The Journal of Chemical Physics*, vol. 103, pp. 8577–8593, Nov. 1995.

[80] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, "A generalized reaction field method for molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 102, no. 13, pp. 5451–5459, 1995.

[81] C. L. Brooks, B. M. Pettitt, and M. Karplus, "Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids," *The Journal of Chemical Physics*, vol. 83, no. 11, pp. 5897–5908, 1985.

[82] H. Schreiber and O. Steinhauser, "Molecular dynamics studies of solvated polypeptides: Why the cut-off scheme does not work," *Chemical Physics*, vol. 168, pp. 75–89, Dec. 1992.

[83] P. Auffinger and D. L. Beveridge, "A simple test for evaluating the truncation effects in simulations of systems involving charged groups," *Chemical Physics Letters*, vol. 234, pp. 413–415, Mar. 1995.

[84] Y. Yonetani, "A severe artifact in simulation of liquid water using a long cut-off length: Appearance of a strange layer structure," *Chemical Physics Letters*, vol. 406, pp. 49–53, Apr. 2005.

[85] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *J. Comput. Chem.*, vol. 21, pp. 1049–1074, Sept. 2000.

[86] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins*, vol. 65, pp. 712–725, Nov. 2006.

[87] L. Yang, C.-h. Tan, M.-J. Hsieh, J. Wang, Y. Duan, P. Cieplak, J. Caldwell, P. A. Kollman, and R. Luo, "New-Generation Amber United-Atom Force Field," *The Journal of Physical Chemistry B*, vol. 110, pp. 13166–13176, July 2006.

[88] B. H. Besler, K. M. Merz, and P. A. Kollman, "Atomic charges derived from semiempirical methods," *J. Comput. Chem.*, vol. 11, pp. 431–439, May 1990.

[89] R. F. W. Bader, "A quantum theory of molecular structure and its applications," *Chem. Rev.*, vol. 91, pp. 893–928, July 1991.

[90] K. B. Wiberg and P. R. Rablen, "Comparison of atomic charges derived via different procedures," *J. Comput. Chem.*, vol. 14, pp. 1504–1518, Dec. 1993.

[91] F. A. Momany, “Determination of partial atomic charges from ab initio molecular electrostatic potentials. Application to formamide, methanol, and formic acid,” *J. Phys. Chem.*, vol. 82, pp. 592–601, Mar. 1978.

[92] U. C. Singh and P. A. Kollman, “An approach to computing electrostatic charges for molecules,” *Journal of Computational Chemistry*, vol. 5, pp. 129–145, Apr. 1984.

[93] R. S. Mulliken, “Electronic Population Analysis on LCAO[Single Bond]MO Molecular Wave Functions. I,” *The Journal of Chemical Physics*, vol. 23, no. 10, pp. 1833–1840, 1955.

[94] D. R. Hartree, “The wave mechanics of an atom with a non-Coulomb central field Part I theory and methods,” *PROCEEDINGS OF THE CAMBRIDGE PHILOSOPHICAL SOCIETY*, vol. 24, pp. 89–110, July 1928.

[95] V. Fock, “Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems,” *Zeitschrift für Physik*, vol. 61, no. 1-2, pp. 126–148, 1930.

[96] C. C. J. Roothaan, “New Developments in Molecular Orbital Theory,” *Reviews of Modern Physics*, vol. 23, pp. 69–89, Apr. 1951.

[97] G. G. Hall, “The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potentials,” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 205, pp. 541–552, Mar. 1951.

[98] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model,” *J. Phys. Chem.*, vol. 97, pp. 10269–10280, Oct. 1993.

[99] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollmann, “Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation,” *J. Am. Chem. Soc.*, vol. 115, pp. 9620–9631, Oct. 1993.

[100] E. Sigfridsson and U. Ryde, “Comparison of methods for deriving atomic charges from the electrostatic potential and moments,” *Journal of Computational Chemistry*, vol. 19, pp. 377+, Mar. 1998.

[101] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, pp. 1157–1174, July 2004.

[102] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, “Automatic atom type and bond type perception in molecular mechanical calculations.,” *Journal of molecular graphics & modelling*, vol. 25, pp. 247–260, Oct. 2006.

[103] F.-Y. Dupradeau, C. Cézard, R. Lelong, E. Stanislawiak, J. Pêcher, J. C. Delepine, and P. Cieplak, “R.E.DD.B.: A database for RESP and ESP atomic charges, and force field libraries,” *Nucleic Acids Research*, vol. 36, pp. D360–D367, Jan. 2008.

[104] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Interaction Models for Water in Relation to Protein Hydration, in Intermolecular Forces*, pp. 331–342. Dordrecht: D. Reidel Publishing Company, 1981.

[105] W. L. Jorgensen, “Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water,” *Journal of the American Chemical Society*, vol. 103, pp. 335–340, Jan. 1981.

[106] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[107] S. L. Shostak, W. L. Ebenstein, and J. S. Muenter, “The dipole moment of water. I. Dipole moments and hyperfine properties of H_2O and HDO in the ground and excited vibrational states,” *The Journal of Chemical Physics*, vol. 94, no. 9, pp. 5875–5882, 1991.

[108] A. V. Gubskaya and P. G. Kusalik, “The total molecular dipole moment for liquid water,” *The Journal of Chemical Physics*, vol. 117, no. 11, pp. 5290–5302, 2002.

[109] M. Orozco and F. J. Luque, “Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems,” *Chem. Rev.*, vol. 100, pp. 4187–4226, Oct. 2000.

[110] V. Makarov, B. M. Pettitt, and M. Feig, “Solvation and Hydration of Proteins and Nucleic Acids: A Theoretical View of Simulation and Experiment,” *Acc. Chem. Res.*, vol. 35, pp. 376–384, Mar. 2002.

[111] A. Vaiana, E. Westhof, and P. Auffinger, “A molecular dynamics simulation study ofanaminoglycoside/A-site RNA complex: conformational andhydration patterns,” *Biochimie*, vol. 88, pp. 1061–1073, Aug. 2006.

[112] P. J. van Maaren and D. van der Spoel, “Molecular Dynamics Simulations of Water with Novel Shell-Model Potentials,” *J. Phys. Chem. B*, vol. 105, pp. 2618–2626, Mar. 2001.

[113] C. Vega, J. L. F. Abascal, M. M. Conde, and J. L. Aragones, “What ice can teach us about water interactions: a critical comparison of the performance of different water models,” *Faraday Discuss.*, vol. 141, pp. 251–276, 2009.

[114] O. Gereben and L. Pusztai, “On the accurate calculation of the dielectric constant from molecular dynamics simulations: The case of SPC/E and SWM4-DP water,” *Chemical Physics Letters*, vol. 507, pp. 80–83, Apr. 2011.

[115] M. Born, “Volumen und Hydratationswärme der Ionen,” *Zeitschrift für Physik*, vol. 1, no. 1, pp. 45–48, 1920.

[116] L. Onsager, “Electric Moments of Molecules in Liquids,” *J. Am. Chem. Soc.*, vol. 58, pp. 1486–1493, Aug. 1936.

[117] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, “Semianalytical treatment of solvation for molecular mechanics and dynamics,” *J. Am. Chem. Soc.*, vol. 112, pp. 6127–6129, Aug. 1990.

[118] C. Cramer and D. Truhlar, “AM1-SM2 and PM3-SM3 parameterized SCF solvation models for free energies in aqueous solution,” *Journal of Computer-Aided Molecular Design*, vol. 6, no. 6, pp. 629–666, 1992.

[119] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, “Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium,” *J. Phys. Chem.*, vol. 100, pp. 19824–19839, Jan. 1996.

[120] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, and A. Onufriev, “Generalized Born Model with a Simple, Robust Molecular Volume Correction,” *J. Chem. Theory Comput.*, vol. 3, pp. 156–169, Dec. 2006.

[121] P. Larsson and E. Lindahl, “A high-performance parallel-generalized born implementation enabled by tabulated interaction rescaling,” *Journal of Computational Chemistry*, vol. 31, pp. 2593–2600, Nov. 2010.

[122] J. Schlitter, “Estimation of absolute and relative entropies of macromolecules using the covariance matrix,” *Chemical Physics Letters*, vol. 215, pp. 617–621, Dec. 1993.

[123] I. Andricioaei and M. Karplus, “On the calculation of entropy from covariance matrices of the atomic fluctuations,” *The Journal of Chemical Physics*, vol. 115, no. 14, pp. 6289–6292, 2001.

[124] G. Behravan, M. Leijon, U. Sehlstedt, B. Nordn, H. Vallberg, J. Bergman, and A. Gruslund, “The interaction of ellipticine derivatives with nucleic acids studied by optical and1H-nmr spectroscopy: Effect of size of the heterocyclic ring system,” *Biopolymers*, vol. 34, pp. 599–609, May 1994.

[125] A. Canals, M. Purciolas, J. Aymam’i, and M. Coll, “The anticancer agent ellipticine unwinds DNA by intercalative binding in an orientation parallel to base pairs,” *Acta Crystallographica Section D*, vol. 61, pp. 1009–1012, July 2005.

[126] Brana, Cacho, Gradillas, de Pascual-Teresa, and Ramos, “Intercalators as Anticancer Drugs,” *Current Pharmaceutical Design*, vol. 7, pp. 1745–1780, Nov. 2001.

[127] D. Řeha, M. Kabeláč, F. Ryjáček, J. Šponer, J. E. Šponer, M. Elstner, S. Suhai, and P. Hobza, “Intercalators. 1. Nature of Stacking Interactions between Intercalators (Ethidium, Daunomycin, Ellipticine, and 4’,6-Diaminide-2-phenylindole) and DNA Base Pairs. Ab Initio Quantum Chemical, Density Functional Theory, and Empirical Potential Study,” *J. Am. Chem. Soc.*, vol. 124, pp. 3366–3376, Mar. 2002.

[128] P. B. Dervan, “Molecular recognition of DNA by small molecules.,” *Bioorganic & medicinal chemistry*, vol. 9, pp. 2215–2235, Sept. 2001.

[129] A. V. Vargiu, P. Ruggerone, A. Magistrato, and P. Carloni, “Dissociation of minor groove binders from DNA: insights from metadynamics simulations,” *Nucleic Acids Research*, vol. 36, pp. 5910–5921, Oct. 2008.

[130] N. Hansen, J. Dolenc, M. Knecht, S. Riniker, and W. F. van Gunsteren, “Assessment of enveloping distribution sampling to calculate relative free enthalpies of binding for eight netropsin-DNA duplex complexes in aqueous solution,” *Journal of Computational Chemistry*, vol. 33, pp. 640–651, Mar. 2012.

[131] S. A. Harris, E. Gavathiotis, M. S. Searle, M. Orozco, and C. A. Laughton, “Cooperativity in DrugDNA Recognition: A Molecular Dynamics Study,” *J. Am. Chem. Soc.*, vol. 123, pp. 12658–12663, Nov. 2001.

[132] A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, “Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of / Conformers,” *Biophysical Journal*, vol. 92, pp. 3817–3829, June 2007.

[133] A. Mládek, J. E. Šponer, P. Jurečka, P. Banáš, M. Otyepka, D. Svozil, and J. Šponer, “Conformational Energies of DNA SugarPhosphate Backbone: Reference QM Calculations and a Comparison with Density Functional Theory and Molecular Mechanics,” *J. Chem. Theory Comput.*, vol. 6, pp. 3817–3835, Nov. 2010.

[134] A. H. Elcock, A. Rodger, and W. G. Richards, “Theoretical studies of the intercalation of 9-hydroxyellipticine in DNA,” *Biopolymers*, vol. 39, pp. 309–326, Dec. 1998.

[135] X. Wei, S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw, “Viral dynamics in human immunodeficiency virus type 1 infection,” *Nature*, vol. 373, pp. 117–122, Jan. 1995.

[136] A. D. Frankel and J. A. T. Young, “HIV-1: Fifteen Proteins and an RNA,” *Annual Review of Biochemistry*, vol. 67, no. 1, pp. 1–25, 1998.

[137] M. A. Navia, P. M. D. Fitzgerald, B. M. McKeever, C.-T. Leu, J. C. Heimbach, W. K. Herber, I. S. Sigal, P. L. Darke, and J. P. Springer, "Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1," *Nature*, vol. 337, pp. 615–620, Feb. 1989.

[138] A. Brik and C.-H. Wong, "HIV-1 protease: mechanism and drug discovery," *Org. Biomol. Chem.*, vol. 1, no. 1, pp. 5–14, 2003.

[139] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions," *J. Phys. Chem. B*, vol. 113, pp. 6378–6396, Apr. 2009.

[140] R. Luo, L. David, and M. K. Gilson, "Accelerated PoissonBoltzmann calculations for static and dynamic systems," *J. Comput. Chem.*, vol. 23, pp. 1244–1253, Oct. 2002.

[141] Q. Lu and R. Luo, "A Poisson–Boltzmann dynamics method with nonperiodic boundary condition," *The Journal of Chemical Physics*, vol. 119, no. 21, pp. 11035–11047, 2003.

[142] A. Klamt, "Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena," *J. Phys. Chem.*, vol. 99, pp. 2224–2235, Feb. 1995.

[143] C. Curutchet, M. Orozco, and F. J. Luque, "Solvation in octanol: parametrization of the continuum MST model," *Journal of Computational Chemistry*, vol. 22, pp. 1180–1193, Aug. 2001.

[144] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, "Molecular Properties That Influence the Oral Bioavailability of Drug Candidates," *J. Med. Chem.*, vol. 45, pp. 2615–2623, May 2002.

[145] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz, "Refinement and Parametrization of COSMO-RS," *J. Phys. Chem. A*, vol. 102, pp. 5074–5085, June 1998.

[146] F. Guthrie, "XXVIII. - On the iodide of iodammonium," *Journal of the Chemical Society*, vol. 16, pp. 239–244, 1863.

[147] O. Hassel, J. Hvoslef, E. H. Vihovde, and N. A. Sörensen, "The Structure of Bromine 1,4-Dioxanate.," *Acta Chemica Scandinavica*, vol. 8, p. 873, 1954.

[148] O. Hassel, K. O. Strømme, H. Haraldsen, A. Grönvall, B. Zaar, and E. Diczfalusy, "Structure of the Crystalline Compound Benzene-Bromine (1:1).," *Acta Chemica Scandinavica*, vol. 12, p. 1146, 1958.

[149] O. Hassel, K. O. Strømme, E. Stenhagen, G. Andersson, E. Stenhagen, and H. Palmstierna, "Crystal Structure of the 1:1 Addition Compound Formed by Acetone and Bromine.," *Acta Chemica Scandinavica*, vol. 13, pp. 275–280, 1959.

[150] H. A. Bent, "Structural chemistry of donor-acceptor interactions," *Chem. Rev.*, vol. 68, pp. 587–648, Oct. 1968.

[151] O. Hassel, "Structural Aspects of Interatomic Charge-Transfer Bonding," *Science*, vol. 170, pp. 497–502, Oct. 1970.

[152] J. P. M. Lommerse, A. J. Stone, R. Taylor, and F. H. Allen, "The Nature and Geometry of Intermolecular Interactions between Halogens and Oxygen or Nitrogen," *J. Am. Chem. Soc.*, vol. 118, pp. 3108–3116, Jan. 1996.

[153] T. Clark, M. Hennemann, J. S. Murray, and P. Politzer, “Halogen bonding: the -hole,” *Journal of Molecular Modeling*, vol. 13, pp. 291–296, Feb. 2007.

[154] P. Politzer, P. Lane, M. Concha, Y. Ma, and J. Murray, “An overview of halogen bonding,” *Journal of Molecular Modeling*, vol. 13, pp. 305–311, Feb. 2007.

[155] K. E. Riley and P. Hobza, “Investigations into the Nature of Halogen Bonding Including Symmetry Adapted Perturbation Theory Analyses,” *J. Chem. Theory Comput.*, vol. 4, pp. 232–242, Jan. 2008.

[156] J. Řezáč, K. E. Riley, and P. Hobza, “Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules,” *J. Chem. Theory Comput.*, vol. 8, pp. 4285–4292, Sept. 2012.

[157] J. Řezáč, K. E. Riley, and P. Hobza, “S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures,” *J. Chem. Theory Comput.*, vol. 7, pp. 2427–2438, July 2011.

[158] Y.-X. Lu, J.-W. Zou, J.-C. Fan, W.-N. Zhao, Y.-J. Jiang, and Q.-S. Yu, “Ab initio calculations on halogen-bonded complexes and comparison with density functional methods,” *Journal of Computational Chemistry*, vol. 30, pp. 725–732, Apr. 2009.

[159] P. Metrangolo, J. S. Murray, T. Pilati, P. Politzer, G. Resnati, and G. Terraneo, “Fluorine-Centered Halogen Bonding: A Factor in Recognition Phenomena and Reactivity,” *Crystal Growth & Design*, vol. 11, pp. 4238–4246, Aug. 2011.

[160] P. Politzer, K. E. Riley, F. A. Bulat, and J. S. Murray, “Perspectives on halogen bonding and other -hole interactions: Lex parsimoniae (Occam’s Razor),” *Computational and Theoretical Chemistry*, vol. 998, pp. 2–8, Oct. 2012.

[161] P. Politzer, J. S. Murray, and M. C. Concha, “-hole bonding between like atoms; a fallacy of atomic charges,” *Journal of Molecular Modeling*, vol. 14, pp. 659–665, Aug. 2008.

[162] E. Corradi, S. V. Meille, M. T. Messina, P. Metrangolo, and G. Resnati, “Halogen Bonding versus Hydrogen Bonding in Driving Self-Assembly Processes,” *Angewandte Chemie International Edition*, vol. 39, pp. 1782–1786, May 2000.

[163] P. Metrangolo, H. Neukirch, T. Pilati, and G. Resnati, “Halogen Bonding Based Recognition Processes: A World Parallel to Hydrogen Bonding†,” *Accounts of Chemical Research*, vol. 38, pp. 386–395, May 2005.

[164] C. M. Reddy, M. T. Kirchner, R. C. Gundakaram, K. A. Padmanabhan, and G. R. Desiraju, “Isostructurality, Polymorphism and Mechanical Properties of Some Hexahalogenated Benzenes: The Nature of HalogenHalogen Interactions,” *Chem. Eur. J.*, vol. 12, pp. 2222–2234, Mar. 2006.

[165] A. C. Legon, “The halogen bond: an interim perspective,” *Phys. Chem. Chem. Phys.*, vol. 12, no. 28, pp. 7736–7747, 2010.

[166] A. R. Voth, P. Khuu, K. Oishi, and P. S. Ho, “Halogen bonds as orthogonal molecular interactions to hydrogen bonds,” *Nature Chemistry*, vol. 1, pp. 74–79, Mar. 2009.

[167] K. Riley, J. Řezáč, and P. Hobza, “Competition between halogen, dihalogen and hydrogen bonds in bromo- and iodomethanol dimers,” *Journal of Molecular Modeling*, pp. 1–5, 2013.

[168] B. Jeziorski, R. Moszynski, and K. Szalewicz, “Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes,” *Chem. Rev.*, vol. 94, pp. 1887–1930, Nov. 1994.

[169] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *The Journal of Chemical Physics*, vol. 132, no. 15, pp. 154104+, 2010.

[170] N. Boden, P. P. Davis, C. H. Stam, and G. A. Wesselink, "Solid hexafluorobenzene," *Molecular Physics*, vol. 25, pp. 81–86, Jan. 1973.

[171] A. Budzianowski and A. Katrusiak, "Pressure-frozen benzene I revisited," *Acta Crystallographica Section B*, vol. 62, pp. 94–101, Feb. 2006.

[172] P. Hobza, H. L. Selzle, and E. W. Schlag, "Potential Energy Surface for the Benzene Dimer. Results of ab Initio CCSD(T) Calculations Show Two Nearly Isoenergetic Structures: T-Shaped and Parallel-Displaced," *J. Phys. Chem.*, vol. 100, pp. 18790–18794, Jan. 1996.

[173] S. Tsuzuki, K. Honda, T. Uchimaru, M. Mikami, and K. Tanabe, "Origin of Attraction and Directionality of the / Interaction: Model Chemistry Calculations of Benzene Dimer Interaction," *J. Am. Chem. Soc.*, vol. 124, pp. 104–112, Dec. 2001.

[174] R. Holliday, "A mechanism for gene conversion in fungi," *Genetics Research*, vol. 5, pp. 282–304, June 1964.

[175] M. Carter and P. S. Ho, "Assaying the Energies of Biological Halogen Bonds," *Crystal Growth & Design*, vol. 11, pp. 5087–5095, Sept. 2011.

[176] F. A. Hays, J. M. Vargason, and P. S. Ho, "Effect of Sequence on the Conformation of DNA Holliday Junctions†," *Biochemistry*, vol. 42, pp. 9586–9597, July 2003.

[177] M. Z. Hernandes, S. M. Cavalcanti, D. R. Moreira, W. F. de Azevedo, and A. C. Leite, "Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design," *Current Drug Targets*, vol. 11, no. 3, pp. 303–314, 2010.

[178] P. Auffinger, F. A. Hays, E. Westhof, and P. S. Ho, "Halogen bonds in biological molecules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 16789–16794, Nov. 2004.

[179] Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang, and W. Zhu, "Halogen BondingA Novel Interaction for Rational Drug Design?," *Journal of Medicinal Chemistry*, vol. 52, pp. 2854–2862, May 2009.

[180] K. E. Riley and P. Hobza, "Strength and Character of Halogen Bonds in ProteinLigand Complexes," *Crystal Growth & Design*, vol. 11, pp. 4272–4278, Sept. 2011.

[181] E. Parisini, P. Metrangolo, T. Pilati, G. Resnati, and G. Terraneo, "Halogen bonding in halocarbon-protein complexes: a structural survey," *Chem. Soc. Rev.*, vol. 40, no. 5, pp. 2267–2278, 2011.

[182] E. I. Howard, R. Sanishvili, R. E. Cachau, A. Mitschler, B. Chevrier, P. Barth, V. Lamour, M. Van Zandt, E. Sibley, C. Bon, D. Moras, T. R. Schneider, A. Joachimiak, and A. Podjarny, "Ultrahigh resolution drug design I: Details of interactions in human aldose reductase-inhibitor complex at 0.66 Å," *Proteins: Structure, Function, and Bioinformatics*, vol. 55, pp. 792–804, Apr. 2004.

[183] A. Wojtczak, V. Cody, J. R. Luft, and W. Pangborn, "Structure of rat transthyretin (rTTR) complex with thyroxine at 2.5Å resolution: first non-biased insight into thyroxine binding reveals different hormone orientation in two binding sites," *Acta Crystallographica Section D*, vol. 57, pp. 1061–1070, Aug. 2001.

[184] M. A. Pagano, M. Andrzejewska, M. Ruzzene, S. Sarno, L. Cesaro, J. Bain, M. Elliott, F. Meggio, Z. Kazimierczuk, and L. A. Pinna, "Optimization of Protein Kinase CK2 Inhibitors Derived from 4,5,6,7-Tetrabromobenzimidazole," *J. Med. Chem.*, vol. 47, pp. 6239–6247, Nov. 2004.

[185] R. W. Dixon and P. A. Kollman, "Advancing beyond the atom-centered model in additive and nonadditive molecular mechanics," *Journal of Computational Chemistry*, vol. 18, pp. 1632–1646, Oct. 1997.

[186] J. S. Rowlinson, "The lattice energy of ice and the second virial coefficient of water vapour," *Trans. Faraday Soc.*, vol. 47, no. 0, pp. 120–129, 1951.

[187] M. A. A. Ibrahim, "Molecular mechanical study of halogen bonding in drug discovery," *Journal of Computational Chemistry*, vol. 32, no. 12, pp. 2564–2574, 2011.

[188] S. Pieraccini, A. Forni, and M. Sironi, "Halogen bonding in ligand-receptor systems in the framework of classical force fields," *Phys. Chem. Chem. Phys.*, 2011.

[189] W. L. Jorgensen and P. Schyman, "Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents," *J. Chem. Theory Comput.*, vol. 8, pp. 3895–3901, Apr. 2012.

[190] K. Riley, J. Murray, J. Fanfrlík, J. Řezáč, R. Solá, M. Concha, F. Ramos, and P. Politzer, "Halogen bond tunability I: the effects of aromatic fluorine substitution on the strengths of halogen-bonding interactions involving chlorine, bromine, and iodine," *Journal of Molecular Modeling*, pp. 1–10, Mar. 2011.

[191] E. De Moliner, N. R. Brown, and L. N. Johnson, "Alternative binding modes of an inhibitor to two different kinases," *European Journal of Biochemistry*, vol. 270, pp. 3174–3181, Aug. 2003.

[192] R. Battistutta, M. Mazzorana, S. Sarno, Z. Kazimierczuk, G. Zanotti, and L. A. Pinna, "Inspecting the Structure-Activity Relationship of Protein Kinase CK2 Inhibitors Derived from Tetrabromo-Benzimidazole," *Chemistry & Biology*, vol. 12, pp. 1211–1219, Nov. 2005.

[193] R. Battistutta, M. Mazzorana, L. Cendron, A. Bortolato, S. Sarno, Z. Kazimierczuk, G. Zanotti, S. Moro, and L. A. Pinna, "The ATP-Binding Site of Protein Kinase CK2 Holds a Positive Electrostatic Area and Conserved Water Molecules," *ChemBioChem*, vol. 8, pp. 1804–1809, Oct. 2007.

[194] P. Dobeš, J. Řezáč, J. Fanfrlík, M. Otyepka, and P. Hobza, "Semiempirical Quantum Mechanical Method PM6-DH2X Describes the Geometry and Energetics of CK2-Inhibitor Complexes Involving Halogen Bonds Well, While the Empirical Potential Fails," *J. Phys. Chem. B*, vol. 115, pp. 8581–8589, June 2011.

[195] E. C. Meng, B. K. Shoichet, and I. D. Kuntz, "Automated docking with grid-based energy evaluation," *J. Comput. Chem.*, vol. 13, no. 4, pp. 505–524, 1992.

[196] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, and I. D. Kuntz, "DOCK 6: Combining techniques to model RNAsmall molecule complexes," *RNA*, vol. 15, pp. 1219–1230, June 2009.

[197] H. Steuber, A. Heine, and G. Klebe, "Structural and Thermodynamic Study on Aldose Reductase: Nitro-substituted Inhibitors with Strong Enthalpic Binding Contribution," *Journal of Molecular Biology*, vol. 368, pp. 618–638, May 2007.

[198] T. O. Fischmann, A. Hruza, J. S. Duca, L. Ramanathan, T. Mayhood, W. T. Windsor, H. V. Le, T. J. Guzi, M. P. Dwyer, K. Paruch, R. J. Doll, E. Lees, D. Parry, W. Seghezzi, and V. Madison, "Structure-guided discovery of cyclin-dependent kinase inhibitors," *Biopolymers*, vol. 89, pp. 372–379, May 2008.

[199] D. M. Himmel, K. Das, A. D. Clark, S. H. Hughes, A. Benjahad, S. Oumouch, J. Guillemont, S. Coupa, A. Poncelet, I. Csoka, C. Meyer, K. Andries, C. H. Nguyen, D. S. Grierson, and E. Arnold, "Crystal Structures for HIV-1 Reverse Transcriptase in Complexes with Three Pyridinone Derivatives: A New Class of Non-Nucleoside Inhibitors Effective against a Broad Range of Drug-Resistant Strains," *J. Med. Chem.*, vol. 48, pp. 7582–7591, Nov. 2005.

List of Publications

Included in the Thesis

1. M. Kolář, J. Fanfrlík, M. Lepšík, F. Forti, F. J. Luque, and P. Hobza "Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches," Submitted to The Journal of Physical Chemistry B.

2. J. Trnka, R. Sedlák, M. Kolář and P. Hobza, "The differences in the sublimation energy of benzene and hexahalogenbenzenes are caused by dispersion energy," Submitted to The Journal of Physical Chemistry A.

3. M. Kolář, P. Hobza and A. K. Bronowska, "Plugging the Explicit σ-Holes in Molecular Docking," *Chemical Communications*, vol. 49, pp. 981–983, Jan. 2013.

4. M. Kolář, "Halogenová vazba aneb popletené náboje novou nadějí pro medicínu," *Vesmír*, vol. 91, pp. 522–523, Oct. 2012.

5. M. Kolář and P. Hobza, "On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds," *The Journal of Chemical Theory and Computation*, vol. 8, pp. 1325–1333, Apr. 2012.

6. M. Kolář, J. Fanfrlík, and P. Hobza "Ligand Conformational and Solvation/Desolvation Free Energy in Protein-Ligand Complex Formation," *The Journal of Physical Chemistry B*, vol. 115, pp. 4718–4724, Apr. 2011.

7. M. Kolář, T. Kubař, and P. Hobza, "Sequence-Dependent Configurational Entropy Change of DNA upon Intercalation," *The Journal of Physical Chemistry B*, vol. 114, pp. 13446–13454, Oct. 2011.

Not Included in the Thesis

1. M. Kolář, "Evoluce sbalování proteinů," *Vesmír*, vol. 92, pp. 263–264, May 2013.

2. M. Kolář and P. Hobza, "Jsou to opravdu vodíkové vazby, které stabilizují DNA?," *Vesmír*, vol. 92, pp. 140–141, Feb. 2013.

3. S. Haldar*, M. Kolář*, P. Sedlák, and P. Hobza, "Adsorption of Organic Electron Acceptors on Graphene-like Molecules: Quantum Chemical and Molecular Mechanical study," *The Journal of Physical Chemistry C*, vol. 116, pp. 25328–25336, Dec. 2012.
   *these authors contributed equally

4. M. Kolář, T. Kubař and P. Hobza "On the Role of London Dispersion Forces in Biomolecular Structure Determination," *The Journal of Physical Chemistry B*, vol. 115, pp. 8038–8046, June 2011.

5. M. Kolář, K. Berka, P. Jurečka, and P. Hobza, "On the Reliability of the AMBER Force Field and its Empirical Dispersion Contribution for the Description of Noncovalent Complexes," *ChemPhysChem*, vol. 11, pp. 2399–2408, Aug. 2010.

6. M. Kolář and P. Hobza, "Accurate Theoretical Determination of the Structure of Aromatic Complexes is Complicated: The Phenol Dimer and Phenol$\cdots$Methanol Cases," *The Journal of Physical Chemistry A*, vol. 111, pp. 5851–5854, July 2007.

Presentations of the Results

Lectures, Seminars

1. “Noncovalent Interactions Involving Halogen Atoms” (Dec 2012) Department of physical and macromolecular chemistry, Faculty of Science, Charles University in Prague, Czech Republic.

2. “Computers in the World of Atoms and Molecules” (Nov 2012) 4th Scientific Conference, Faculty of Science, Charles University in Prague, Czech Republic.

3. “Molecular Mechanical Treatment of Halogen Bonding’ (Oct 2012) Computational Biomedicine group, Forschungszentrum Jülich, Federal Republic of Germany.

4. “The Devil is in the Details – Halogen Bonding, σ-holes, and Rational Drug Design” (Sept 2012) Annual Meeting of the German Biophysical Society, Göttingen, Federal Republic of Germany, presented by Agnieszka K. Bronowska.

5. “Halogen Bonds in the Context of Current Empirical Force Fields” (June 2012) Steinbuch Center for Computing, Karlsruher Institut für Technologie, Karlsruhe, Germany

6. “Halogen Bonds and their Description by Molecular Mechanics” (Apr 2012) Workshop of Computer Simulation and Theory of Macromolecules, Hünfeld, Germany.

7. “Halogen Bonds and their Description by Molecular Mechanics” (Feb 2012) Molecular Biomechanics group, Heidelberger Institute für Teoretische Studien, Federal Republic of Germany.

8. “Treatment of Halogen Bonding with Current Biomolecular Empirical Force Field” (Jan 2012) Theoretical Chemical Biology group, Karlsruher Institut für Technologie, Federal Republic of Germany.

9. “Treatment of Halogen Bonding with Current Biomolecular Empirical Force Fields” (Nov 2011) Ninth Annual Congress of International Drug Discovery Science and Technology, Shenzhen, People’s Republic of China.

10. “Treatment of Halogen Bonding with Current Biomolecular Empirical Force Field” (Jul 2011) Ninth Triennial Congress of the World Association of Theoretical and Computational Chemists; Santiago de Compostela, Kingdom of Spain.

11. “Ligand Binding: Entropy Plays an Important Role” (May 2010) Conference of the Institute of Organic Chemistry and Biochemistry AS CR, v.v.i., Frymburk, Czech Republic.

Posters

1. “Water-Octanol Transfer Free Energies: Role of Conformational Flexibility” (Apr 2012) Workshop of Computer Simulation and Theory of Macromolecules, Hünfeld, Federal Rapublic of Germany.

2. “Treatment of Halogen Bonding with Current Biomolecular Empirical Force Fields” (Jan 2012) Isolated Biomolecules and Biomolecular Interactions. Les Diablerets, Swiss Confederation.

3. “The Effect of Dynamics on Calculated Stabilities of Ligand Conformations and Energies of Protein-Ligand Interactions” (Jan 2012) Isolated Biomolecules and Biomolecular Interactions. Les Diablerets, Swiss Confederation, presented by Martin Lepšík.

4. “Ligand Conformational and Solvation Free Energy in Protein-Ligand Complex Formation” (Jul 2011) Computer Aided Drug Design The Impact of Computational Sciences along the Drug Discovery Process, West Dover, USA, presented by Jindřich Fanfrlík.

5. “Configurational Entropy Change of DNA upon Intercalation of a Small Drug Molecule” (Aug 2010) Methods in Molecular Simulations Summer School, Queen’s University Belfast, United Kingdom.

A

Declaration of the Shared Authorship

ÚOCHB AV ČR, v.v.i.

ÚSTAV ORGANICKÉ CHEMIE A BIOCHEMIE
AKADEMIE VĚD ČESKÉ REPUBLIKY, v.v.i.
INSTITUTE OF ORGANIC CHEMISTRY AND BIOCHEMISTRY
ACADEMY OF SCIENCES OF THE CZECH REPUBLIC, v.v.i.

Prohlášení

Prohlášení spoluautorů upřesňující podíl Mgr. Michala Koláře na publikacích přiložených k disertaci:

1. M. Kolář, J. Fanfrlík, M. Lepšík, F. Forti, F. J. Luque, P. Hobza „Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches“ Submitted to The Journal of Physical Chemistry B.

2. J. Trnka, R. Sedlák, M. Kolář, P. Hobza „The differences in the sublimation energy of benzene and hexahalogenbenzenes are caused by dispersion energy“ Submitted to The Journal of Physical Chemistry A.

3. M. Kolář, P. Hobza, A. K. Bronowska „Plugging the Explicit σ-Holes in Molecular Docking” Chemical Communications, 2012, 49 (10), 981–983.

4. M. Kolář, „Halogenová vazba aneb popletené náboje novou nadějí pro medicínu“ Vesmír, 91, 522–523.

5. M. Kolář, P. Hobza „On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds“ The Journal of Chemical Theory and Computation, 2012, 8 (4), 1325–1333.

6. M. Kolář, J. Fanfrlík, P. Hobza „Ligand Conformational and Solvation/Desolvation Free Energy in Protein-Ligand Complex Formation“ The Journal of Physical Chemistry B, 2011 115 (16), 4718–4724.

7. M. Kolář, T. Kubař, P. Hobza „Sequence-Dependent Configurational Entropy Change of DNA upon Intercalation“ The Journal of Physical Chemistry B, 2010, 114 (42), 13446–13454.

Mgr. Michal Kolář je prvním autorem na většině publikací přiložených k disertaci, což jednoznačně vymezuje jeho podíl. Ve všech případech je tento podíl dominantní a to ve všech fázích přípravy publikace, od zadání tématu až k jejímu sepsání.

V Praze, 3. dubna 2013

Pavel Hobza

Prof. Ing. Pavel Hobza, Dr.Sc., dr. h. c., FRSC

B

Publication 1 – Conformational Entropy

Sequence-Dependent Configurational Entropy Change of DNA upon Intercalation

Michal Kolář,[†] **Tomáš Kubař,**[‡] **and Pavel Hobza***[,†,§]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague 6, Czech Republic, Institute of Physical Chemistry, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany, and Department of Physical Chemistry, Palacký University, Olomouc, 771 46 Olomouc, Czech Republic

Received: March 3, 2010; Revised Manuscript Received: August 25, 2010

We investigated the intercalation of an antitumor drug ellipticine into four adenine−thymine (AT) rich DNA duplexes with the focus on the configurational entropy, by means of molecular dynamics (MD) simulations. Two possible binding orientations of ellipticine in a DNA double helix were studied, and the orientation with the pyrrole nitrogen exposed into a major groove was identified as the more probable. The configurational entropy change of DNA is shown to contribute significantly to the binding free energy. The magnitude of this contribution depends on the exact DNA sequence. A detailed analysis revealed that the largest flexibility changes occurred in the sugar−phosphate backbone, resulting in an entropy gain in the most cases. The nucleobases were not involved in the changes of flexibility and entropy. BI/BII-like conformational transitions were observed after the intercalation of ellipticine, and the consequences of these transitions for the evaluation of entropy are discussed.

Introduction

As the carrier of genetic information, DNA constantly lies at the center of research interest in the life sciences. Particular attention has been focused on the interaction of small organic molecules with double-stranded DNA. The binding of a ligand changes the structure and/or dynamics of DNA, possibly affecting its function in the living cell. For example, the bound ligand itself or the distorted structure of DNA may interfere with the action of enzymes facilitating replication or transcription. Such an event may severely affect the rate and fidelity of those processes and thus influence the operation of the living cell. The mesoscopic effects could include cell death or, on the contrary, uncontrolled cell proliferation. Consequently and rather controversially, some of DNA-binding drugs act as mutagens and carcinogens, whereas others have found their application in the therapy of certain diseases, including cancer.[1,2]

The ligand can bind into DNA in a number of ways.[3−5] Apart from covalently bound adducts, the ligand molecule can be held in place by the action of noncovalent interactions, that is, electrostatic attraction, van der Waals forces, and hydrophobic effects. These processes governed by noncovalent forces are of a reversible nature and can be conveniently studied by means of molecular dynamics (MD) simulations.

Ellipticine (Figure 1) is a plant alkaloid with antitumor activity, which has been the object of many experimental and theoretical studies concerning its activity, sequence selectivity, and metabolism; various ellipticine derivatives have been studied as well.[6−14] As an intercalator, it exhibits a lower sequence specificity as compared to minor-groove binders;[15] the pyrimidine−purine steps are slightly preferred because of their larger flexibility.[16] The first experimental evidence of the interaction of ellipticine with a DNA-like structure was given by Jain et al. who obtained a crystal of ellipticine bound to two dinucleoside monophosphates.[7] More recently, Canals et al. suggested a possible binding mode of ellipticine to a DNA hexamer (PDB ID 1Z3F)[14] admitting, however, that this structure need not necessarily be the only one possible. Owing to the asymmetry of the ellipticine molecule, there are two families of intercalative modes into palindromic DNA double helices. In this work, we studied both possible intercalative motives—first, with the pyrrole nitrogen oriented into the major groove (denoted “int1”) and, second, with the pyrrole nitrogen oriented into the minor groove (“int2”). The latter motif is found in the mentioned X-ray structure by Canals et al. although the DNA sequences studied in this work differ from the X-ray one significantly. Both binding modes are sketched in Figure 2.

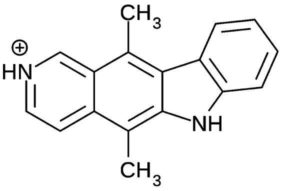


Figure 1. Protonated form of ellipticine.

The strength of the ligand binding to DNA can be quantified by the equilibrium constant or, equivalently, by the binding free energy. These quantities are accessible by experimental techniques[17−21] as well as by computer simulations.[22,23] The binding free energy, which amounts to −5 to −6 kcal·mol^{-1} for the intercalation of ellipticine under various conditions,[6] describes the overall binding, which, however, might be a very complex process. The free energy is composed of the enthalpic and the entropic parts. The change of enthalpy upon complex formation relates to the creation or breaking of hydrogen bonds within the ligand molecule, DNA, and the solvent and also to the electrostatic and dispersion interaction, especially (but not

[*] Corresponding author. Tel.: (+420) 220 410 311; e-mail: pavel.hobza@uochb.cas.cz.
[†] Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems.
[‡] Karlsruhe Institute of Technology.
[§] Palacký University.

10.1021/jp1019153
Published on Web 10/06/2010

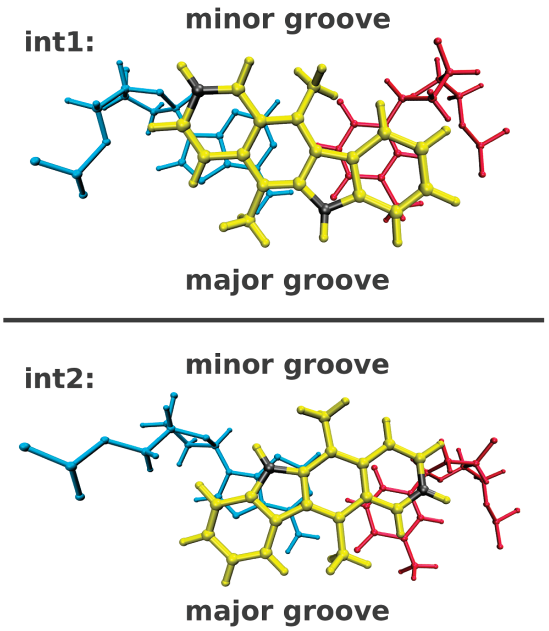


Figure 2. Schematic representation of the studied orientations of ellipticine in the intercalation site; top view with a base pair below the intercalator. Top: "int1", the nitrogen atom in the five-member ring (pyrrole) is oriented toward the major groove; bottom: "int2", the pyrrole nitrogen atom is oriented toward the minor groove. Red: thymine nucleotide; blue: adenine nucleotide; yellow: intercalator (gray: nitrogen atoms).

only) between the ligand and the DNA.[24] These contributions may be conveniently separated according to the pairs of interacting partners (i.e., Δ*E*(ligand...DNA), Δ*E*(ligand...solvent), etc.), and their evaluation is relatively straightforward, in principle.

Being a typical noncovalent binding motif, the intercalation of a small molecule requires two consecutive base pairs in the DNA double helix to be separated before the ligand can be accommodated.[1,2,24] Such a separation of base pairs is energetically unfavorable, with a reaction enthalpy of 20−24 kcal·mol^{-1}, as evaluated previously.[25] Among all of the components of the total reaction free energy, the interaction energy of the intercalator with the DNA species plays a major role in compensating for this energetic penalty. In a previous study, we showed that this interaction energy is substantial, reaching a value of approximately −70 kcal·mol^{-1} for a charged intercalating molecule of ethidium; the situation with ellipticine is expected to be very similar.

The binding entropy is commonly evaluated as the difference of the previously obtained values of the free energy and the enthalpy of binding, while direct measurements of entropy changes are quite rare.[26] Moreover, it seems to be difficult to express the entropy as a sum of physically meaningful components.[27] A reasonable decomposition of entropy appears to be that into the solute and solvent contributions, although the components need not be strictly additive because of the correlation between the solute and the solvent degrees of freedom. The term "configurational entropy" has been used for the entropy of the solute, usually without transitional and rotational contributions.[28] Upon ligand binding, the flexibility of the target (DNA in our case) is likely to change, with the thermodynamic consequence being the change of the configurational entropy. In other words, a calculation of configurational entropy makes it possible to assess the importance of the flexibility changes for the thermodynamics of ligand binding. The essential role of flexibility of the molecules involved in biochemical processes has been recognized. Even in the computer-aided drug design where static models have been applied in the docking-scoring schemes, approaches to consider the changes of configurational entropy upon binding tend to attract distinct attention.[29] For instance, Schlitter's scheme to estimate entropy (see bellow) is used in several free energy estimators or scoring functions,[30−32] although the entropy of the ligand is considered and that of the target is often ignored.[28]

TABLE 1: Palindromic DNA Duplex Sequences Studied[a]

A	5′-CGA**TA**T*(int)*A**TA**TCG-3′
B	5′-CGT**TA**T*(int)*A**TA**ACG-3′
C	5′-CG**TA**AT*(int)*AT**TA**CG-3′
D	5′-CG**TA**TT*(int)*AA**TA**CG-3′

[a] The TA steps are in boldface, and *(int)* denotes the ellipticine intercalation site.

In the past decades, computational approaches to estimate the configurational entropy have been developed. The pioneering method by Karplus and Kushick[33] is based on the evaluation of covariance matrix of fluctuations in internal coordinates. More feasible approaches were proposed by Schlitter[34] and by Andricioaei and Karplus,[35] which both of work with fluctuations in Cartesian coordinates and are based on the quasi-harmonic approximation; they have been thoroughly tested and frequently used[36−41] and provide nearly identical results. One of their advantages is the possibility to evaluate the components of total configurational entropy related to the various parts of the molecule[40] (e.g., the sugar−phosphate backbone and nucleobases in the case of DNA).

In this work, we investigate four adenine−thymine (AT) rich DNA species by means of MD simulations. The changes of configurational entropy upon the intercalation of ellipticine are evaluated, for the entire molecular system as well as for various parts of it, to assess the role that the flexibility of these parts plays in the thermodynamics of binding. The configurational entropy changes of DNA upon intercalation are shown to depend on the DNA sequence, indicating the possible importance of a correct account for the structural flexibility of DNA in the studies of ligand binding or even in the design of efficient ligands.

Methods

MD Simulations. The B-like structure of each of the four AT-rich DNA sequences (see Table 1) was generated with the Nucgen program from the AMBER 9 suite.[42] With the Xleap module of AMBER 9, the intercalation site was prepared manually by increasing the distance between two respective hexamers, while a spurious deformation of the phosphates connecting the hexamers was avoided by a simple relaxation available in the program. A protonated molecule of ellipticine was manually docked into the resulting cavity within the central TA step. Both possible orientations with the pyrrole nitrogen oriented into the major or the minor groove were prepared. Note that each of the sequences possesses two further TA steps (apart from the intercalation site), which are located two base pairs away from the intercalation site in the case of A and B and three base pairs away in case of C and D. The structures are provided as PDB files in Supporting Information.

For the simulations, the Gromacs 4 program package[43] along with the parm99 force field[44] including corrections for the backbone dihedral angles[45] was used systematically. The atomic charges of the ligand were derived by means of the RESP technique[46] based on the electrostatic potential calculated at the HF/6-31G* level using the Gaussian 03 package (revision C.02).[47] The remaining parameters of the ligand were assigned from the GAFF force field[48] designed for organic molecules, which is fully compatible with parm99.

The solute was solvated with approximately 10 000 TIP3P water molecules[49] in a cubic box with the edge length of 6.7 nm and neutralized with an appropriate number of Na^+ ions placed according to the electrostatic potential. The MD simulations were performed under periodic boundary conditions, at a constant temperature of 300 K and a constant pressure of 1 bar, maintained by the Nosé−Hoover[50,51] and the Parrinello−Rahman[52] algorithms, respectively. The length of all of the bonds involving a hydrogen atom was constrained to the respective equilibrium value by a parallel version of the LINCS algorithm,[53] which made it possible to use an integration time step of 2 fs. The long-range electrostatic interactions were treated with the PME algorithm[54] with a 1.2 nm direct-space cutoff. The Lennard−Jones (LJ) interactions were cut off at the separation of 1.2 nm, and the neighbor list was updated every 10 steps. The center-of-mass translation was removed.

Prior to the 60 ns production run, an equilibration was performed. Water molecules were minimized in 100 cycles, followed by a minimization of the whole system (including the DNA). Subsequently, the water was heated to 300 K within a 20 ps constant-volume simulation with position restraints on the DNA. In a subsequent 20 ps constant-pressure simulation, the position restraints were released, and the entire system was heated to 300 K. The equilibration was concluded with a free 1 ns constant-pressure simulation. The structure of the solute was recorded every picosecond.

Estimation of Entropy. The configurational entropy can be estimated according to both Schlitter and Andricioaei and Karplus, as the entropy of the one-dimensional quantum-mechanical harmonic oscillator (eqs 1 and 2)[34,35]

$$S_{\mathrm{HO}} = \frac{k\alpha}{e^{\alpha} - 1} - k\ln(1 - e^{-\alpha}) \tag{1}$$

$$\alpha = \frac{\hbar\omega}{kT} \tag{2}$$

where $\hbar$ is the reduced Planck constant and ω is the harmonic oscillator frequency. Schlitter further introduced the approximation in eq 3 and proved that $S \leq S_{\mathrm{HO}} < S'$.[34]

$$S' = \frac{1}{2}k\ln\left(1 + \frac{e^2}{\alpha^2}\right) \tag{3}$$

A complex molecular system with many degrees of freedom is approximated by a system of uncoupled harmonic oscillators. The necessary frequencies ω can be obtained by means of the quasi-harmonic analysis, employing a diagonalization of the mass-weighted covariance matrix. This approximation assumes a multivariate normal distribution of the atomic fluctuations. Following Andricioaei and Karplus, the entropy is obtained as

$$S_{\mathrm{AK}} = k\sum_{i}^{3N-6} \frac{\hbar\omega_i/kT}{e^{\hbar\omega_i/kT} - 1} - \ln(1 - e^{-\hbar\omega_i/kT}) \tag{4}$$

$$\omega_i = \sqrt{kT/\lambda_i} \tag{5}$$

where the sum runs over all $3N - 6$ nonzero λ_i eigenvalues of the mass-weighted covariance matrix in Cartesian coordinates. This provides a tighter upper bound to the true entropy than Schlitter's formula. The advantage of Schlitter's procedure is that it may be transformed to a simple calculation of a determinant of a matrix, which was more computationally efficient than diagonalization at that time. Still, both approaches yield results that are identical within numerical precision, and Schlitter's formula was tested for DNA by Harris and Laughton[41] with excellent results.

In this work, the mass-weighted covariance matrix was evaluated and diagonalized by the tools provided with Gromacs 4 simulation package.[43] The obtained eigenvalues were converted to frequencies (eq 5), and the entropy was calculated (eq 4).

In our calculations, the global translational and rotational entropy were removed with a least-squares fitting to a reference structure. As the reference in every calculation of entropy, we used a structure calculated as an average over all frames fitted to the first frame of the particular trajectory. Such an average structure might be iteratively improved until convergence is reached. Our resulting entropies are, however, rather insensitive to the iterative improvement of the reference structures; thus, they may be considered converged with respect to the quality of the reference structure.

It has to be noted that the separation of the translational entropy of the solute is correct without any approximation, whereas the separation of rotational entropy assumes a negligible correlation between the internal motion and the overall rotation of the solute, as discussed by Schäfer et al.[37]

We estimated the configurational entropies for various parts of the DNA double-helical species. In all of the calculations, only non-hydrogen atoms were considered. The error brought about with this assumption may be neglected as we focus on the entropy changes rather than absolute values; this approximation is discussed in refs 27 and 36, where the contribution of the fast motion of hydrogen atoms to configurational entropy is found to be negligible. The cytosine−guanine pairs at the end of the DNA strands were excluded from the analysis. The entropy was estimated for the following parts of the DNA species (Figure 3):

1. Helix: all of the non-hydrogen atoms of the DNA double-helix excluding the outer C and G nucleotides.
2. Backbone: the non-hydrogen atoms of the sugar and phosphate moieties belonging to the A and T nucleotides.
3. Bases: the non-hydrogen atoms of the A and T nucleobases.
4. Base pair: the non-hydrogen atoms of two hydrogen-bonded bases (A and T).
5. Step: the non-hydrogen atoms of two neighboring base pairs.

The fitting of the structure was performed for each part of interest (of those presented in Figure 3) separately, and this procedure prevented an undesired correlation of the obtained entropies with other parts of the system.

As reported previously, the calculated entropy depends on the length of the MD trajectory used for the analysis and should converge with increasing trajectory length. Therefore, for the calculation of entropy, the trajectories were divided to non-

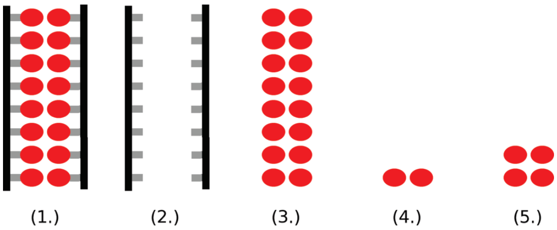


Figure 3. Parts of the DNA double helix for which the configurational entropies were estimated. Red: DNA bases; gray: sugars; black: phosphates. 1. "Helix", 2. "Backbone", 3. "Bases", 4. "Base pair", 5. "Step".

overlapping intervals of varied length, namely, 2, 5, 10, 20, and 60 ns; then, the obtained entropies were averaged for every respective length of the time interval t. Following Harris et al.,[38,41] we extrapolated the entropies to an infinitely long simulation by using an empirical equation (eq 6).

$$S(t) = S_{\text{inf}} - \frac{A}{t^{\text{B}}} \tag{6}$$

The parameters S_{inf}, A, and B in eq 6 were obtained by fitting S_t to the values of entropy obtained from 2, 5, 10, 20, and 60 ns long simulations. S_{inf} represents the estimation of entropy for an infinitely long simulation, whereas no physical interpretation has yet been established for parameters A and B.[41]

Calculation of Interaction Energy. A crucial component of the enthalpy of formation of a molecular complex is the interaction energy. For the two binding modes of ellipticine (Figure 2), the interaction energy of ellipticine and DNA double helix was estimated and averaged along the respective MD trajectory. The interaction energy E_i was calculated for each trajectory frame as the sum of Coulomb and LJ terms:

$$E_i = \sum_i^{\text{int}} \sum_j^{\text{DNA}} \left(\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \frac{C_{12,ij}}{r_{ij}^{12}} - \frac{C_{6,ij}}{r_{ij}^{6}} \right) \tag{7}$$

(The first sum runs over the atoms of the intercalator, the second sum runs over all atoms of DNA, q_i and q_j are atomic charges, ε_0 is the permittivity of vacuum, r_{ij} is the interatomic distance, and $C_{12,ij}$ and $C_{6,ij}$ are LJ pair parameters.) The parameters as well as the atomic charges for DNA were assigned from the AMBER parm99 force field;[44] LJ parameters for ellipticine were taken from the GAFF force field,[48] and the charges for ellipticine were obtained with the RESP fitting technique as described above. An increased cutoff distance of 3.3 nm was applied for both Coulomb and LJ terms in the calculation of interaction enthalpy.

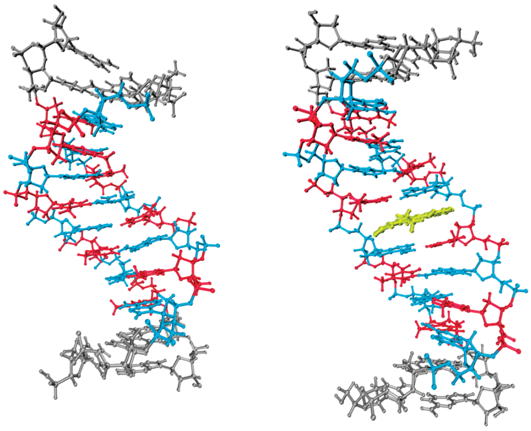


Figure 5. Representative snapshots from the simulation of a DNA double helix with (right) and without (left) the intercalator. Blue: adenine nucleotides, red: thymine nucleotides, gray: cytosine and guanine nucleotides, yellow: ellipticine ("int1" orientation).

Results and Discussion

Simulations. The simulations of bare DNA showed a stable behavior, with all studied sequences. The root-mean-squared deviation (rmsd) calculated with respect to the first frame of the trajectory displays no drift (Figure 4). The same is true about the simulations of DNA...ellipticine complexes with the "int1" orientation. In sequences A and B, however, the simulations of the "int2" orientation switched to "int1" after about 55 and 25 ns, respectively. These simulations were repeated with different starting coordinates and velocities taken from the equilibrated ensemble, resulting in a stable "int2" orientation, in both A and B. In case of sequences C and D, the "int2" orientation remained stable within the simulation time. Representative snapshots taken from simulations of bare DNA as well as of the complex with ellipticine ("int1" orientation) are shown in Figure 5.

During all simulations, the outermost CG pairs assumed a non-Watson−Crick arrangement temporarily (for a few nanoseconds) on several occasions; the occurrence of such events is well-known in the MD simulations of DNA and affects neither

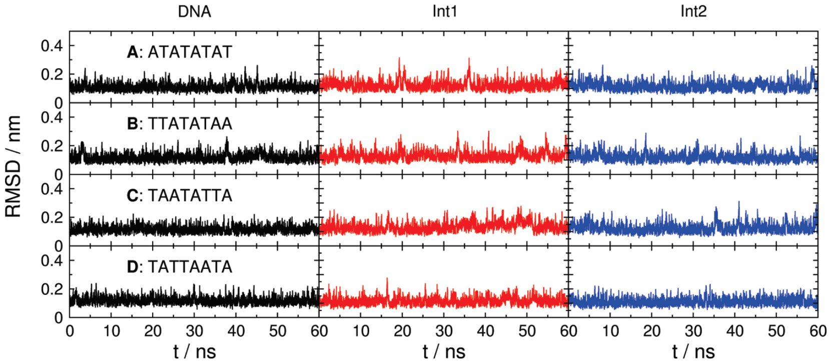


Figure 4. rmsd of non-hydrogen atoms with respect to the first frame of the trajectory.

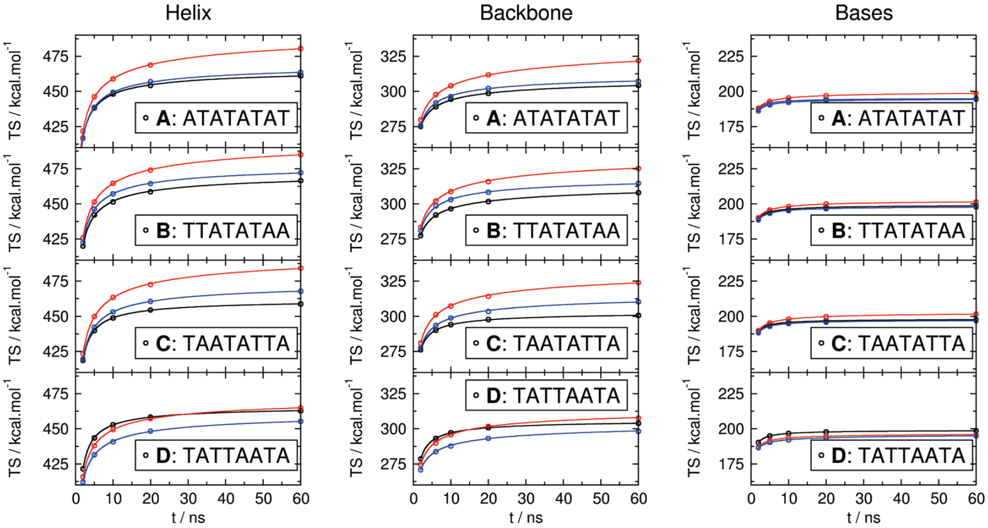


Figure 6. Dependence of calculated configurational entropy on the length of the MD trajectory. Black lines and circles: bare DNA, red: DNA...ellipticine complex with "int1" orientation, blue: DNA...ellipticine complex with "int2" orientation. Note the identical *y*-axis range throughout every column. Circles: entropic contributions to the free energy ($T{\cdot}S$ in kcal·mol^{-1}) calculated at 300 K, solid lines: fitted functions (see the Methods section).

TABLE 2: Change of Configurational Entropy upon Intercalation at 300 K: $T\Delta S$ in kcal·mol^{-1}, Calculated for Various Parts of the Molecular System (See Text or Figure 3 for a Definition)

sequence	A	B	C	D
helix (int1)−helix (DNA)[a]	33.2	27.0	38.3	8.5
helix (int2)−helix (DNA)	3.8	5.6	14.0	−1.4
backbone (int1)−backbone (DNA)	39.3	27.4	38.5	12.6
backbone (int2)−backbone (DNA)	2.2	3.3	16.6	2.5
bases (int1)−bases (DNA)	5.2	3.5	5.5	−2.1
bases (int2)−bases (DNA)	−0.4	−1.1	−0.7	−2.2

[a] Difference of entropies extrapolated to an infinitely long simulation.

the stability of the double helix nor the aim of this study, which is to discuss only the inner AT base pairs. The rarely registered intervals of trajectory with a non-Watson−Crick arrangement of the inner AT pairs were excluded from the analysis. These intervals were identified by calculating the stretch helical parameter for all AT base pairs as defined by Olson et al.[55,56] The trajectory frame was excluded from the analysis if the stretch parameter was outside of the interval of regular fluctuations (usually larger than 0.06 nm).

Entropy of the Entire Double Helix. The configurational entropies multiplied by the temperature (300 K) for the four different AT-rich DNA sequences are plotted in Figure 6 (left panel), and the entropy changes for an infinite simulation are summarized in Table 2. An extended version of Table 2 containing absolute entropies is provided in the Supporting Information.

The convergence of entropies is satisfactory and is slightly better for bare DNA (black lines) than for the DNA...ellipticine complexes (red and blue lines). The fitting of a function in eq 6 to the obtained data resulted in a correlation coefficient almost equal to one. The parameter B for the simulations of bare DNA (i.e., without intercalator) reaches a value close to 2/3 as used in ref 38 and discussed in ref 41, whereas it is significantly smaller for the trajectories with the intercalator (ca. 0.45 for "int1" and 0.54 for "int2").

The intercalation of ellipticine in the "int1" orientation (the pyrrole nitrogen oriented toward the major groove, Figure 2) is accompanied by an increase of configurational entropy of the DNA of all studied sequences, although the increase is considerably smaller in the case of D than in A, B, and C (see the extrapolated values in Table 2). The configurational entropy contribution to the free energy change for the "int1" orientation varies in the range of 8−38 kcal·mol^{-1}. For the "int2" orientation (the pyrrole nitrogen oriented toward the minor groove, Figure 2), the calculated entropies are smaller, in the range of −1 to 14 kcal·mol^{-1}; the value is negative for the sequence D.

This indicates that the magnitude of the changes of flexibility of the DNA helix upon intercalation depends not only on the targeted sequence but also on the orientation of the ligand. The increased flexibility is in contrast with the increased rigidity of the DNA helix upon minor groove binding, observed previously.[38,57]

Entropy of the Backbone and of the Nucleobases. The configurational entropies of the backbone and of the bases (see the Methods section) multiplied by the temperature are plotted in Figure 6 (middle and right panels). The values extrapolated for infinite simulation are summarized in Table 2. Obviously, the major part of the change of entropy upon intercalation into all sequences is carried by the sugar−phosphate backbone, while the system of nucleobases exhibits a significantly smaller change of entropy. In the case of sequence D, the change of entropy of the backbone is clearly smaller than in the other sequences, which is also reflected in the smaller change of entropy of the whole DNA ("Helix"), as described in the previous section. In all sequences, the configurational entropy of the backbone increases for both orientations of ellipticine, and the increase is smaller with the "int2" than with "int1".

Unlike the case of backbone, the entropy change of the system of nucleobases is quite small; $T\Delta S$ increases by 3−5 kcal·mol^{-1} for sequences A, B, and C, whereas it decreases by 2 kcal·mol^{-1} for sequence D ("int2" orientation). Again, the changes are more significant for the "int1" orientation of the ligand. This shows that the intercalation induces dramatic changes of the dynamics of the DNA backbone while leaving the dynamics of the

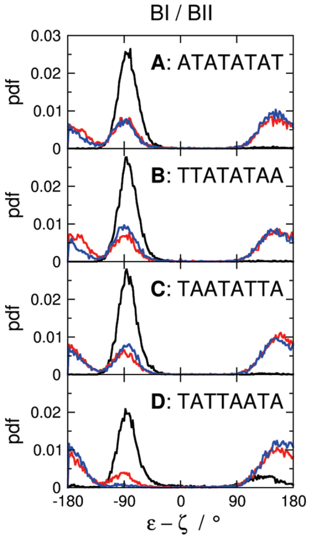


Figure 7. Probability distribution function of the difference of backbone dihedral angles $\varepsilon-\zeta$ calculated for one of the two phosphates in the intercalation site (between the sixth and the seventh base pair). Black: bare DNA, red: "int1" orientation of the ligand, blue: "int2" orientation of the ligand.

TABLE 3: Fractions of Simulation Time Spent in the BI and BII Conformation, by One of the Two Phosphates in the Intercalation Site, in %. Results for Bare DNA ("DNA") as well as for Intercalative Complexes (with Both Orientations, "int1" and "int2")

	A	B	C	D
sequence	BI[a]/BII[b]	BI[a]/BII[b]	BI[a]/BII[b]	BI[a]/BII[b]
DNA	98/2	99/1	99/1	79/21
"int1"	35/65	31/69	28/72	18/82
"int2"	33/67	42/58	33/67	5/95

[a] Integral of the probability distribution function over the interval (−130°, +70°). [b] Integral of the pdf over the interval (0°, +180°) and (−180°, −130°).

nucleobases rather untouched. Upon the creation of the binding site, that is, the separation of two base pairs accompanied by the reorganization of the sugar−phosphate backbone, the backbone experiences an increased flexibility, hence "feeling" the presence of the ligand much more strongly than the base pairs do. This is true for the sequences with the central ATAT tetramer (A, B, and C) but not quite so for the sequence D with the TTAA tetramer, where the possible increase of dynamical flexibility is effectively damped.

To identify the flexibility changes within the sugar−phosphate backbone, we evaluated the difference of backbone dihedral angles ε (C4′−C3′−O3′−P′) and ζ (C3′−O3′−P−O5′). This difference defines the BI and BII conformations of B-DNA as described by Hartmann et al.[58,59] The typical value of $\varepsilon-\zeta$ is −90° for BI and +90° for BII. The probability distributions of $\varepsilon-\zeta$ for one of the phosphates in the intercalation site are plotted in Figure 7, and the fractions of BI and BII conformations are summarized in Table 3. The behavior of both phosphates in an intercalation site is identical (data not shown). The distributions display a single peak at around −90° for all simulations of bare DNA (Figure 7, black lines), indicating the common BI conformation. The situation changes upon the intercalation into sequences A, B, and C, where a bimodal distribution with peaks around −90° and +160° is observed (Figure 7, red and blue lines). So, the phosphate in the intercalation site spends a certain part of the simulation time in the BI conformation, but it undergoes transitions to a BII-like conformation very often (many times per nanosecond). This transition is, however, involved only within the intercalation site, and the other phosphates are unaffected by the intercalation (data not shown).

A different distribution of $\varepsilon-\zeta$ is observed for the sequence D. Upon intercalation, the peak at −90° corresponding to the BI conformation vanishes almost completely, especially with the "int2" orientation of the ligand (Figure 7, sequence D, blue line). This points out that intercalation into the central TTAA tetramer is connected with a complete transition of the phosphates in the intercalation site into a BII-like conformation. According to Hartmann et al., DNA is notably more rigid in a BII conformation than in BI;[59] our finding of the small or even nonexistent increase of configurational entropy upon intercalation into the TTAA tetramer in sequence D might be explained by the conformational transition of the backbone.

A methodological remark has to be made in this place. The quasi-harmonic approximation assumes the fluctuations of atomic coordinates adopting a multivariate normal distribution, which means that there should be only one equilibrium structure of the molecule, or in other words, that each atomic coordinate should fluctuate around a single equilibrium value. Unfortunately, the bimodal distribution of $\varepsilon-\zeta$ dihedral angles indicates that this was most likely not the case, in the simulations of intercalation complexes A, B, and C, and the entropy calculated here was most likely overestimated. Indeed, this becomes clear when comparing the calculated entropy changes upon intercalation with the expected enthalpic cost of the creation of an intercalation site. In our earlier work[25] we estimated this deformation energy of DNA at 20−24 kcal·mol^{-1}. Subtracting the favorable change of entropy (of around 30 kcal·mol^{-1}) from this value, we would obtain a negative free energy, meaning that the DNA double helix would unwind spontaneously, creating an intercalation site; undoubtedly, this would be an absurd conclusion. Whereas it seems to be obvious that the (possibly quite frequent) conformational transitions of phosphates within the intercalation site between the BI- and BII-like states will increase the configurational entropy of DNA as the phosphate groups experience increased conformational freedom, the calculated value of entropy change is clearly overestimated. This finding illustrates that the approach based on quasi-harmonic approximation truly yields merely an upper bound of the configurational entropy, and care must be taken in the analysis of the results, in particular with respect to the possibility of local violation of the involved approximation(s).

Entropy of the Nucleobase Pairs. The configurational entropy calculated for each of the AT pairs (according to the definition in the Methods section) is plotted in Figure 8, left panel. The values obtained for 60 ns trajectories are presented; the calculated entropy converges very quickly with the length of simulation for these systems, reaching 98% of S_{inf} already in a 5 ns simulation. The presented quantity covers the entropy changes within a particular base pair and is unaffected by any correlation with the other base pairs; no information is provided here on how the motion of one base pair affects the motion of another.

The black lines in Figure 8 represent the entropy obtained from the simulations of bare DNA. All of the sequences are palindromic, and so symmetrical lines could be expected. Indeed, the left−right symmetry with respect to the intercalating site

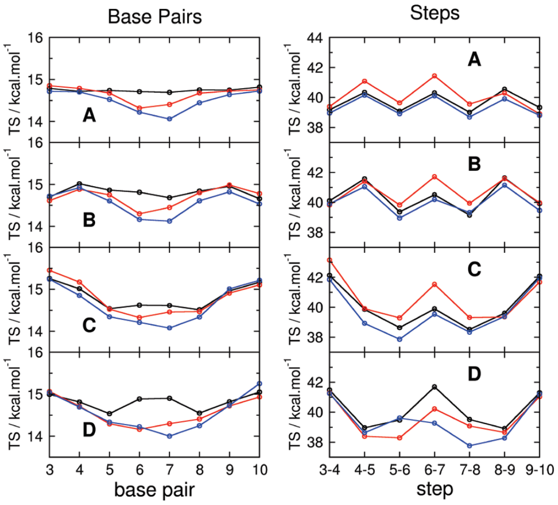


Figure 8. Entropy contributions of “Base pairs” and “Steps” at 300 K—$T\cdot S$ in kcal/mol^{-1}. Black: bare DNA, red: DNA...ellipticine complex with “int1” orientation of the ligand, blue: DNA...ellipticine complex with “int2” orientation of the ligand.

TABLE 4: Configurational Entropy Calculated for the Individual Base Pairs (“Intra-base-pair” Entropy) and the Difference from the Entropy of the System of Bases (“Inter-base-pair” Contribution), as $T\cdot S$ in kcal·mol^{-1} [a]

sequence	A	B	C	D
bases (DNA)[b]	195.3	199.6	198.2	199.1
bases (int1)	200.6	203.0	203.7	197.0
bases (int2)	194.9	198.5	197.6	197.0
bases (int1−DNA)	**5.2**	**3.5**	**5.5**	**−2.1**
bases (int2−DNA)	**−0.4**	**−1.1**	**−0.7**	**−2.2**
base pairs (DNA)[c]	118.0	118.5	118.7	118.6
base pairs (int1)	117.2	117.6	118.4	116.6
base pairs (int2)	116.0	116.5	117.3	116.4
base pairs (int1−DNA)	**−0.8**	**−1.0**	**−0.3**	**−2.0**
base pairs (int2−DNA)	**−1.9**	**−2.0**	**−1.4**	**−2.2**
difference (DNA)[d]	77.4	81.0	79.6	80.6
difference (int1)	83.4	85.5	85.3	80.4
difference (int2)	78.9	82.0	80.3	80.6
difference (int1−DNA)	**6.0**	**4.5**	**5.7**	**−0.2**
difference (int2−DNA)	**1.5**	**0.9**	**0.7**	**0.0**

[a] In bold typeface is presented the change of entropies upon intercalation, to be compared with the data in Table 2. [b] Extrapolated entropies of the system of all base pairs (“Bases”) for an infinitely long simulation. [c] Sum of individual base pair entropies (cf. Figure 8, left panel)—the “intrabase-pair” entropy. [d] Difference of the values “Bases” and “Base Pairs”—an estimate of the “inter-base-pair” entropy.

located between the sixth and the seventh base pair is evident in all sequences.

The sequence A is composed of regularly alternating adenines and thymines, as a result of which each base pair contains the same amount of entropy (resulting in a straight black line). The sequences B, C, and D are more heterogeneous, and thus the course of base-pair entropies along the strand is more complex.

Upon intercalation (red and blue lines), the entropy of the base pairs nearest to the binding site decreases by about 0.5 kcal·mol^{-1} per base pair, in all sequences. The red and blue lines need not be symmetric anymore because of the left−right asymmetry of the ellipticine molecule (Figure 1). However, the asymmetry of the red and blue lines is fairly weak, indicating a weak effect of the asymmetry of the intercalator.

The sum of entropies of the individual base pairs (lines “Base Pairs” in Table 4) is relatively large though, but nearly identical for all DNA sequences studied: it was found in an interval narrower than 1 kcal/mol for the simulations of bare DNA and in an interval of 2.5 kcal/mol for the simulations of intercalation complexes. Such an agreement may be explained simply by the fact that, for every simulation, the calculations are performed on eight identical molecular systems, AT base pairs, and the rest of the DNA is not taken into account. From another point of view, this agreement suggests that the variation of entropy of the system of nucleobases among the studied DNA sequences is hidden exclusively in the motion of hydrogen-bonded base pairs relative to the others, that is, the “interbase-pair” motion, rather than that within one or several particular isolated base pair(s) (“intrabase-pair” motion). Much the same, this intrabase-pair entropy hardly changes upon intercalation, in particular in case of the “int1” binding into sequences A, B, and C, where the touched entropy change is negative and of a magnitude of up to 1 kcal·mol^{-1}.

The sum of entropies of the individual base pairs can be compared with the entropy of the system composed of all of the bases, and the difference between these figures quantifies the thermodynamic role of the interbase-pair motion, see the “difference” lines in Table 4. These values are clearly larger than those obtained for the intrabase-pair motion (“Base Pairs” in Table 4), for the intercalation in the “int1” mode into sequences A, B, and C; no clear trend is seen in the remaining cases. Thus, the increased intrabase-pair conformational flexibility of the DNA double strand contributes favorably to the free energy of intercalation. However, it must be noted that this contribution is tiny in comparison with the change of configurational entropy of the backbone.

Entropy of the Base-Pair Steps. Configurational entropy of the base-pair steps is plotted in Figure 8, right panel. As in the case of the base pairs, the values from 60 ns simulations are considered to be converged and are presented here.

The entropies obtained from the simulations of bare DNA (black lines in Figure 8) are symmetric with respect to the central 6−7 steps, in accordance with the symmetry of the sequences. In all of the sequences, the TA steps “contain” more entropy than the AT steps, which is consistent with the previous observation of the larger flexibility of these steps.[16] The entropic contribution to the free energy change upon intercalation amounts to 1−3 kcal·mol^{-1} per step, depending on the sequence.

The entropy of steps increases upon intercalation of the ligand in orientation “int1” (red lines), with the exception of sequence D, where a decrease of about 1 kcal·mol^{-1} is observed. The entropy remains nearly constant upon intercalation of the ligand in orientation “int2”, again with the exception of sequence D where the entropy decreases by about 1 kcal·mol^{-1}. In general, the shape of the black, red, and blue lines is very similar. A weak asymmetry of entropies is seen for the intercalated DNA, being most apparent at the binding site and decaying with increasing distance from the binding site. Also, it may be inferred from the data in Figure 8 that the effect of the ligand to the flexibility of nucleobases is localized to the intercalation site and its nearest neighborhood.

Enthalpic Changes. The calculated contributions to the changes of enthalpy upon intercalation in both modes “int1” and “int2” are presented in Table 5. These values represent the interaction enthalpy evaluated as the ensemble average of the interaction energy (see the Methods section), and they exhibit

TABLE 5: Interaction Enthalpies of the DNA...Intercalator Complex, in kcal·mol^{-1}

sequence	A	B	C	D
int1	−571.8	−571.3	−573.3	−576.2
int2	−560.1	−560.6	−566.1	−563.0
int2−int1	11.7	10.7	7.2	13.2

good convergence with respect to the length of the simulation (data not shown). The large magnitude of the interaction energies is caused by the strong electrostatic interaction between the cationic intercalator and the negatively charged phosphates; in our analysis, we will concentrate merely on the difference of these interaction energies, as a component of the difference of enthalpy change upon intercalation.

Once we assume that the de/solvation changes of DNA and the ligand are independent of the exact DNA sequence and the orientation of intercalator, the enthalpies in Table 5 together with the configurational entropy changes in Table 2 may be used to assess the affinity of the intercalator to the various sequences as well as the preferred mode of intercalation. As for the difference of reaction enthalpy, the orientation "int1" seems to be preferred to "int2" in all studied sequences, with the difference of contributions to the free energy amounting to about 10 kcal·mol^{-1}. Recalling that the calculated changes of configurational entropy were more favorable for the binding in the "int1" orientation, as well, it may be concluded that this binding mode of ellipticine appears to be thermodynamically more stable than "int2".

This statement seems to be in partial agreement with the report by Elcock et al.[12] who studied the binding of 9-hydroxyellipticine into poly(AT) DNA and obtained a structure of intercalative complex with the pyridine nitrogen oriented into the major groove, however, on the basis of quite short MD simulations. The orientation of the pyrrole nitrogen was not clearly specified, and as discussed in ref 12 the resulting structure was stabilized by the interaction of the hydroxyl group with water in the minor groove—an interaction that cannot occur with ellipticine lacking the hydroxyl group. In addition, the pyridine nitrogen in our simulations tends to be oriented close to the edge of the major groove, creating a hydrogen bond with the O4′ atom of one of the sugars in the backbone occasionally.

Summary

The change of the configurational entropy of four AT-rich double-helical DNA species upon the intercalation of ellipticine was studied. It was shown that the entropy change favors the binding of ellipticine into all of the presented sequences except for one (into sequence D in the "int2" orientation). While the entropy changes are comparable for the sequences A, B, and C, it was found substantially smaller for the sequence D. This may be explained by the appearance of BI-/BII-like conformational transitions. Upon the intercalation of ellipticine, DNA species with the central tetramer TTAA undergoes a conformational change from BI- to BII-like. Consequently, the sugar−phosphate backbone within the intercalation site adopts a more rigid arrangement. An incomplete transition occurs in the DNA species with the central tetramer ATAT, accompanied by an increase of conformational flexibility.

On the basis of the calculated entropic and enthalpic changes, we suggest the "int1" orientation of ellipticine (with the pyrrole nitrogen atom oriented toward the major groove) to be more stable than the "int2" orientation. Taking also previous results of X-ray experiments and MD simulations[7,12,14] into account, our results indicate that the actual binding motif is certainly dependent not only on the chemical identity of the ligand but also on the targeted sequence.

The change of conformational flexibility of the DNA contributes to the binding free energy (by way of configurational entropy) as much as 38 kcal·mol^{-1} (for sequence C, orientation "int1"). Although the presented values constitute the upper bound to the entropy and their overestimation due to the local violation of the quasi-harmonic approximation cannot be ruled out (especially for the central pair of phosphates in sequences A, B, and C), the contribution of configurational entropy must still be considered crucial with respect to the magnitude of the total binding free energies of common noncovalent biomolecular complexes of up to 15 kcal·mol^{-1}, for ellipticine for instance about 6 kcal·mol^{-1} as measured by Kohn et al. in ref 6. Hence, the changes of configurational entropy should be properly accounted for in studies of ligand binding processes.

We have observed that the major part of the increase of configurational entropy comes from the increased flexibility of the sugar−phosphate backbone, whereas the entropy change of the system of nucleobases is much smaller. Again, the properties of the central tetramers TTAA and ATAT differ substantially in this respect.

The magnitude of conformational flexibility changes assessed in this work manifest the important role of the configurational entropy term in the free energy estimators as well as in the scoring functions, comprehending not only the changes of flexibility of the ligand but mainly that of the target. A correct account for this effect does not need to be straightforward, especially in the case of flexible molecules like proteins where the quasi-harmonic approximation may be invalid.

Acknowledgment. We are grateful to Dr. Filip Lankaš for stimulating discussions. An important part of the calculations was performed on the "Chinook" supercomputer at the Environmental Molecular Sciences Laboratory of the Pacific Northwest National Laboratory. This work was supported by the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic (Z40550506), the Ministry of Education, Youth and Sports of the Czech Republic (LC512 and MSM6198959216), the Czech Science Foundation (203/06/1727), and the Korea Science and Engineering Foundation (World Class Univ. program: R32-2008-000-10180-0). The support of the Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged.

Supporting Information Available: Structures of all studied DNA double helices and DNA...ellipticine complexes provided as PDB files. An extended version of Table 2 containing absolute values of entropy estimates for infinite simulation times is also available. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Martínez, R.; Chacón-García, L. *Curr. Med. Chem.* **2005**, *12*, 127–151.

(2) Brana, M. F.; Cacho, M.; Gradillas, A.; de Pascual-Teresa, B.; Ramos, A. *Curr. Pharm. Des.* **2001**, *7*, 1745–1780.

(3) Kopka, M. L.; Goodsell, D. S.; Baikalov, I.; Grzeskowiak, K.; Cascio, D.; Dickerson, R. E. *Biochemistry* **1994**, *33*, 13593–13610.

(4) Wemmer, D. E.; Dervan, P. B. *Curr. Opin. Struct. Biol.* **1997**, *7*, 355–361.

(5) Chaires, J. B. *Curr. Opin. Struct. Biol.* **1998**, *5*, 314–320.

(6) Kohn, K. W.; Waring, M. J.; Glaubiger, D.; Friedman, A. *Cancer Res.* **1975**, *35*, 71–76.

(7) Jain, S. C.; Bhandary, K. K.; Sobell, H. M. *J. Mol. Biol.* **1979**, *135*, 813–840.

(8) Auclair, C. *Arch. Biochem. Biophys.* **1987**, *259*, 1–14.

(9) Dodin, G.; Schwaller, M. A.; Aubard, J.; Paoletti, C. *Eur. J. Biochem.* **1988**, *176*, 371–376.

(10) Bailly, C.; Ohuigin, C.; Rivalle, C.; Bisagni, E.; Hénichart, J. P.; Waring, M. J. *Nucleic Acids Res.* **1990**, *18*, 6283–6291.

(11) Behravan, G.; Leijon, M.; Selhlstedt, U.; Nordén, B.; Vallberg, H.; Bergamn, J.; Gräslund, A. *Biopolymers* **1994**, *34*, 599–609.

(12) Elcock, A. H.; Rodger, A.; Richards, W. G. *Biopolymers* **1996**, *39*, 309–326.

(13) Stiborová, M.; Sejbal, J.; Bořek-Dohalská, L.; Aimová, D.; Poljaková, J.; Forsterová, K.; Rupertová, M.; Wiesner, J.; Hudeček, J.; Wiessler, M.; Frei, E. *Cancer Res.* **2004**, *64*, 8374–8380.

(14) Canals, A.; Purciolas, M.; Aymami, J.; Coll, M. *Acta Crystallogr.* **2005**, *61*, 1009–1012.

(15) Dervan, P. B. *Bioorg. Med. Chem.* **2001**, *9*, 2215–2235.

(16) Lankaš, F.; Šponer, J.; Langowski, J.; Cheatham, T. E. *Biophys. J.* **2003**, *85*, 2872–2883.

(17) Jelesarov, I.; Bosshard, H. E. *J. Mol. Recognit.* **1999**, *12*, 3–18.

(18) Leavitt, S.; Freire, E. *Curr. Opin. Struct. Biol.* **2001**, *11*, 560–566.

(19) Haq, I.; Ladbury, J. *J. Mol. Recognit.* **2000**, *13*, 188–197.

(20) Haq, I.; Ladbury, J. E.; Chowdhry, B. Z.; Jenkins, T. C.; Chaires, J. B. *J. Mol. Biol.* **1997**, *271*, 244–257.

(21) Leng, F.; Chaires, J. B.; Waring, J. *Nucleic Acids Res.* **2003**, *31*, 6191–6197.

(22) Mukherjee, A.; Lavery, R.; Bagchi, B.; Hynes, J. T. *J. Am. Chem. Soc.* **2008**, *130*, 9747–9755.

(23) Singh, S. B.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 3267–3271.

(24) Řeha, D.; Kabeláč, M.; Ryjáček, F.; Šponer, J.; Šponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.

(25) Kubař, T.; Hanus, M.; Ryjáček, F.; Hobza, P. *Chem.—Eur. J.* **2006**, *12*, 280–290.

(26) Stone, M. J. *Acc. Chem. Res.* **2001**, *34*, 379–388.

(27) Zhou, H. X.; Gilson, M. K. *Chem. Rev.* **2009**, *109*, 4092–4107.

(28) Chang, C. A.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.

(29) Schneider, G. *Nat. Rev. Drug Discov.* **2010**, *9*, 273–276.

(30) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 238–250.

(31) Singh, N.; Warshel, A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1705–1723.

(32) Fanfrlík, J.; Bronowska, A.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666−12678.

(33) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 332–335.

(34) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

(35) Andricioaei, I.; Karplus, M. *J. Chem. Phys.* **2001**, *115*, 6289–6292.

(36) Schäfer, H.; Mark, A. E.; van Gunsteren, W. F. *J. Chem. Phys.* **2000**, *113*, 7809–7817.

(37) Schäfer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 45–56.

(38) Harris, S. A.; Gavathiotis, E.; Searle, M. S.; Orozco, M.; Laughton, C. A. *J. Am. Chem. Soc.* **2001**, *123*, 12658–12663.

(39) Rueda, M.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **2005**, *127*, 11690–11698.

(40) Dolenc, J.; Baron, R.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. F. *Biophys. J.* **2006**, *91*, 1460–1470.

(41) Harris, S. A.; Laughton, C. A. *J. Phys.: Condens. Matter* **2007**, *19*, 076103.

(42) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, CA, 2008.

(43) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindhal, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(44) Cheatham, T. E.; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.

(45) Peréz, A.; Marchan, I.; Svozil, D.; Šponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.

(46) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossiv, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(49) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

(50) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(51) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(52) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(53) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(54) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(55) Olson, W. K.; Bansal, M.; Burley, S. K.; Dickerson, R. E.; Gerstein, M.; Harvey, S. C.; Heinemann, U.; Lu, X. J.; Neidle, S.; Shakked, Z.; Sklenar, H.; Suzuki, M.; Tung, C. S.; Westhof, E.; Wolberger, C.; Berman, H. M. *J. Mol. Biol.* **2001**, *313*, 229–237.

(56) Lu, X. J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

(57) Wang, H.; Laughton, L. A. *Methods* **2007**, *42*, 196–203.

(58) Hartmann, B.; Piazzola, D.; Lavery, R. *Nucleic Acids Res.* **1993**, *21*, 561–568.

(59) Heddi, B.; Foloppe, N.; Bouchemal, N.; Hantz, E.; Hartmann, B. *J. Am. Chem. Soc.* **2006**, *128*, 9170–9177.

JP1019153

Table 2. Change of configurational entropy upon intercalation at 300 K – T.D*S* in kcal/mol^{-1}., calculated for various parts of the molecular system.

Sequence	A	B	C	D
Helix (DNA)[a]	468.0	474.8	462.6	466.6
Helix (int1)	501.2	501.7	501.0	475.1
Helix (int2)	471.8	480.4	476.6	465.2
Helix (int1-DNA)	33.2	27.0	38.3	8.5
Helix (int2-DNA)	3.8	5.6	14.0	–1.4
Backbone (DNA)	313.3	319.3	303.8	306.9
Backbone (int1)	352.6	346.7	342.3	319.5
Backbone (int2)	315.3	322.6	320.4	309.4
Backbone (int1-DNA)	39.3	27.4	38.5	12.6
Backbone (int2-DNA)	2.2	3.3	16.6	2.5
Bases(DNA)	195.3	199.6	198.2	199.1
Bases (int1)	200.6	203.0	203.7	197.0
Bases (int2)	194.9	198.5	197.6	197.0
Bases (int1-DNA)	5.2	3.5	5.5	–2.1
Bases (int2-DNA)	–0.4	–1.1	-0.7	–2.2

[a] extrapolated entropies for infinitely long simulation

C

Publication 2 – HIV-1 Inhibitors

THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

Ligand Conformational and Solvation/Desolvation Free Energy in Protein−Ligand Complex Formation

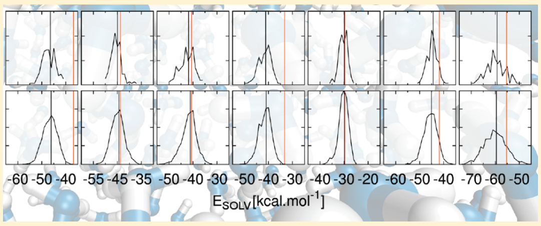
Michal Kolář,[†] Jindřich Fanfrlík,[†] and Pavel Hobza*,[†,‡,§]

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, Flemingovo nam. 2, 166 10 Prague, Czech Republic

[‡]Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

[§]Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

S *Supporting Information*



ABSTRACT: In this study, an extensive sampling of the conformational space of nine HIV-1 protease inhibitors was performed to estimate the uncertainty with which a single-conformation scoring scheme approximates the ligand−protein binding free energy. The SMD implicit solvation/desolvation energy and gas-phase PM6-DH2 energy were calculated for a set of 1600 conformations of each ligand. The probability density functions of the energies were compared with the values obtained from the single-conformation approach and from a short *ab initio* molecular dynamics simulation. The relative uncertainty in the score within the set of nine inhibitors was calculated to be 3.5 kcal·mol^{-1} and 2.7 kcal·mol^{-1} for the single-conformation and short dynamics, respectively. These results, though limited to the consideration of flexible ligands, provide a valuable insight into the precision of rigid models in the current computer-aided drug design.

1. INTRODUCTION

Computer-assisted drug design represents an attractive and useful tool for pharmaceutical research, and the main advantage of the procedure is the expected reduction of the number of systems that should be synthesized. The modeling of the formation of the protein−ligand (P−L) complex from the free subsystems in a water environment,

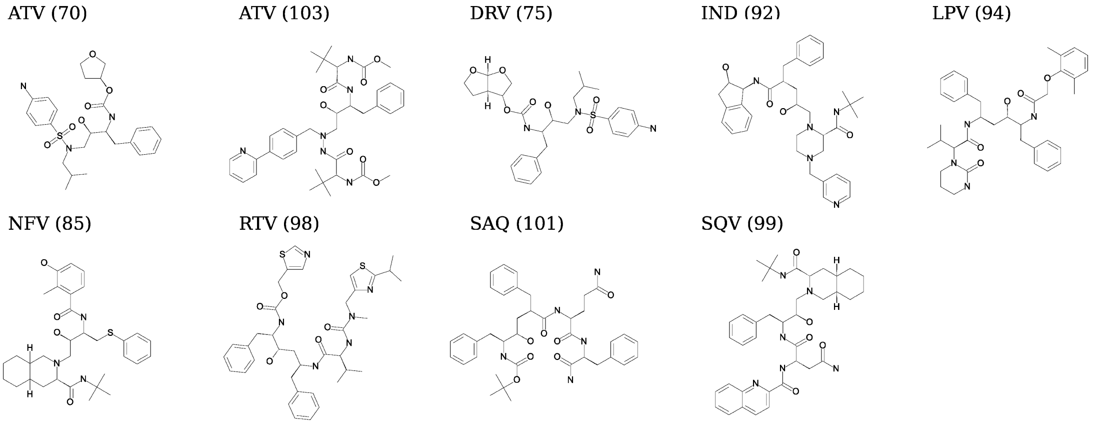
$$P_w + L_w \leftrightarrow (P - L)_w \qquad (1)$$

represents a crucial step in the drug-design process.[1] The aim of the theoretical description is the evaluation of the binding free energy, which is expected to be directly proportional to the ligand potency.[2] The evaluation of the absolute values of the binding free energies is impractical. The relative binding free energies for similar ligands acting on the same target can be estimated by using thermodynamic integration[3,4] or free-energy perturbation techniques.[4,5] The use of these advanced molecular dynamics methods for different ligands acting on different proteins is computationally demanding and thus limited, hence other, simpler procedures should be applied. To save computer resources, often a single-conformation approach is adopted, which means that the flexibility of the object (target, ligand or both) is neglected. Especially for high-throughput studies, the flexibility issues are beyond the limit.[6−8] Recently, we have introduced a novel protein−ligand scoring procedure based on a semiempirical quantum mechanical (SQM) Hamiltonian.[9] Here the score, which approximates the binding free energy and stands for a measure of the ligand affinity, is constructed as a sum of various contributions:[9]

1 the binding enthalpy of the P−L complex in a water environment;
2 the solvation/desolvation free energy of a ligand and protein;
3 the deformation energies of the ligand and protein; and
4 the change of the entropy accompanying the P−L complex formation in a water environment.

All of the contributions are important, and none of them can be neglected. Most attention is paid to the evaluation of the first two energies, and in the majority of cases the empirical potentials or even their simplifications are used.[8,10,11] Their main drawback consists of their neglecting the quantum effects (proton and electron transfer, description of the halogen bond etc.), which is, however, correctly covered by our SQM PM6-DH2 method. Another important feature of our new score is that every physical term is calculated using the most accurate method available. The score is thus constructed as a sum of the PM6-DH2 interaction enthalpy,[12,13] changes in the SMD solvation and PM6-DH2 deformation energies[14] and the empirical-force-field-based vibrational entropy change. No adjustable empirical

Received: January 31, 2011
Revised: March 22, 2011
Published: April 05, 2011


ATV (70) ATV (103) DRV (75) IND (92) LPV (94)

NFV (85) RTV (98) SAQ (101) SQV (99)

Figure 1. The chemical formulas of the HIV-1 PR inhibitors with their abbreviations. The number of atoms is provided in the parentheses.

parameters for the particular energetic terms and for specific ligands are used as was the case, for instance, in references 15−17.

This scoring approach was successfully applied on two series of diverse inhibitors, namely, HIV-1 protease (PR)[9] and CDK2 kinase inhibitors.[18] In both studies, the interaction enthalpies of the P−L complexes represent the dominant terms, and, owing to the reliable PM6-DH2 technique, we expect these to be sufficiently accurate. (For the twenty-two complexes included in the S22 data set, the PM6-DH2 method provides interaction energies within 1 kcal·mol^{-1} of the benchmark CCSD(T) values.)[13]

In the case of evaluation of the solvation/desolvation free energies, the situation is, however, different. In both studies mentioned above, the desolvation energies of the ligands were very large, comparable to the interaction enthalpies, but with the opposite sign. While the interaction enthalpies are negative and thus encourage binding, the ligand desolvation free energies are positive and thus oppose the binding. The main problem here, however, originates not in the choice of the solvation model but in the choice of the ligand structure used for the evaluation of the change of the solvation free energy.

The solvation free energy is quite well-defined for rigid molecules such as benzene or methane. It represents the free energy change connected with the transfer of the molecule from the gas phase (vacuum) into bulk water. For a flexible molecule, the interpretation of the solvation free energy is not straightforward. In this case, we cope with an ensemble of distinct conformations, and the solvation energy calculated for a single conformation stands for the free energy of the vacuum−water transfer under the assumption that the single conformation represents an equilibrium structure of the molecule in both the vacuum and solvated states. It is, however, not very clear to what extent this assumption is valid.

Our score contains a term which describes a change of the solvation free energy upon ligand binding. Apart from the protein solvation energy change, this is constructed as the difference of the solvation energy of a ligand conformation in water and of the ligand conformation restrained by protein surroundings. The conformation in a water environment is often approximated by a structure taken from the P−L complex that is optimized with an implicit solvent.[9,17] More reliable ligand conformations are expected from the molecular dynamics (MD) simulations followed by a gradient optimization (quenching technique). In the case of rigid ligands, the optimization with an implicit solvent model is justified. However, for a flexible ligand, the latter approximation seems to be better suited.

In the present paper, we have reexamined the choice of the ligand conformation for which the solvation energy is computed by performing a standard MD simulation and by calculating the conformational energy. We selected complexes of HIV-1 protease with nine inhibitors which had been briefly studied in our previous paper. All of the inhibitors considered are very flexible. Here, the conformational energy denotes the sum of the solvation energy calculated for a particular ligand conformation and of the gas phase electronic energy which also relates to the particular ligand conformation. The difference of the electronic energies between the two different conformations would be called "deformation energy" and the difference of the solvation energies between the two conformations would be called the "change in solvation energy". When considering the P−L binding, one of the two conformations represents a bonded state in a protein environment and the other represents the free state in water.

The molecular dynamics simulations of the nonstandard residues with an empirical potential usually face a problem of partial charges. In MD, the concept of the point charges is claimed to be essential even though the results might depend on the choice of the atomic partial charges. Several publications have tackled this issue.[19−21] Throughout the study, the General Amber Force Field (GAFF)[22] was employed. The designers of the force field recommend partial charges evaluated on the bases of the RESP technique[23] or calculated at the AM1-BCC level of theory.[24,25] We chose the RESP charges since the AM1-BCC are parametrized to reproduce the RESP charges, as a result of which the RESP charges should be more reliable. However, the charges of both methods might be conformationally dependent, and it can be expected that this dependence increases with the increasing number of possible conformers. To eliminate any possible dependency, we sampled the conformations with ten different charge sets.

2. METHODS

We have studied the conformational energies of nine HIV-1 protease inhibitors, namely, amprenavir (APV) 1HPV,[26] atazanavir (ATV) 2AQU,[27] darunavir (DRV) 1T3R,[28] indinavir (IDV) 2BPX,[29] lopinavir (LPV) 1MUI,[30] nelfinavir (NFV) 1OHR,[31] ritonavir (RTV) 1HXW,[32] Boc-Phe-Psi[(S)-CH(OH)CH2NH]-Phe-Gln-Phe-NH2 (SAQ) 1IIQ[33] and saquinavir (SQV) 3CYX.[34] All nine ligands (see Figure 1) were considered to be neutral, which is consistent with our previous work.[9] The biological activities of the ligands are presented in references 35−39.

The estimations of the conformational energies of nine HIV PR inhibitors were performed in several steps: presampling, sampling, and energy estimation. We plotted the probability density functions of the energies and calculated their mean values. The probability density function describes how likely is to find the conformation with such an energy. Since called "probability", it is normalized to yield 1 when integrated.

The probability density functions of the energies were compared with the results obtained from a short MD simulation with a PM6-DH2 potential and with a single value calculated with the ligand conformation presented in the P−L complex optimized with an implicit solvent. In that sense, the mean values of the probability density functions are considered as the "correct" ones (i.e., not suffering from the single-conformation approximation) and the comparison of the values obtained by other protocols is presented with respect to them.

2.1. Presampling. The P−L complex's experimental geometry was optimized at the PM6-DH2 level. The grid of the electrostatic potential (ESP) points was calculated around the bare ligand structure on the HF/6-31G* level.[40,41] The partial charges were fitted onto the grid according to the RESP methodology; typically about 8000 grid points were used for the fit. The bond, angle, torsion and atomic Lennard-Jones parameters of each of the nine HIV PR inhibitors were assigned from the GAFF force field using the Antechamber program from the Amber program package[42] with the default setup.

Each ligand was surrounded by TIP3P water molecules[43] in a cubic periodic box. The distance of the ligand from the edge of the box was 1 nm, which resulted in approximately 1500 water molecules in the box. A short minimization of the ligand and water molecules was performed to avoid any possible close contacts. The system was heated during a 50 ps simulation with the box volume kept constant, which was followed by a 200 ps equilibration at a temperature of 300 K and under a pressure of 1 bar. A Berendsen thermostat and barostat were employed.[44] The production consisted of a 200 ps simulation at a temperature of 700 K and under a pressure of 1 bar. The time step of 1 fs was used, and the structure of the ligand was saved every 20 ps.

For each ligand, this yielded ten structures. Owing to the high temperature, a variety of conformations was visited during rather short MD simulations. It may be a question if the GAFF force field, originally proposed for simulations at 300 K, is well-behaved for simulations at significantly elevated temperature. However, the conformations obtained at 700 K need not represent the dynamics accurately here, since they only serve as the initial points for the sampling, which was indeed done at 300 K. The variability of the ten conformations is demonstrated by a root-mean-square deviation from the starting structure, which was typically 0.35 nm.

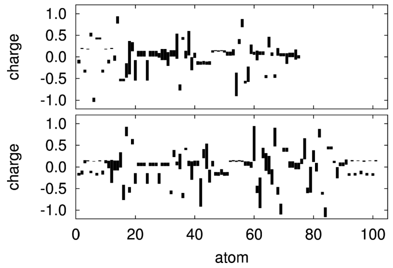


Figure 2. The variations of the atomic partial charges. For each atom, the range between the minimum and maximum charge within the ten charge sets is plotted. The ligand with the most negative binding free energy (DRV) and the ligand with the least negative binding free energy (SAQ)[35−39] were chosen for illustration. The complete set of ligands is available in the Supporting Information.

2.2. Sampling. For each ligand, ten conformations were used for a reevaluation of the RESP charges. Each ligand structure was optimized at the PM6-DH2 level with the COSMO implicit solvation model,[45] the HF/6-31G* ESP points were calculated and the partial charges were fitted onto them. The other parameters were taken from the GAFF force field as mentioned above.

The water box preparation and equilibration protocol was the same as described in the presampling section. Hereafter, ten 40 ns MD simulations of ligands with different RESP charge sets were performed for each ligand at a temperature of 300 K and under a pressure of 1 bar. A Nosé−Hoover thermostat[46,47] and Parrinello−Rahmann barostat[48] were used to obtain the correct isobaric−isothermal ensemble. The ligand structures were saved every 250 ps, which yielded 160 structures for each charge set. For all of the MD simulations, the Gromacs program package was used.[49]

2.3. SMD and Gas Phase Energies. Each structure was optimized at the PM6-DH2 level with the COSMO implicit solvation model until the convergence criteria (the energy difference between the two consecutive steps lower than 6 × 10^{-3} kcal·mol^{-1} and a maximal gradient lower than 1.2 kcal·mol^{-1}·A^{-1}) were satisfied. With the final structure, the gas-phase PM6-DH2 energy (denoted E_{vac}) and solvation SMD/HF/6-31G* energy (denoted E_{SMD}) were calculated. The sum of the two respective energies is presented as the conformational energy (denoted E_{conf}). In total, 1600 conformations for each of the nine HIV PR inhibitors were evaluated. The normalized probability density functions of the solvation, gas-phase electronic and conformational energies were calculated.

2.4. PM6-DH2 Optimization and Quenching. The ligand coordinates were taken from the PM6-DH2 optimized P−L complex and reoptimized with the implicit COSMO model at the PM6-DH2 level. The solvation SMD, gas-phase PM6-DH2 and conformational energies—the sum of the previous two—were calculated on the final structure. These values are referred to as the "PM6-DH2 optimization" energies.

The PM6-DH2 simulations were performed with the implicit COSMO water model. A temperature of 500 K was kept constant by an Andersen thermostat.[50] The total simulation time was 50 ps with a time step of 1 fs. The ligand coordinates were saved every picosecond, after which they were optimized at the PM6-DH2 level with the COSMO water model. On the final

Table 1. The Standard Deviations (std) of the Mean Values of the Probability Density Functions[a]

ligand	APV	ATV	DRV	IND	LPV	NFV	RTV	SAQ	SQV
std(E_{SMD})	0.76	0.44	0.79	1.16	0.53	0.58	0.59	0.82	2.03
std(E_{vac})	2.03	0.45	1.29	3.98	1.02	0.89	0.69	0.81	3.00
std(E_{conf})	1.37	0.44	0.55	3.57	0.87	1.10	0.54	0.63	1.74

[a] E_{SMD} stands for SMD solvation energy, E_{vac} stands for gas-phase PM6-DH2 energy, and E_{conf} stands for conformational energy, the sum of the previous two. The standard deviations are calculated for the set of ten mean values obtained from ten 40 ns long simulations. The simulations differ in the charge set describing the ligand and the initial conformation of the MD. All of the values are in kcal·mol^{-1}.

structures, the E_{SMD}, E_{vac} and E_{conf} were calculated and are referred to as the "PM6-DH2 quench" energies. All of the PM6-DH2 optimizations were performed with the same convergence criteria (mentioned above).

For the PM6-DH2 and SMD calculations, Gaussian09[51] and Mopac[52] programs were used.

3. RESULTS AND DISCUSSIONS

3.1. MD Simulations. For each ligand, we performed a series of MD simulations which differed in the set of atomic partial charges. The charges were calculated for ten different conformations (see Methods). Figure 2 shows the range of the charges for two ligands. The difference between the minimal and maximal charge within the set is plotted for each atom. About 10% of the atoms usually embody a large conformational dependence while the rest remain quite independent. The largest variation reaches 0.802 e for a carbon atom in one of the SAQ carbonyl groups. This extremely large variation is, however, not accompanied by a comparably large change in the conformational energy (see below). We expect the lower variations of the charges of the other atoms to compensate for the rarely occurrences of large variations. It is surprising that the conformational energy is not very sensitive to the particular charge distribution. Part of the conformational energy insensitivity might also be "hidden" in the subsequent geometry optimization, which is always done at the semiempirical PM6-DH2 level with a relatively correct charge distribution and which might buffer the inaccuracy of the RESP charge models.

All of the simulations were stable; the fluctuations in temperature and pressure were within the normal range. In total, 400 ns for each ligand were simulated. We believe that the simulations sufficiently converge with respect to the mean values of particular energies. The mean values of the 40 ns long simulations differ only slightly. Their standard deviations are summarized in Table 1. The entire probability density functions of darunavir (DRV) are shown in Figure 3; the probability density functions of all of the ligands are available in the Supporting Information.

The standard deviations from Table 1 might serve as a measure of two characteristics: first, the convergence of the simulations, and second, the conformational dependence of the RESP charges. The fact that some atomic partial charges vary significantly seems to be quite unimportant in the sense of the SMD solvation energy as well as the vacuum PM6-DH2 energy. The probability density functions of all of the ten simulations (Figure 3, DRV) are normal-like and localized with very similar mean values. One has to bear in mind that there were ten different starting conformations differing in partial charges on the atoms. In the case of DRV, the RMSDs of the heavy atoms with respect to the conformation 0 were within the range of 0.171 to 0.384 nm.

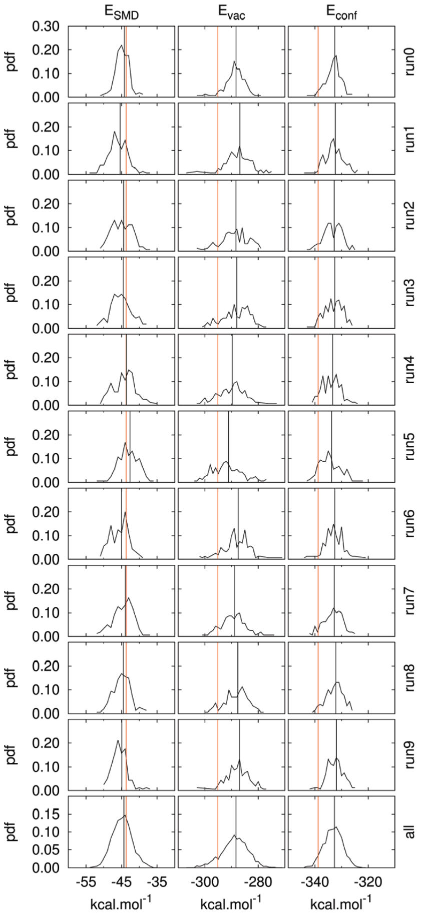


Figure 3. The probability density functions (pdf) of the SMD solvation energy (E_{SMD}), gas-phase PM6-DH2 energy (E_{vac}) and their sum, conformational energy (E_{conf}). The pdfs for the different charge sets calculated for ten ligand conformations are plotted. The pdfs are constructed from the set of 160 values of each particular energy. The 1600 values of all of the runs were used for the last row ("all"). The black vertical line indicates the mean value of pdf; the orange lines indicate the values from the PM6-DH2 quench. The plots describing the most potent drug, darunavir (DRV), were chosen for illustration. All of the ligands are available in the Supporting Information.

3.2. Energies. The probability density functions are plotted in Figure 4. They were calculated for the set of 1600 values obtained by the extensive MD sampling with an empirical potential (see Methods). This does not, however, mean that the energies are based on an empirical potential. The minimization at the PM6-DH2 level ensures the relaxation of the snapshot from the empirical MD onto a semiempirical potential energy surface (PES). The conformation then represents the nearest semiempirical minimum of the empirical PES.

The mean values of the probability density functions in Figure 4 are also summarized in Tables 2 and 3 together with

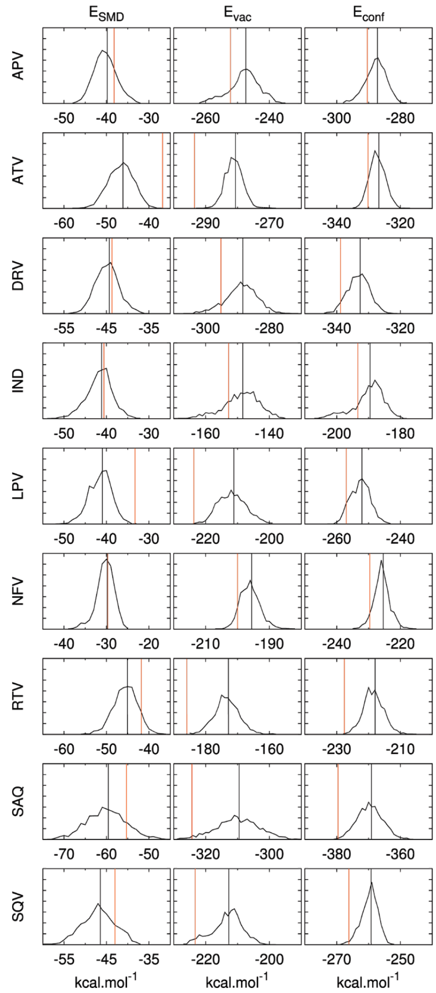


Figure 4. The probability density functions (pdfs) of the SMD energy (E_{SMD}), the gas-phase PM6-DH2 energy (E_{vac}) and their sum (E_{conf}). The pdfs are calculated for each ligand from the set of 1600 conformations obtained from the ten MD simulations with the various charge sets. Note the same range of *x*-axes (in kcal·mol^{-1}). The black vertical lines indicate the mean value of pdf; the orange lines indicate the values from the PM6-DH2 quench.

the PM6-DH2 quench and PM6-DH2 optimization values, respectively. In Figure 4, the mean values are plotted as vertical black lines, whereas the orange lines represent the values from the PM6-DH2 quench. For each ligand and each energy, we calculated the mean value and the standard deviation over the set of ten simulations.

Omitting the flexibility of the ligand, i.e. considering those values not obtained from extensive conformation sampling, we introduce an error. Here, we present the errors connected genuinely with the single-conformation approximation of various extents (PM6-DH2 optimization and PM6-DH2 quench) when compared with MD sampling, which is indeed multiple-conformation approach.

Table 2. The Distinctions between the Extensive MD Sampling and the PM6-DH2 Optimization[a]

ligand	APV	ATV	DRV	IND	LPV	NFV	RTV	SAQ	SQV	avg	std
ΔE_{SMD}	3.6	1.9	3.1	−4.9	3.9	−6.9	2.3	2.2	−1.3	0.4	3.9
ΔE_{vac}	0.8	−0.7	−1.7	11.3	−1.3	2.9	−4.3	2.9	−0.1	1.1	4.5
ΔE_{conf}	4.4	1.3	1.4	6.5	2.6	−4.0	−2.0	5.1	−1.3	1.5	3.5

[a] The differences E(optim) − E(MD) are shown in kcal·mol^{-1}. The averages (avg) and standard deviations (std) are calculated over the set of ligands. ΔE_{SMD} stands for the difference of the SMD solvation energies, ΔE_{vac} of the gas-phase PM6-DH2 energies and ΔE_{conf} of the conformational energies. The absolute values of the energies are provided in the Supporting Information.

Table 3. The Distinctions between the Extensive MD Sampling and the Short PM6-DH2 Quench[a]

ligand	APV	ATV	DRV	IND	LPV	NFV	RTV	SAQ	SQV	avg	std
ΔE_{SMD}	1.6	9.3	0.7	0.5	7.6	0.2	3.3	4.2	3.5	3.4	3.2
ΔE_{vac}	−4.8	−12.8	−6.8	−4.5	−12.5	−4.4	−13.0	−14.8	−10.6	−9.4	4.2
ΔE_{conf}	−3.2	−3.5	−6.2	−4.0	−4.9	−4.3	−9.7	−10.6	−7.1	−5.9	2.7

[a] The difference of E(quench) − E(MD) is shown in kcal·mol^{-1}. The averages (avg) and standard deviations (std) are calculated over the set of ligands. ΔE_{SMD} stands for the difference of the SMD solvation energies, ΔE_{vac} of the PM6-DH2 gas-phase energies and ΔE_{conf} of the conformational energies. The absolute values of the energies are provided in the Supporting Information.

For the set of nine HIV PR inhibitors, the average error in the SMD solvation energy of the PM6-DH2 optimization (see Table 2) is 0.4 kcal·mol^{-1} with a standard deviation of 3.9 kcal·mol^{-1}. The average of E_{vac} is 1.1 kcal·mol^{-1} with a standard deviation of 4.5 kcal·mol^{-1}. The sum of ΔE_{SMD} and ΔE_{vac} increases the average to 1.5 kcal·mol^{-1} while it decreases the standard deviation to 3.5 kcal·mol^{-1}.

The aim of the scoring is to rank the ligands according to their binding free energy or, in other words, on the bases of the score to distinguish the effective binders from the weak binders or nonbinders. In that sense, the average values of the errors (i.e., absolute errors) are rather unimportant, because they represent an average shift of all of the ligands. For a correct ranking, one tries to minimize the relative shift between the ligands, as a consequence of which the standard deviations (relative errors), which tell us with what uncertainty the ligands might be sorted, deserve greater attention.

The PM6-DH2 optimization yields structures which are similar to those in the P−L complex. This results in a quite small absolute error with a significant relative error. What seems to be important is the fact that the differences between the MD sampling and PM6-DH2 optimization are not of the same "sign". Undesirably, the ΔE_{conf} is negative for NFV, RTV and SQV, whereas it is positive for the rest of the ligands. That fact is reflected in the standard deviation of ΔE_{conf} being 3.5 kcal·mol^{-1}.

Table 3 and Figure 4 show that even a short PM6-DH2 quench (MD followed by geometry optimization, see Methods) can improve the results reasonably. In the case of ΔE_{SMD}, the error is systematically shifted (cf. Figure 3) for all of the ligands. When comparing the ΔE_{SMD} energies from the MD sampling and the PM6-DH2 quench, the former are more negative by an average value of 3.4 kcal·mol^{-1}, with the standard deviation being 3.5 kcal·mol^{-1}. The E_{vac} are shifted in the opposite direction by an average value of -9.4 kcal·mol^{-1}, with a standard deviation of 4.2 kcal·mol^{-1}. These shifts are partially compensated for, which results in an average ΔE_{conf} of -5.9 kcal·mol^{-1}, with the standard deviation being 2.7 kcal·mol^{-1}.

The range of the scores for the series of HIV PR binders was about 50 kcal·mol^{-1} (from -15 to $+35$ kcal·mol^{-1}).[9] A similar difference in the score between the strong and weak binders was also obtained for the CDK2 ligands.[18] Hence, the uncertainty arising from the choice of ligand conformation representing the equilibrium conformation in a water environment is 2.7 kcal·mol^{-1}, which is about 5% of the range of the total scores calculated. The implication of this uncertainty for our scoring is that if the score differs by less than about 2.7 kcal·mol^{-1}, the only conclusion is that such ligands are predicted to have similar binding activity. This limitation seems to be in most cases minor. It should be noted that HIV PR ligands are of a peptidomimetic character (a protein-like chain that mimics a peptide) and are thus extremely flexible, which is also evident in the probability density functions (Figure 4). The variation might be smaller when more rigid molecules are scored.

Even though the HIV PR inhibitors exhibit quite large conformational flexibility, none of the studied ligands shows a multiple-maxima probability density function of energy. This indicates that it is possible to choose a single conformation which would represent the ligand conformational energy of the equilibrium ensemble in an aqueous environment. Nevertheless, we admit that the choice of the proper conformation might not be straightforward and should be the object of further studies.

4. CONCLUSIONS

We estimated the uncertainty with which the conformational energy is calculated in our scoring function based on the quantum mechanical PM6-DH2 method. From extensive MD sampling at 300 K, we calculated the average values and standard deviations of the SMD solvation energy, gas-phase PM6-DH2 energy and their sum—the conformational energy—of nine HIV PR inhibitors. These energies were compared with the energies obtained by two other approaches.

The simplest approach (PM6-DH2 optimization) provides a relative uncertainty of about 3.5 kcal·mol^{-1} while the more demanding PM6-DH2 quenching improves this to 2.7 kcal·mol^{-1}. The extent of the flexibility of the inhibitors was demonstrated by the probability density functions of energy, based on 400 ns of simulations in total. From the simulations, the choice of the conformation which is used for the calculation of the atomic partial charges appears to be rather unimportant with respect to the conformational energy.

In the scoring studies based on rigid molecules (i.e., single-conformation approach), our results indicate an error brought by the approximation of neglect of flexibility. The results suggest that the resolution with which the scores can be interpreted is limited. The difference within the range of single kilocalories in the score might be misleading. However, this uncertainty is considerably smaller than the discrimination between binders and nonbinders, or between strong and weak binders. The uncertainty presented in this study then represents only a few percent of the score difference between those groups of ligands.

■ ASSOCIATED CONTENT

Supporting Information. The values of the E_{SMD}, E_{vac} and E_{conf} of all of the ligands, the plots of the charge variations for all of the ligands and the probability density functions of E_{SMD}, E_{vac} and E_{conf} for all of the ligands. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pavel.hobza@uochb.cas.cz.

■ ACKNOWLEDGMENT

This work was supported by the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic [Z40550506], the Ministry of Education, Youth and Sports of the Czech Republic [LC512 and MSM6198959216], the Czech Science Foundation [P208/11/0295] and Korea Science and Engineering Foundation [World Class Univ. program: R32-2008-000-10180-0]. This work was also supported by the Operational Program Research and Development for Innovations—European Science Fund (CZ.1.05/2.1.00/03.0058). The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged.

■ REFERENCES

(1) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(2) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(3) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.

(4) Chipot, C.; Pohorille, A. *Free Energy Calculations*, Springer-Verlag: Berlin, 2007.

(5) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(6) Huang, D.; Caflisch, A. *J. Mol. Recognit.* **2009**, *23*, 183–193.

(7) Zhou, T.; Caflish, A. *Chem. Med. Chem.* **2010**, *5*, 1007–1014.

(8) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 15–26.

(9) Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.

(10) DeWitte, R. S.; Shakhnovich, E. I. *J. Am. Chem. Soc.* **1996**, *118*, 11733–11744.

(11) Essex, J. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.

(12) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

(13) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.

(14) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–96.

(15) Friedman, R.; Caflisch, A. *Chem. Med. Chem.* **2009**, *4*, 1317–1326.

(16) Huang, D.; Caflisch, A. *J. Mol Recognit.* **2009**, *23* (2), 183–193.

(17) Hayik, S. A.; Dunbrack, R.; Merz, K. M. *J. Chem. Theory Comput.* **2010**, *6*, 3079–3091.

(18) Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Otyepka, M.; Hobza, P. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 223–235.

(19) Okamoto., Y.; Tanaka, T.; Kokubo, H. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 699–712.

(20) Söderhjelm, P.; Ryde, U. *J. Comput. Chem.* **2009**, *30*, 750–760.

(21) Shivakumar, D.; Deng, Y. Q.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

(22) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(23) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(24) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132–146.

(25) Jakalian, A.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(26) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. *J. Am. Chem. Soc.* **1995**, *117*, 1181–1182.

(27) Clemente, J. C.; Coman, R. M.; Thiaville, M. M.; Janka, L. K.; Jeung, J. A.; Nukoolkarn, S.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; Leelamanit, W.; Goodenow, M. M.; Dunn, B. M. *Biochemistry* **2006**, *45*, 5468–5477.

(28) Surleraux, D. L. N. G.; Tahri, A.; Verschueren, W. G.; Pille, G. M. E.; de Kock, H. A.; Jonckers, T. H. M.; Peeters, A.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. T. P. *J. Med. Chem.* **2005**, *48*, 1813–1822.

(29) AMunshi, S.; Chen, Z.; Li, Y.; Olsen, D. B.; Fraley, M. E.; Hungate, R. W.; Kuo, L. C. *Acta Crystallogr., Sect. D* **1998**, *54*, 1053–1060.

(30) Stoll, V.; Qin, W.; Stewart, K. D.; Jakob, C.; Park, C.; Walter, K.; Simmer, R. L.; Helfrich, R.; Bussiere, D.; Kao, J.; Kempf, D.; Sham, H. L.; Norbeck, D. W. *Bioorg. Med. Chem.* **2002**, *10*, 2803–2806.

(31) Kaldor, S. W.; Kalish, V. J.; Davies, J. F., II; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patick, A. K.; Reich, S. H.; Su, K. S.; Tatlock, J. H. *J. Med. Chem.* **1997**, *40*, 3979–3985.

(32) Kempf, D. J.; Marsh, K. C.; Denissen, J. F.; McDonald, E.; Vasavanonda, S.; Flentge, C. A.; Green, B. E.; Fino, L.; Park, C. H.; Kong, X. P. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2484–2488.

(33) Dohnálek, J.; Hašek, J.; Dušková, J.; Petroková, H.; Hradílek, M.; Souček, M.; Konvalinka, J.; Brynda, J.; Sedláček, J.; Fabry, M. *J. Med. Chem.* **2002**, *45*, 1432–1438.

(34) Liu, F. L.; Kovalevsky, A. Y.; Tie, Y.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. *J. Mol. Biol.* **2008**, *381*, 102–115.

(35) Petroková, H.; Dušková, J.; Dohnálek, J.; Skálová, T.; Vondráčková-Buchtelová, E.; Souček, M.; Konvalinka, J.; Brynda, J.; Fábry, M.; Sedláček, J.; Hašek, J. *Eur. J. Biochem.* **2004**, *271*, 4451–61.

(36) Urban, J.; Konvalinka, J.; Stehlíková, J.; Gregorová, E.; Majer, P.; Souček, M.; Andreánský, M.; Fábry, M.; Strop, P. *FEBS Lett.* **1992**, *298*, 9–13.

(37) Majer, P.; Urban, J.; Gregorová, E.; Konvalinka, J.; Novek, P.; Stehlíková, J.; Andreánský, M.; Sedláček, J.; Strop, P. *Arch. Biochem. Biophys.* **1993**, *304*, 1–8.

(38) Konvalinka, J.; Litera, J.; Weber, J.; Vondrášek, J.; Hradílek, M.; Souček, M.; Pichová, I.; Majer, P.; Strop, P.; Sedláček, J.; Heuser, A. M. *Eur. J. Biochem.* **1997**, *250*, 559–566.

(39) Kottler, H.; Kräusslich, H. G. *Eur. J. Biochem.* **1997**, *250*, 559–566.

(40) Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

(41) Hariharan, P. C.; Pople, J. A. *Theor. Chem. Acc.* **1973**, *28*, 213–22.

(42) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz,K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.

(43) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

(44) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(45) Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799–805.

(46) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(47) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(48) Parrinello, M.; Rahman, A. *J. App. Phys.* **1981**, *52*, 7182–7190.

(49) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(50) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(52) Stewart, J. J. P. *Stewart Computational Chemistry*, Colorado Springs, CO, MOPAC2009; http://OpenMOPAC.net.

Ligand Conformational and Solvation/Desolvation Free Energy in Protein-Ligand Complex Formation

Michal Kolář, Jindřich Fanfrlík and Pavel Hobza

Supporting Information

Table S 1: The average values and standard deviations of the energies calculated for the set of 160 conformations of each simulation. “Avg.” stands for the average over the set of all of the 1600 coformation of particular inhibitor. All of the values are in kcal.mol^{-1}.

E_{SMD}

	0	1	2	3	4	5	6	7	8	9	avg.
APV	-40.74	-40.57	-40.18	-40.13	-39.20	-40.11	-39.00	-39.18	-38.55	-40.38	-39.80
	2.17	2.52	2.99	2.24	2.47	1.80	2.48	3.34	2.47	2.59	2.64
ATV	-45.93	-46.50	-45.53	-45.77	-46.20	-45.38	-46.46	-46.14	-46.75	-46.42	-46.11
	2.80	2.97	3.15	3.48	2.90	3.19	2.63	2.49	2.86	2.68	2.96
DRV	-44.23	-45.42	-44.41	-44.52	-43.66	-42.62	-45.03	-43.93	-44.50	-44.93	-44.32
	1.87	2.75	2.85	2.84	3.16	2.73	2.56	2.71	2.40	2.24	2.74
IND	-39.25	-41.27	-43.52	-41.49	-41.35	-40.84	-41.41	-39.99	-40.46	-41.84	-41.14
	3.67	2.90	3.65	2.46	2.13	1.60	2.96	3.16	2.59	2.94	3.07
LPV	-39.82	-40.74	-40.92	-41.73	-41.17	-41.49	-40.94	-41.08	-40.48	-41.15	-40.95
	2.45	2.56	2.47	2.44	2.61	2.45	2.82	2.59	2.82	2.67	2.64
NFV	-29.43	-30.22	-29.00	-29.46	-29.44	-29.68	-30.95	-30.49	-29.55	-29.94	-29.82
	2.16	2.27	1.75	1.62	1.88	1.76	2.04	2.69	1.70	1.81	2.07
RTV	-45.70	-44.22	-45.05	-45.59	-44.90	-45.31	-44.44	-45.84	-45.35	-44.37	-45.08
	2.79	2.24	2.58	2.84	3.03	3.40	2.79	2.72	2.77	2.25	2.81
SAQ	-59.11	-59.71	-60.57	-58.55	-60.15	-58.02	-59.98	-60.14	-60.17	-59.21	-59.56
	4.57	5.00	4.68	4.63	4.06	5.29	4.30	4.30	4.65	4.00	4.63
SQV	-48.27	-45.62	-46.46	-41.39	-45.94	-46.62	-48.88	-46.63	-47.81	-46.84	-46.45
	3.27	3.82	3.86	2.43	4.07	3.83	4.57	2.18	3.06	2.51	3.95

E_{vac}

	0	1	2	3	4	5	6	7	8	9	avg.
APV	-245.24	-244.85	-246.78	-245.99	-247.65	-246.48	-250.18	-248.25	-251.19	-247.35	-247.40
	3.36	4.83	4.71	3.44	4.13	2.94	4.42	5.17	5.53	4.39	4.77
ATV	-280.75	-279.65	-281.07	-280.97	-280.55	-281.03	-281.06	-280.21	-280.68	-280.52	-280.65
	2.41	2.67	2.83	3.32	2.56	3.48	2.58	2.40	2.30	2.51	2.77
DRV	-288.32	-286.90	-288.35	-288.03	-289.69	-291.12	-287.50	-288.81	-287.65	-286.97	-288.33
	3.80	5.08	5.12	5.14	5.32	5.68	4.89	4.69	4.38	4.28	5.02
IND	-151.44	-149.10	-144.05	-147.56	-146.60	-158.22	-146.09	-147.70	-146.31	-146.37	-148.35
	3.52	4.27	4.42	3.56	2.61	3.98	5.86	4.65	4.47	3.45	5.63
LPV	-211.91	-211.92	-209.92	-209.43	-211.42	-211.30	-211.20	-209.98	-212.08	-212.20	-211.14
	3.73	3.98	4.78	4.13	3.97	4.24	4.32	4.53	3.91	4.31	4.31
NFV	-195.80	-194.81	-196.31	-195.59	-195.86	-195.50	-197.44	-194.36	-194.76	-195.11	-195.55
	3.01	3.21	2.51	2.68	2.49	2.78	3.78	3.64	3.02	2.61	3.12
RTV	-172.67	-173.69	-172.46	-171.95	-173.86	-172.14	-173.79	-172.70	-172.52	-172.78	-172.86
	3.80	3.47	3.24	4.15	4.60	4.51	3.63	2.98	4.02	4.20	3.95
SAQ	-309.81	-309.78	-309.31	-310.40	-309.86	-310.94	-308.47	-309.35	-308.43	-308.82	-309.52
	7.25	6.57	6.47	6.76	5.73	7.25	6.50	6.20	6.71	6.34	6.64
SQV	-210.41	-212.93	-211.94	-220.87	-212.11	-211.95	-213.78	-211.41	-210.82	-211.14	-212.73
	3.60	4.13	4.14	2.76	4.37	4.17	5.37	2.68	3.45	3.21	4.81

E_{conf}

	0	1	2	3	4	5	6	7	8	9	avg.
APV	-285.98	-285.42	-286.96	-286.13	-286.85	-286.59	-289.18	-287.43	-289.73	-287.73	-287.20
	2.59	3.09	2.74	2.58	2.64	2.59	3.24	3.19	4.22	2.87	3.29
ATV	-326.68	-326.15	-326.60	-326.75	-326.75	-326.41	-327.52	-326.35	-327.43	-326.94	-326.76
	2.35	2.49	2.67	2.59	2.40	2.50	2.24	2.65	2.27	2.40	2.50
DRV	-332.55	-332.33	-332.75	-332.54	-333.35	-333.75	-332.53	-332.74	-332.15	-331.91	-332.66
	2.86	3.61	3.37	3.40	3.50	3.77	3.55	3.58	3.18	3.08	3.44
IND	-190.69	-190.37	-187.57	-189.05	-187.95	-199.06	-187.49	-187.69	-186.77	-188.22	-189.49
	2.71	3.95	3.66	3.27	2.63	3.77	4.34	3.11	3.49	4.03	4.91
LPV	-251.74	-252.66	-250.84	-251.17	-252.59	-252.79	-252.14	-251.05	-252.56	-253.35	-252.09
	2.79	2.69	3.33	2.85	2.85	3.14	3.13	3.04	2.66	3.03	3.07
NFV	-225.22	-225.03	-225.31	-225.05	-225.29	-225.18	-228.39	-224.85	-224.32	-225.04	-225.37
	2.25	2.48	2.02	2.30	2.03	2.07	2.85	2.64	2.44	2.20	2.56
RTV	-218.37	-217.90	-217.50	-217.55	-218.76	-217.45	-218.23	-218.54	-217.87	-217.15	-217.93
	3.21	2.96	2.78	3.11	3.74	3.73	2.86	2.79	3.17	3.35	3.22
SAQ	-368.92	-369.49	-369.88	-368.95	-370.01	-368.96	-368.45	-369.49	-368.60	-368.03	-369.08
	4.26	3.95	3.70	3.95	3.50	3.80	4.14	3.79	3.86	4.16	3.96
SQV	-258.67	-258.55	-258.40	-262.26	-258.06	-258.57	-262.66	-258.04	-258.63	-257.98	-259.18
	1.97	1.88	2.02	1.93	2.39	1.86	2.43	1.67	2.02	1.70	2.60

Figure S 1: The variations of the atomic partial charges. For each atom, the range between the minimum and maximum charge within the ten charge sets is plotted.

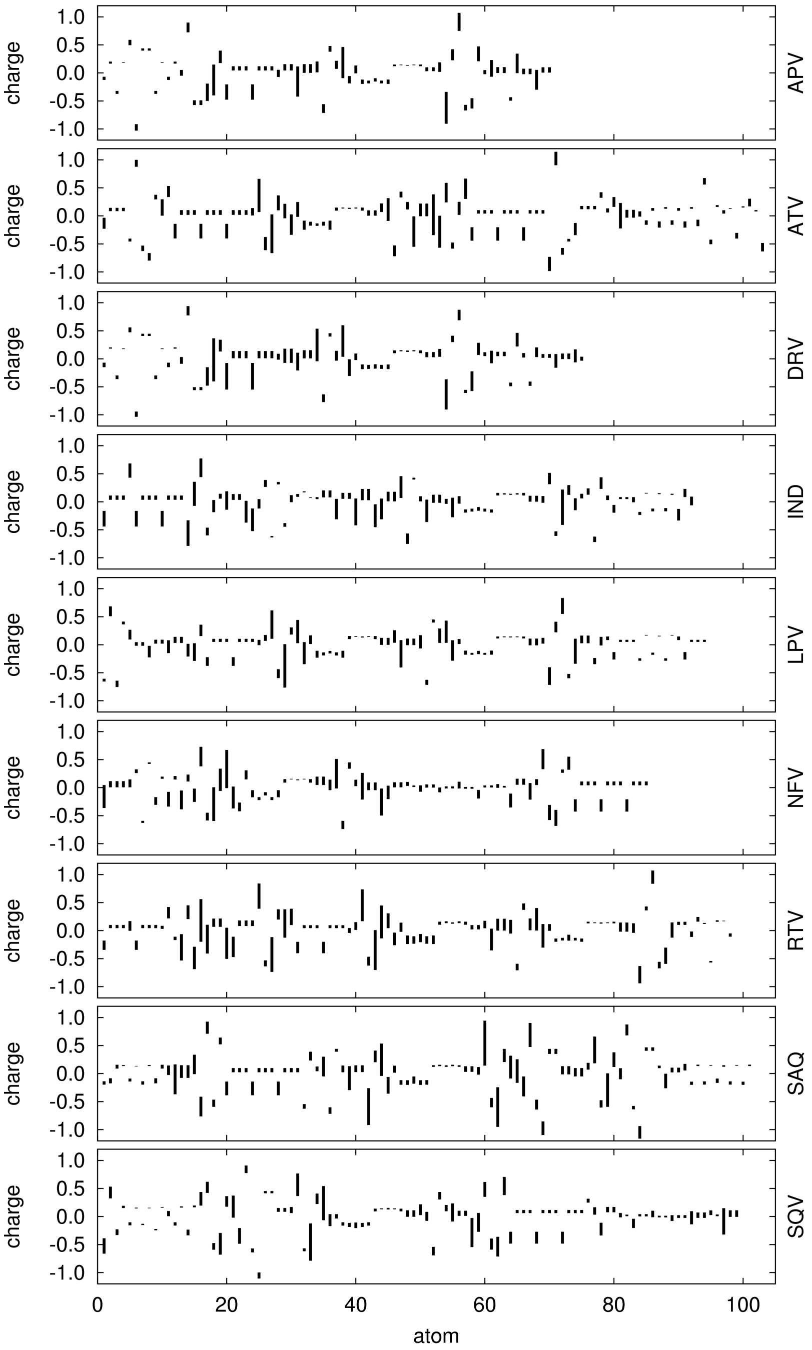


Figure S 2: The probability density functions (pdf) of the SMD solvation energy (E_{SMD}). The pdfs for the different charge sets calculated for ten ligand conformations are plotted. The pdfs are constructed from the set of 160 values of the energy. The 1600 values of all of the runs were used for the last row ("all"). The black vertical line indicates the mean value of the distribution; the orange lines indicate the values from the PM6-DH2 quench.

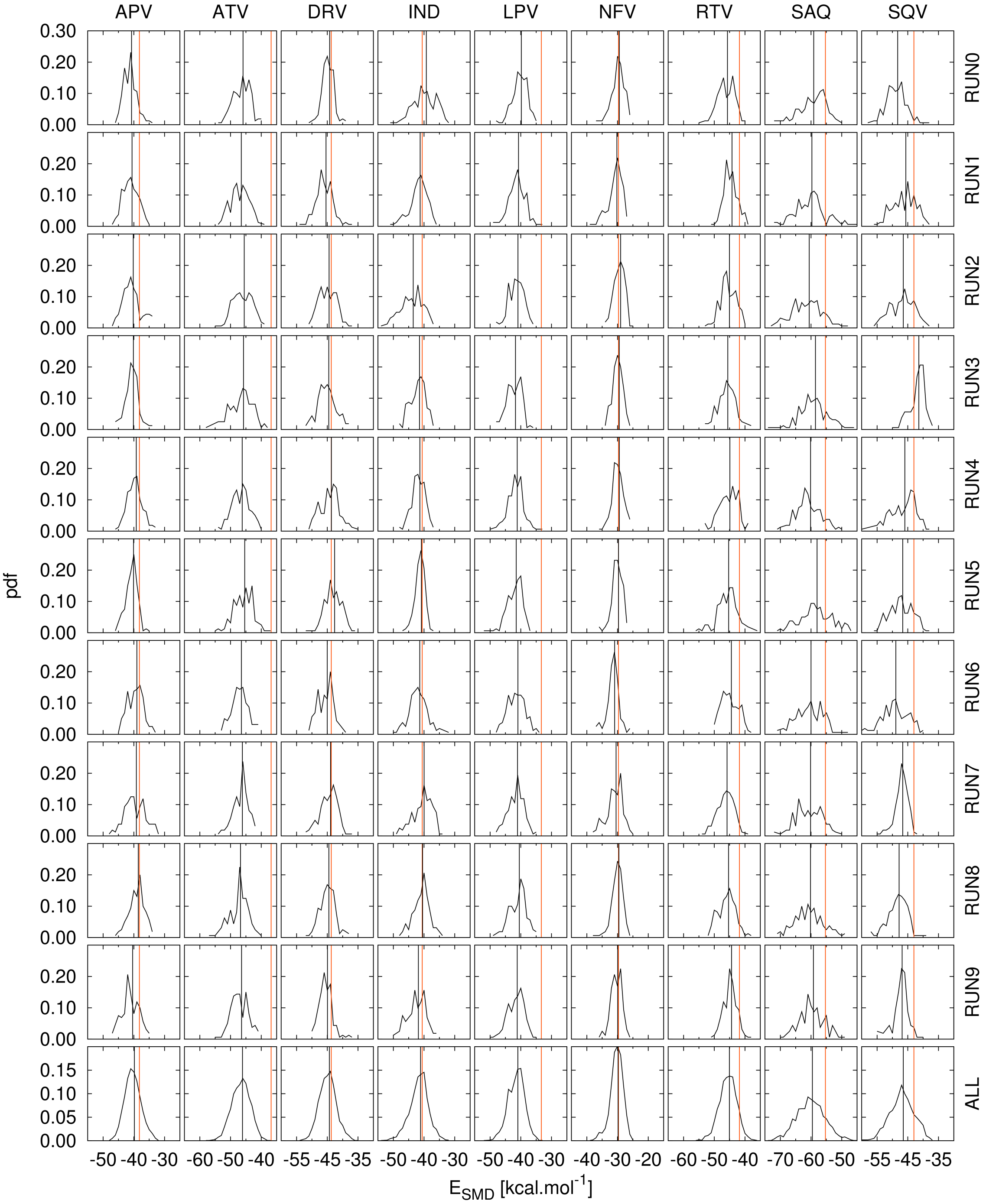


Figure S 3: The probability density functions (pdf) of the SMD solvation energy (E_{vac}). The pdfs for the different charge sets calculated for ten ligand conformations are plotted. The pdfs are constructed from the set of 160 values of the energy. The 1600 values of all of the runs were used for the last row ("all"). The black vertical line indicates the mean value of the distribution; the orange lines indicate the values from the PM6-DH2 quench.

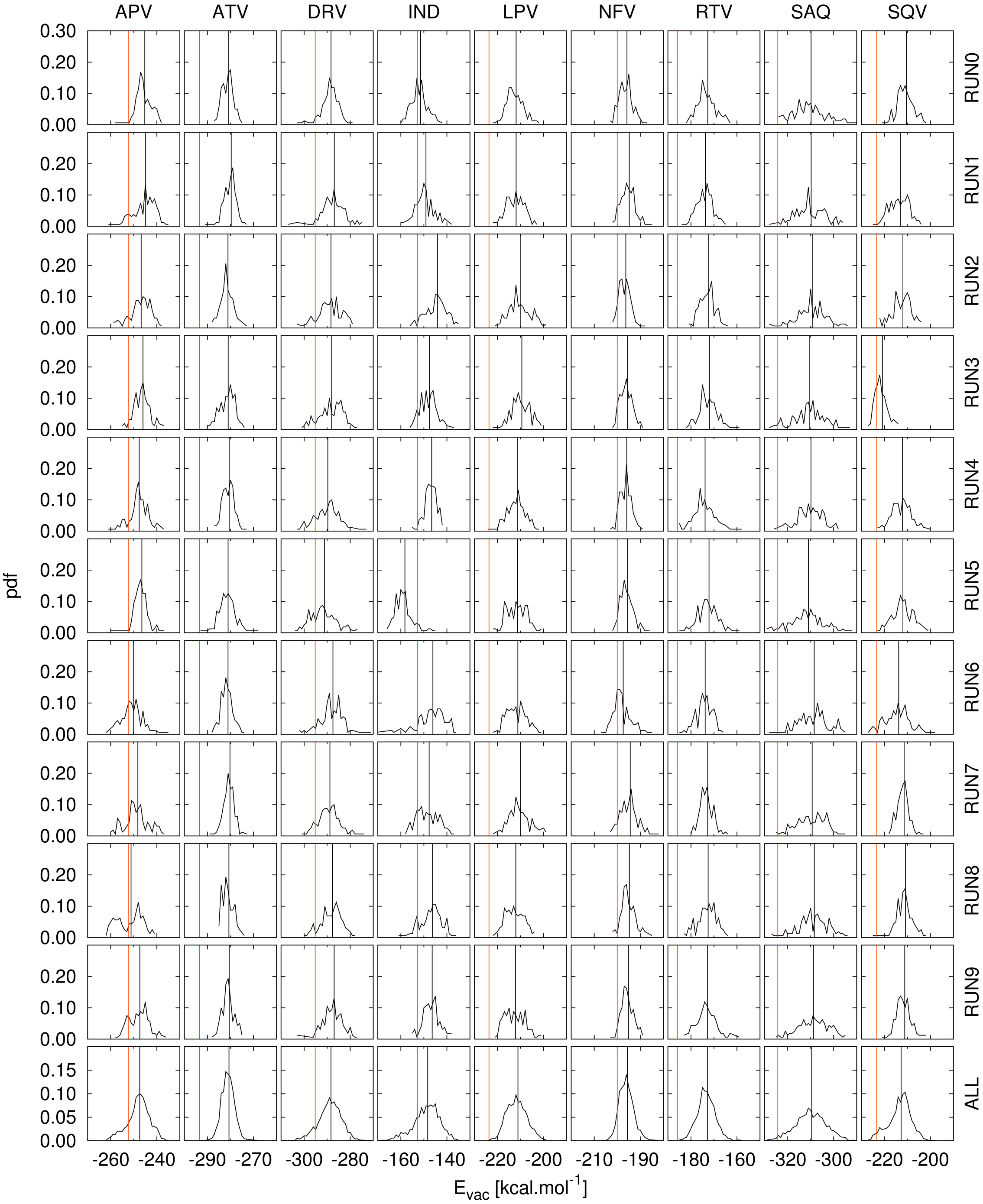
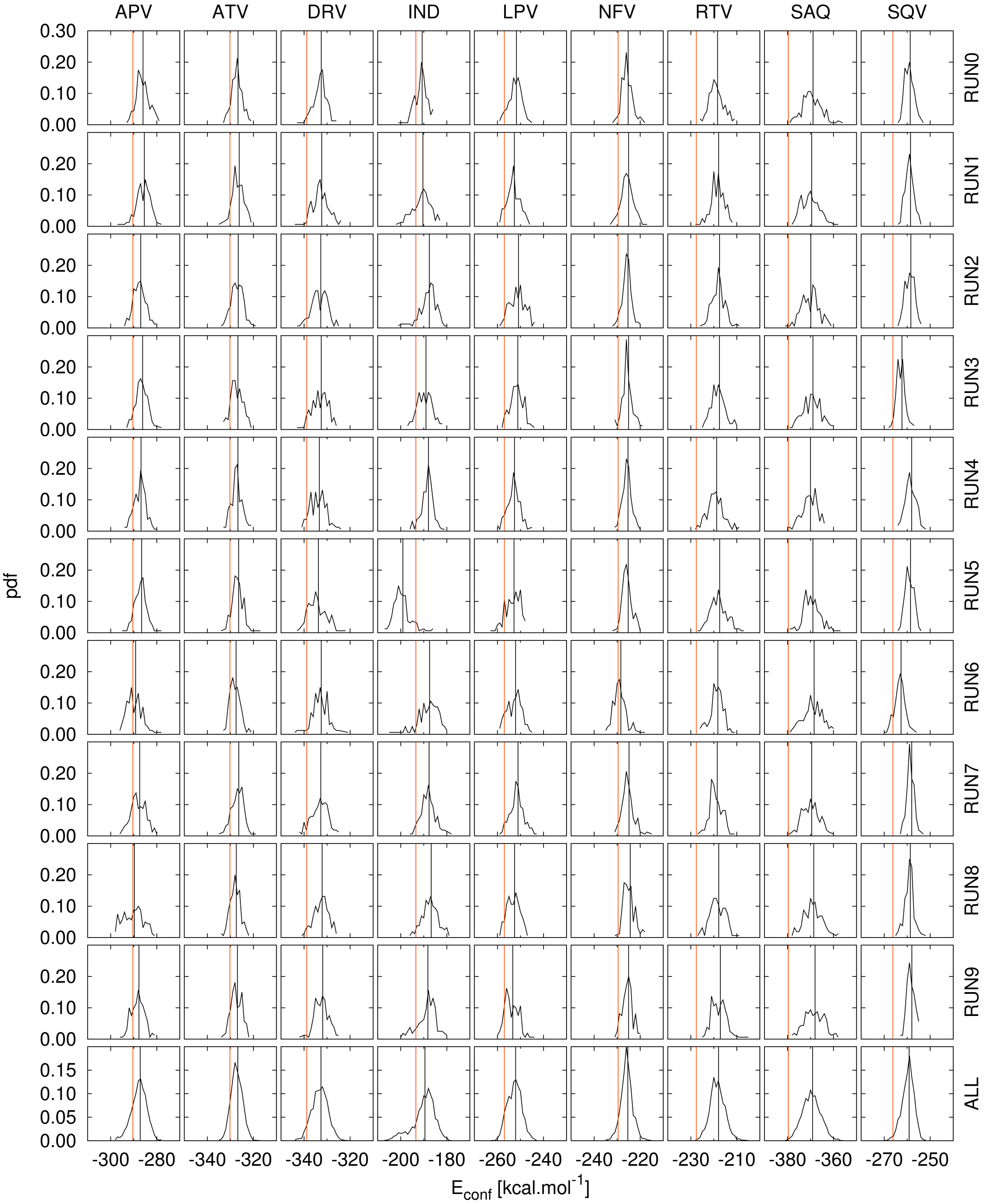


Figure S 4: The probability density functions (pdf) of the SMD solvation energy (E_{conf}). The pdfs for the different charge sets calculated for ten ligand conformations are plotted. The pdfs are constructed from the set of 160 values of the energy. The 1600 values of all of the runs were used for the last row ("all"). The black vertical line indicates the mean value of the distribution; the orange lines indicate the values from the PM6-DH2 quench.



D

Publication 3 – Water–Octanol Transfer Free Energies

Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches

Michal Kolář[a,b,*], Jindřich Fanfrlík[a], Martin Lepšík[a], Flavio Forti[c], F. Javier Luque[c] and Pavel Hobza[a,d,*]

[a] Institute of Organic Chemistry and Biochemistry and Gilead Sciences & IOCB Research Center, Academy of Sciences of the Czech Republic, v. v. i., 166 10 Prague 6, Czech Republic, mail: michal.kolar@uochb.cas.cz, pavel.hobza@uochb.cas.cz

[b] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague 2, Czech Republic

[c] Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, Campus de l'Alimentació, Santa Coloma de Gramenet, Spain

[d] Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

0 Abstract

The accuracy and performance of implicit solvent methods for solvation free energy calculations were assessed on a set of 20 neutral drug molecules. Molecular dynamics (MD) provided ensembles of conformations in water and water-saturated octanol. The solvation free energies were calculated by popular implicit solvent models based on quantum mechanical (QM) electronic densities (COSMO-RS, MST, SMD) as well as on molecular mechanical (MM) point-charge models (GB, PB). The performance of the implicit models was tested by a comparison with experimental water–octanol transfer free energies (ΔG_{ow}) by using single- and multi-conformation approaches. MD simulations revealed difficulties in *a priori* estimation of the flexibility features of the solutes from simple structural descriptors, such as the number of rotatable bonds. An increasing accuracy of the calculated ΔG_{ow} was observed in the following order: GB1 ~ PB < GB7 << MST < SMD ~ COSMO-RS with a clear distinction identified between MM- and QM-based models, although for the set excluding three largest molecules, the differences between COSMO-RS, MST and SMD were negligible. It was shown that the single-conformation approach applied to crystal geometries provides a rather accurate estimate of ΔG_{ow} for rigid molecules yet fails completely for the flexible ones. The multi-conformation approaches improved the performance, but only when the deformation contribution was ignored. It was revealed that for large-scale calculations on small molecules a recent GB model, GB7, provided a reasonable accuracy/speed ratio. In conclusion, the study contributes to the understanding of solvation free energy calculations for physical and medicinal chemistry applications.

1 Introduction

Implicit solvation models have found a distinguished place in computational chemistry. It was proven that for some types of chemical problems, the description of solvent as a dielectric polarizable continuum is reliable enough[1–4] while being orders of magnitude faster than the simulations with explicit solvent molecules. However, the implicit solvent models suffer from some deficiencies. For instance, when a specific interaction (e.g. a hydrogen bond) between a solute and solvent plays a role, the use of structureless continuum is problematic.[5] The limitations were also reported in description of hydrophobic/hydrophilic interfaces.[6]

Most of the implicit solvent models were parametrized to reproduce experimental solvation free energies, partition coefficients, or other macroscopic properties of simple organic compounds and/or ions.[7–9] In many schemes including those used in the present study, the solvation free energy is divided into contributions coming from electrostatic and non-electrostatic interactions. However, it is worth noting that the solvation free energy, which includes the average response of the solvent molecules, depends parametrically on the molecular geometry of the solute. Thus, as stated by Mennucci, for a given geometry of the solute "...continuum models automatically give configurationally sampled solvent effect".[10] Nevertheless, the assumption that the molecular structure (i.e. geometry) of a single conformation of the solute properly represents the statistical ensembles in both gas and solvent phases is questionable for flexible molecules.

The continuum solvent models have been popular in the calculations of protein–ligand or protein–protein affinities[11–13] applying the so-called single-conformation approach. The total binding free energy between the protein and the ligand consists of various energy terms,[13–15] among which the interaction energy between the protein and the ligand and the solvation free energy of the ligand are clearly dominant. The former term is negative (favoring the complex formation) while the latter one is positive, and there is a significant cancellation in the net contribution of these terms to the binding free energy. Today's computational chemistry determines the former term with a much higher accuracy than the latter one. Hence, an improvement of the calculation of solvation free energies is of high importance.

With the single-conformation approach, there appears another deficiency of the implicit solvent models: the physical base of the solvation free energy calculation is justified in the rigid molecule case since the conformation/geometry used for the solvation free energy calculation represents both the gas and solvated conformational ensembles well. In the case of flexible molecules, the situation may be different, because the single-conformation implicit solvation free energy cannot represent correctly both conformational ensembles (gas and aqueous one, for instance). In fact, we have shown that the conformation-dependent variance of solvation free energy can reach several kcal/mol for flexible peptidomimetic protein inhibitors when calculated for the ensemble of conformations, which is comparable with their total binding free energies.[16]

Water–octanol transfer free energies (ΔG_{ow}) are experimentally accessible quantities which are often used to characterize the hydrophobic/hydrophilic features of organic molecules. Such a property gives an account of the behavior of organic molecules, often drugs or drug candidates, in the vicinity of cell membranes and/or in the protein environment.[17, 18] Further, the ΔG_{ow} comprises in its definition the solvation/desolvation free energies used in computer-aided drug design as a part of the energetics of protein–ligand binding.[2, 11] Consequently, the determination of ΔG_{ow} is a key physicochemical parameter for understanding the pharmacodynamic and pharmacokinetic properties of drugs, which justifies its widespread impact in medicinal chemistry studies.

Experimentally, the ΔG_{ow} of a compound can be estimated from the equilibrium molar fractions. The uncharged molecule of interest (pH is adjusted to match this criterion) is solvated in the water-octanol mixture, and the equilibrium concentrations in water and in octanol, respectively, are measured after the liquid-phase separation, providing the equilibrium constant. The transfer free energy is then calculated by Equation 1:

$$\Delta G_{ow} = -RTln\frac{x_{oct}}{x_{wat}} , \qquad 1.$$

where x_{oct} and x_{wat} are the equilibrium molar fractions of the solute in water and octanol. The ratio of x_{oct} and x_{wat} is interpreted as the partition coefficient *P*, the decadic logarithm of which is used instead of ΔG_{ow}. There has been an abundance of methods proposed for *logP* estimation highlighting the relevance for the biological activity of the studied compounds.[19, 20] Apart from simple empirical approaches relying on the atom additivity (like AlogP[21] and XLogP[22]) or fragment additivity (e.g. ClogP)[23, 24] of certain properties, more elaborate methods were developed based on quantum chemically derived molecular properties,[25, 26] or molecular dynamics simulations.[27–29]

For the calculations of ΔG_{ow}, or equivalently *logP*, a simple thermodynamic cycle can be utilized (Figure 1A). This cycle implies the following equality (Equation 2):

$$\Delta G_{ow} = G_E(octanol) - G_E(water) = \Delta G_o - \Delta G_w , \qquad 2.$$

where ΔG_o stands for the solvation free energy of solute E in octanol (i.e. the transfer from the gas phase to octanol) and ΔG_w stands for the hydration free energy of solute E (i.e. the transfer from the gas phase to water). Once the geometry of the solute is known, these two terms can be conveniently calculated by an implicit solvation model.

The partition coefficients, or equivalently ΔG_{ow}, are not the subject of the study *per se*, but instead they serve as a tool for understanding the accuracy of the implicit solvation models. Experimentally, it is possible to measure solvation free energies for small molecules directly while it is much more complicated for large flexible molecules. The measurements of partition coefficients seem to suffer from this deficiency much less. Moreover, we claim that those implicit solvent models which are able to describe both water and low-dielectric media are suitable for a faithful description of biomolecule-ligand interactions.

To reflect this refined view on the molecular flexibility, the thermodynamic cycle of solvation in Figure 1A can be modified as shown in Figure 1B. Subsequently, the solvation free energies in water and in octanol are calculated for distinct conformations C1 and C2, which denote representative structures in water and octanol, respectively. The energy contribution due to the conformational change in the two phases is then accounted for by the term ΔG_{def}, which stands for the free energy of deformation in the gas phase. Clearly, the contribution of ΔG_{def} to the solvation free energy can be significant and should not be a priori neglected. A generalization of the scheme shown in Figure 1B can be made when multiple conformations are sampled by the solute in water and octanol.

The aim of the present study is to explore the practical aspects of solvation free energy calculations by continuum solvation models. The thermodynamic cycle in Figure 1B is probed in order to interpret single-conformation solvation free energies as well as those of conformational ensembles. The experimental transfer free energies are used as a reliable reference to the values calculated and the accuracy of the continuum solvent models is addressed.

To this end, we have examined three implicit quantum mechanical (QM) solvent models: the Miertus, Scrocco and Tomasi (MST) extension of the Integral Equation Formalism,[8, 30] the Solvent Model D (SMD)[9] and the Conductor-like Screening Model for Real Solvents (COSMO-RS).[7] While the two former methods rely exclusively on the single-conformation approach, the latter one provides a workflow of multi-conformational treatment. Indeed, COSMO-RS was previously used for the water-octanol *logP* and micelle-water *logP* estimations of organic molecules using conformational ensembles.[31, 32] Finally, we also calculated the transfer free energies using two flavors of the Generalized-Born (GB) model[33, 34] and Poisson-Boltzmann (PB) model[35, 36] both in conjunction with the molecular mechanical point-charge treatment of the molecules.

Being interested primarily in the accuracy and performance of implicit solvation models for computer-aided drug development, we employed a set of approved drugs gathered previously by Wang et al.[22] and extended the set by three HIV-1 protease inhibitors with known water-octanol partition coefficients. Running classical all-atom molecular dynamics (MD) simulations in explicit water and water-saturated octanol, we sampled the conformational spaces of the compounds, and continuum solvation models were subsequently used to calculate single- and multi-conformational water-octanol transfer free energies. The results were analyzed (a comparison with the experimental transfer free energy values, a comparison of single-conformation vs. ensemble-average approaches, etc.) to give an account of their physical significance and accuracy.

2 Methods

2.1 Studied Molecules

With an emphasis on computer-aided drug design, a set of 20 neutral organic compounds was compiled. It contains organic molecules of various sizes, flexibility and chemical diversity, thus representing a challenging set to be explored. The set of 17 approved drugs previously studied by Wang et al.[22] was extended by three HIV-1 protease inhibitors, which exhibit a large conformational flexibility and hence represent a more demanding test. Accordingly, the analysis was performed for the whole set of compounds and for the subset obtained upon the exclusion of the HIV-1 protease inhibitors. The water–octanol transfer free energies were calculated from the experimentally derived water-octanol *logP* values according to Equation 3:

$$\Delta G_{ow} = -RT\ln 10 \cdot \log P \ , \qquad 3.$$

where *R* is the gas constant and *T* is temperature. The *logP* values were taken from Hansh et al.[37] and Refs. 38–40. The starting geometries were taken either from the Cambridge Structural Database (pure molecule solid phase)[41–54] or from Protein Databank (in complex with a protein).[55–57] The crystal structures of dph, dzp and ptn compounds were not available. Hence, these molecules were built manually, and the geometries optimized at the M062X/6-31G* level of theory with the SMD served as the starting structures for further conformational sampling. All of the molecules studied are summarized in Table 1. The structural formulas are provided in the Supplementary Information (Figure S1).

2.2 Explicit Solvent Molecular Dynamics

The conformational space of the individual molecules was sampled by all-atom classical molecular dynamics (MD) in explicit solvent under periodic boundary conditions. The molecules were solvated in a cubic box of TIP3P water[58] and in a box with octanol saturated by water (water molar fraction 0.25).[59] For the solute molecules, General Amber Force Field (up-to-date version 2011) (GAFF) was used[60] in conjunction with an automatic atom-type assignment of the Antechamber program.[61] For all-atom octanol simulations, we prepared our own combination of GAFF and OPLS-AA to reproduce the liquid properties of pure-octanol and water–octanol mixtures keeping the solvent-solute interactions consistent. Indeed, our effort has overcome the problem of different combination rules of atomic Lennard-Jones parameters in GAFF and OPLS-AA force fields.[60, 62, 63] When compared to the experimental values, the signed relative errors in pure-octanol density and vaporization enthalpy are only -2% and +3%, respectively, and the error in water–octanol mixture density is -3%, which all seem to be sufficiently accurate for the conformational sampling of the organic compounds taking into account the previous simulations of liquid octanol.[64, 65] The octanol parameters are available in the Supplementary Information. In the discussion below, the phrase "octanol" MD simulations is used, if not mentioned otherwise, as a short name for the simulations in water-saturated octanol.

The solvated drug molecules were subjected to the standard equilibration, which was finalized by a high-temperature simulation (330 K). The purpose of the high-temperature simulation was to generate a variety of starting conformations for the production runs. In the case of water, ten starting conformations were collected every 100 ps, while in the case of octanol it was every 300 ps. These conformations were then used for 3 ns (water) and 6 ns (octanol) long simulations at a temperature of 300 K and a pressure of 1 bar. From the last two thirds of each simulation, ten snapshots were generated resulting in 100 snapshots per molecule, for which the solvation free energies were calculated by the implicit solvation models (see below). The snapshots are provided in the xyz file format as the Supplementary Information. The simulation scheme, i.e. running several uncorrelated MD simulations, was previously termed as Multiple Molecular Dynamics, and it was shown to be an efficient methodology for conformational sampling.[66, 67] For each compound, a total of 20 ns of trajectories in water and 40 ns of trajectories in water-saturated octanol were used for the preparation of the snapshots.

The simulation details were as follows: a time step of 2 fs, all bonds constrained by the LINCS algorithm,[68] the electrostatics were treated by the Particle Mesh Ewald algorithm with a direct-space cut-off of 1.2 nm in octanol and 1.0 nm in water simulations; Lennard-Jones interactions were included within a cut-off of 1.2 nm in octanol and 1.0 nm in water simulations. The temperature was maintained by the Nosé-Hoover thermostat[69, 70] and the pressure by the Parrinello-Rahman barostat.[71] The compressibility of water and octanol was $4.5 \cdot 10^{-5}$ bar^{-1} and $7.43 \cdot 10^{-5}$ bar^{-1},[72] respectively. The MD simulations were conducted in the Gromacs program package.[73]

2.3 Implicit Solvent Models

For the set of 100 snapshots per molecule, the implicit solvation free energies were calculated in the manner of a standard single-conformation approach. The computational schemes adopted for QM solvation methods were chosen as those recommended by the developers of each model. Below, the particular implicit solvation free energy calculations are summarized. Next, the calculations performed using two widely used MM-based implicit solvent methods, generalized Born (GB) and Poisson-Boltzmann (PB), are also described.

2.3.1 The COSMO-RS and COSMO Implicit Models

COSMO-RS[7, 74] involves a statistical-mechanics post-processing of QM implicit solvent model COSMO[75] calculations of solutes to obtain their free energies of solvation (in water or octanol in our case). We employed AM1 optimizations in COSMO using high-solvent screening (a dielectric constant of 78.4) in MOPAC 2009[76] and a DFT single point at the BP86/SVP level using a perfect insulator in COSMO (a dielectric constant of infinity) in Turbomole 6.3.[77] Further, the vacuum optimizations (and single-point calculations) are needed as well. The resulting G_w' and G_o' (the notation highlights the difference of origin of the numbers when compared to the other estimators) were subtracted, yielding transfer free energies.

A built-in treatment of the ensembles in COSMO-RS is done via the auto-conformer option in which the representative conformers are submitted for post-processing together. We test here three approaches for the selection of conformers, all the 100 snapshots, and 50 snapshots from the first and second halves, the latter two for addressing the issue of convergence.

For the sake of extended comparison, we calculated hydration free energies with the COSMO method alone. These were later compared with other hydration free energies to gain insight into its performance, since the original COSMO model is not well-suitable for low-dielectric media such as octanol. The hydration free energies were calculated in MOPAC 2009[76] after the full PM6-D3H4[78–80] gradient optimization in the implicit water environment.

2.3.2 The MST Implicit Model

The snapshots were optimized at the B3LYP/6-31G* level in vacuo with harmonic restraints on all heavy atoms following the procedure described by Butler et al.[81] The aim of this strategy was to avoid the occurrence of drastic conformational changes in the sampled molecular structure during geometry optimization. To this end, the optimization was performed by using constraints related to the Debye-Waller temperature factors, in conjunction with the keyword *opt=loose* in Gaussian calculations,[82] which sets less strict criteria for convergence and expedites job completion. Test computations showed that this approach has a negligible impact on the relative energies.[81] Single-point calculations were subsequently done at the B3LYP/6-31G* level in the gas phase and in the solvent (octanol, water) according to the MST model.[8, 30]

2.3.3 The SMD Implicit Model

The snapshots were optimized at the M062X/6-31G* level[83] in the SMD implicit model[9] of the particular solvent (water or octanol) using the standard convergence criteria (the change of energy of 0.006 kcal/mol and the maximum gradient of 1.2 kcal/mol/Å) in Gaussian09.[82]

As the performance of the SMD model is limited by the use of a rather demanding M06X functional here, we also calculated transfer free energies with a faster electronic energy method, namely at the PM6-D3H4 level.[78–80] These results are presented separately, since they were not initially recommended for the use with the SMD.

2.3.4 The GB and PB Models

The solvation free energies were calculated in the Amber program package[84] using i) the Hawkins et al. GB model (abbrev. GB1),[34, 85] ii) the GB model described by Mongan et al. (abbrev. GB7)[86] and iii) a finite-difference numerical PB solver.[35, 36] The corresponding Sander options were i) igb=1 ii) igb=7 iii) igb=10. Igb=7 and igb=10 were used with modified Bondi radii as was recommended by Kongsted et al. [87] The dielectric constant of water and octanol was set to 78.5 and 9.8629 (identical with the SMD octanol), respectively. No cut-off for pairwise interactions was adopted. Other molecular mechanical parameters such as atomic partial charges, Lennard-Jones parameters, bonding, angle and torsional parameters were taken from GAFF and were thus kept identical with those used for the conformational MD sampling. The surface tension for the non-electrostatic free energy contributions was 0.005 kcal/mol/Å^2.

For the solvation free energies, we used unoptimized geometries from the trajectories as used in MM-GBSA studies.[88, 89] Further, also optimized (i.e. energy-minimized) structures were used to remain consistent with the rest of the methods in this study. The optimization was performed in the particular implicit solvent model using the Sander program[84] with the default optimization setup (10 steps of the steepest descent followed by 90 steps of the conjugate gradient algorithm) and convergence criteria (gradient RMSD < $1 \cdot 10^{-4}$ kcal/mol/Å).

2.4 Analysis

2.4.1 The Radius of Gyration and Rotatable Bonds

We calculated the radius of gyration R_{gyr} for each snapshot, which was used to follow the conformational changes in flexible molecules. The results were analyzed by considering the number of *relevant rotatable bonds*. This parameter was defined as the number of bonds between sp^3- or sp^2- hybridized heavy atoms to which at least one additional heavy atom was bound. This means that -OH, $-NH_2$ and $-CH_3$ groups were not considered as *relevant rotatable bonds*. The peptide bond (-CO-NH-) was not included either.

2.4.2 Cluster Analysis – Rigid and Flexible Drugs

A cluster analysis was performed for the series of 100 conformations taken for each molecule (by using the Gromacs g_cluster tool).[73] The conformation was included into the cluster if the heavy-atom RMSD with respect to any member of the cluster was lower than 1 Å (the "linkage" method in the g_cluster). The cluster can be viewed also as a conformational family.
According to the number of clusters in water, the set was divided into two subsets: i) rigid, with only one cluster, and ii) flexible, with more than one cluster. The rigid subset included the following molecules: cpr, cth, dlt, dzp, imp, ldc, pam, pbl, ptn and tpm. The flexible subset was formed by the following molecules: atr, cam, dph, ffa, hpd, idv, nfv, ppl, sqv and tcn. Some molecules from the rigid subset do

have rotatable bonds; thus they are not strictly rigid and exhibit some extent of conformational freedom. For the sake of simplicity, however, we keep the subset notation as "rigid" and "flexible". We note that the subsets would remain identical if the numbers of clusters in octanol were taken as the criterion.

2.4.3 Transfer Free Energy Estimators

The solvation free energies in water ΔG_w and in octanol ΔG_o were calculated for the initial (crystal) conformation as well as for conformational ensembles. The respective transfer free energies were estimated several times, following Equations 4, 5, 6, 7 and 8, and were abbreviated by simpler notation. First, the single-conformation solvation free energies in water and in octanol were calculated for the series of initial geometries (see Section 2.1) according to Equation 4.

$$G_0 = \Delta G_{ow}(0) = \Delta G_o(\text{Xray}) - \Delta G_w(\text{Xray}) \qquad 4.$$

$$G_1 = \Delta G_{ow}(1) = \langle \Delta G_o \rangle_o - \langle \Delta G_w \rangle_w \qquad 5.$$

$$G_2 = \Delta G_{ow}(2) = \langle \Delta G_o \rangle_o - \langle \Delta G_w \rangle_w + \langle \Delta E_d \rangle \qquad 6.$$

$$G_3 = \Delta G_{ow}(3) = \langle \Delta G_o - \Delta G_w \rangle_w \qquad 7.$$

$$G_4 = \Delta G_{ow}(4) = \langle \Delta G_o - \Delta G_w \rangle_o \qquad 8.$$

ΔG_o and ΔG_w are the single-conformation solvation free energies in octanol and water, respectively, calculated for each snapshot and averaged over the set from octanol MD $< >_o$ and water MD $< >_w$. The combination of the average values in octanol and water yields the estimate G_1 (Equation 5). The additional contribution in G_2, the deformation free energy in Figure 1B, was approximated by the average deformation electronic energy calculated by Equation 9:

$$\langle \Delta E_d \rangle = \langle \Delta E_o \rangle_o - \langle \Delta E_w \rangle_w \qquad 9.$$

where E_o and E_w stand for the internal electronic energy of the particular snapshot.

The G_3 values stand for the average transfer free energy assuming that the water and octanol conformational ensembles are identical and represented by the water ones. In other words, for G_3, the solvation free energy was calculated in octanol (ΔG_o) and water (ΔG_w), but only for those 100 snapshots obtained from water MD ($< >_w$). A similar calculation was done for octanol MD, providing G_4. We note that the G_3 and G_4 values tend to be those which are commonly used in the computer-aided drug design, i.e. ignoring the different conformational ensembles in water and in vacuo.

2.4.4 Statistical Evaluation

The correlation coefficients (R^2) between the experimental and theoretical values were calculated for the whole set of compounds as well as for the two subsets. The mean signed absolute error (*MSAE*) and root-mean-square error (*RMSE*) were calculated for the entire set as well as for the subsets with respect to the experimental data according to Equations 10 and 11, where N stands for the number of set/subset members.

$$MSAE = \frac{1}{N} \sum \left[G_i(\text{calc}) - G_i(\exp) \right] \qquad 10.$$

$$RMSE = \left(\frac{1}{N} \sum \left[G_i(\text{calc}) - G_i(\exp) \right]^2 \right)^{0.5} \qquad 11.$$

In order to elucidate the convergence of the simulations (i.e. the representability of the 100 snapshots), we calculated R^2, *MSAE* and *RMSE* and compared the values arising from the first and second halves of the MD simulations.

3 Results and Discussion

This section is organized as follows: First we report the analysis of the explicit-solvent classical MD simulations of 20 drug molecules in water and water-saturated octanol with respect to their flexibility. Second, the performance of the implicit solvation models is discussed based on the estimated values of the water–octanol transfer free energy for the entire set of solutes, followed by the analysis of the rigid/flexible subsets. The outliers were interesting in their own right and are presented separately. To eliminate a possible effect of the cancellation of errors between water and octanol solvation free energies, a comparison of the hydration free energies only is also provided. Finally, the representability of the limited set of 100 snapshots, the issue of convergence and the CPU cost of the individual methods are examined.

3.1 The Flexibility Features of the 20 Drugs

The radii of gyration of the conformations from individual MD snapshots are shown in Figure 2. The correlation between the standard deviations (std) of R_{gyr} in water and octanol is high (a correlation coefficient of 0.93), which shows a similar extent of the molecule-size fluctuations for the molecules in the two solvents. Not surprisingly, there are some compounds with distinct behavior in water and in octanol. The haloperidol (hpd) drug is an example of a molecule which prefers a compact conformation in water, most likely because of the hydrophobic effect, and an extended conformation in octanol (Figure 3). As mentioned below, the existence of different populations of conformers in the two solvents poses a challenge for the prediction the solvation free energies by implicit solvent models.

Highly flexible molecules are expected to have a high number of clusters. The number of clusters in water and octanol is shown in Table 1, where the size (i.e. number of members) of the largest cluster is provided in parentheses. The number of the clusters is slightly lower in octanol than in water, and both are well correlated (R^2 of 0.96). We admit that some clusters contain only one snapshot. Nevertheless, the snapshots represent reasonably 1 % of the entire simulation time, i.e. 200 ps in the case of the water simulation or 400 ps in the case of the octanol simulation, and thus are accounted for in further considerations.

Based on the number of clusters, the two most flexible compounds from the set are indinavir (idv) and saquinavir (sqv). Both are peptidomimetic HIV-1 protease inhibitors and possess the highest number of atoms and rotatable bonds. Surprisingly, nelfinavir (nfv) has a much lower number of clusters, even though this compound has a similar number of atoms and rotatable bonds to idv and sqv. This can hint at an undersampling of the largest molecules in this set. The convergence of the results is therefore discussed below.

We show here that the actual flexibility defined by the number of clusters is not directly related to the number of rotatable bonds. An example may be procainamide (pam) and haloperidol (hpd), both with six rotatable bonds. The former drug exhibits lower conformational flexibility (one cluster in both octanol and water) than the latter one (15 clusters in water, 8 in octanol). Moreover, hpd shows two distinct conformations in water and a number of intermediates between them (Figure 3). This finding may complicate the debate about the possibility to estimate conformational freedom *a priori* without any sampling, because it shows that the flexibility features of a molecule strikingly depend on the *type* of rotatable bonds and not on their *number*. Consequently, the flexibility depends on the overall molecular topology.

3.2 The Overall Performance of Implicit Solvation Methods

The statistical indicators, R^2, *MSAE* and *RMSE*, for all the implicit solvation methods are summarized in Figure 4. Various transfer free energy estimators (G_0–G_4, Section 2.4.3.) are shown. Generally, the methods based on electronic density (COSMO-RS, MST, SMD) perform better than the methods based on molecular mechanical partial charges (GB1, GB7, PB) as shown by all the indicators.

Unexpectedly, the performance of the models is not dramatically changed passing from the single-conformational approach to the ensemble of conformations. In all the cases, the G_0 estimators provide only slightly worse results than the best multi-conformational estimator (G_1–G_4).

The COSMO-RS and SMD provide acceptable agreement with the experimental data (R^2 of about 0.75) while another electronic density model – MST – gives worse agreement (R^2 of about 0.55). GB and PB models perform poorly (R^2 < 0.40). COSMO-RS and MST models tend to underestimate the water–octanol transfer free energies having a negative *MSAE*, which reflects an overestimation of the hydration free energy, or alternatively an underestimation of the solvation free energy in octanol.

Surprisingly, there is no big difference between free energy estimators. Thus, there is a negligible difference between the G_3 and G_4 values across all the implicit solvent models. These estimators do not cover deformations, because both ΔG_w and ΔG_o are calculated for the same geometries. The correlation between G_3 and G_4 is higher than 0.98 and the *RMSE* is lower than 0.3 kcal/mol. The amount of water in water-saturated octanol (a molar fraction of 0.25) may be sufficient for the solvation of small molecules to mimic locally pure water properties. Presumably, the conformational ensembles in water and octanol, which do not differ considerably (see above), yield almost the same averages. Moreover, in all instances, the contribution of deformation worsens the performance of the estimator. For all of the models, the correlation coefficient is the lowest (i.e. the worst) for the G_2 estimator, and for the MST model also the *MSAE* and *RMSE* worsen as well when the deformation is included (Figure 4).

A comparison of the hydration free energies is shown in Figure 5. For our purposes here, the SMD hydration free energies were arbitrarily chosen as reference data. The $<G_w>_w$ values are plotted in Figure 5 and besides the methods discussed so far, the original COSMO method is shown as well. We note that in the case of COSMO-RS, the G_w values arising from the multi-conformation *a posteriori* weighting do not strictly correspond to the other methods.

All of the models provide values in the similar intervals, although the GB models have a slightly wider range. On average, the MST model has hydration free energies less negative by 5.9 kcal/mol when compared to the SMD model. On the other hand, COSMO yields hydration free energies more negative (about 2.5 kcal/mol) than the SMD model. This is in contrast with COSMO-RS, which gives more positive hydration free energies (by 2.0 kcal/mol). The difference must lie in the post-processing scheme, which in the case of bare COSMO is substituted by the simple ensemble average. The GB models are on average closer than 2 kcal/mol to the SMD model. The correlation coefficients are higher than 0.8 in all the instances, the highest value being found for the COSMO-RS, PB and COSMO models. The correlation coefficients of the $<G_o>_o$ are lower (about 0.7).

Overall, it can be stated that the differences in the performance of the models in the prediction of the water–octanol transfer free energies arise mostly from the various extent of error cancellation between aqueous and octanol solvation free energies and also from the varying accuracy of the octanol-phase description.

Figure 6 shows the typical probability density functions (pdf) of the transfer free energies for atropine (atr). The gaussian pdfs with the mean values and the standard deviations are depicted for G_1-G_4 estimators. The histograms of the solvation free energies in water and octanol are also shown (Figure 6, left).

Two important aspects of our calculations are demonstrated. First, the pdf of solvation free energies do not need to be normally distributed, which agrees with our previous findings.[16] Indeed, the octanol pdf (Figure 6, right) shows a bimodal-like shape and thus the gaussian pdf is only an approximation to this. Second, there are striking differences in the pdf half-width across the estimators (Figure 6, left). This seems to be the reason why G_2 results with a large standard deviation show lower correlation while improving the *RMSE* than the other estimators with smaller standard deviations.

What needs to be mentioned as well is that R^2 is related to the relative order of the compounds, unlike the *MSAE* and *RMSE*, which show the ability of the model to describe rather absolute values of the solvation free energies. Hence, the deformation (as calculated in G_2) on average disrupts the relative order of the solvation free energies but shifts them closer to the experimental values. The disruption may originate from the fact that the internal energies of the compounds are not calculated equally well/badly or that the cancellation of errors is not systematic. This is

supported by several studies that have reported significant errors (around 12-18%

when Pople-type basis set are used) in predicting the conformational energies from

density functional calculations. [90–92]

In Table S1 in the Supplementary Information, the performance of the models based on molecular mechanical point charges is shown for both optimized and unoptimized geometries. There are only minor changes for the estimators without a deformation contribution, but a large deviation is found for G_2, which covers the gas-phase deformation (c.f. Figure 1). The use of unoptimized geometries in the MM-PB/GBSA[88, 89] studies is justified thanks to a slightly better agreement with the experimental data. Nevertheless, as mentioned above, the performance is still poor.

Table 2 summarizes the distributions of error sizes for the implicit solvent model-ΔG_{ow} estimator combinations. According to the distributions, the SMD with G_3 estimator might be evaluated as the most successful approach, which has a high number of low-deviating molecules and where only a few instances have an error > 3 kcal/mol. MST suffers from a rather large number of molecules with a high absolute error. The performance of the GB and PB models is typically poor, except for the G_2 estimator with a favorable distribution of error magnitudes.

3.3 The Performance of Implicit Solvation Methods for Rigid vs. Flexible Drugs

The complete set of correlation plots for the entire set of molecules and for rigid and flexible subsets is provided as Supplementary Information (Figures S2–S6). There is a very good agreement between the calculations and experiment for the rigid subset of molecules. A typical correlation coefficient of the electronic density based models exceeds 0.80 and also for the GB models is larger than 0.60. One source of errors (i.e. conformational freedom) in implicit solvation is eliminated here by the inherent nature of the compounds; hence the deviations have to come from different sources. Most likely, the limit of the parametrization process has been reached and the solvation free energies suffer from errors arising from the insufficient quality of such parameters as atomic radii (i.e. the solute/solvent boundary) or atomic surface tensions. Slightly worse results were obtained when the deformation contribution was included (Figure S4), and this effect was more pronounced in molecular mechanical point-charge based methods.

Such agreement was, however, observed for the single-conformational approach as well (Figure S2). For the rigid subset, the overall best performance was achieved by the COSMO-RS method (R^2 = 0.87, *RMSE* = 1.84 kcal/mol). For the rigid subset, the worst performance of GB7 was still acceptable in the sense of the relative order of the molecules in the set (R^2 = 0.67, *RMSE* = 5.19 kcal/mol).

The flexible compounds were, as expected, described with a lower accuracy than the rigid ones. For the single-conformational G_0 estimator, there was no agreement at all, which points to the real need for adopting more advanced methods. For the multi-conformational estimators, except for a few cases, the correlation was poor (R^2 lower than 0.4). Again, the G_1 estimator provided systematically better values than G_2. Overall, the best description of flexible molecules was reached by a SMD G_1 model–estimator combination (R^2 = 0.66, *RMSE* = 2.02), and also COSMO-RS yielded acceptable agreement (R^2 = 0.48, *RMSE* = 2.16 kcal/mol).

3.4 Outliers

In the correlation plots, a few outliers could be identified. The compounds with the highest absolute error are shown in Table 2. Within the molecules, two reasons can be recognized for the failure of implicit solvent methods: i) the molecules are either large/flexible (hpd, idv, nfv, sqv) or ii) they contain problematic chemical fragments (cth, ffa).

The transfer free energies of chlorothiazide (cth) exhibit large deviations from the experiment no matter what implicit solvent model and estimator are used. For COSMO-RS and SMD, cth was identified as the molecule with the highest absolute deviation from the experimental value (Table 2). The conformational flexibility of the compound is low with only one cluster in water and in water-saturated octanol, too. The R_{gyr} standard deviations are the third lowest. However, the cth compound contains a sulfonyl group (R-SO_2-R'), which was previously shown to be problematic. In the SAMPL blind challenge,[93] the solvation energies of compounds containing the sulfonyl group disagreed with the experiment markedly,[94–96] and even the experimental values themselves were questioned.[95]

3.5 The Exclusion of Three the Largest Drugs

The subset of compounds obtained upon the exclusion of the three HIV-1 protease inhibitors (idv, nfv, sqv) which exhibit the largest size and conformational flexibility was also analyzed. The results of the analysis for this subset are shown in Figure 7. Clearly, the differences between the methods diminished when the three largest molecules were excluded from the analysis. In other words, for the small molecules in our set (i.e. of fewer than 50 atoms), the performance of the *ab initio* electronic density methods (COSMO-RS, MST, SMD) was practically identical, with the correlation coefficient being about 0.8 and the *RMSE* about 2 kcal/mol. The methods based on the molecular mechanical point charges improved as well (R^2 being about 0.6 and the *RMSE* being lower than 6 kcal/mol: cf. Figures 4 and 7).

For the GB1, GB7 and MST models, the most problematic compounds were nelfinavir (nfv) and saquinavir (sqv), and it is thus not surprising that the performance increased upon their exclusion. In MST, a certain role may be played by the specific optimization setup scheme as it relies on the restrained optimization,[81] contrary to the unrestrained optimizations of COSMO-RS and SMD, although we cannot rule out that the deviation found for these compounds reflects a bias in the solvation contribution of a specific functional group (i.e. amide).

3.6 SMD with Electronic Energy from Semiempirical Quantum Mechanics

As shown below, the SMD recommended scheme employing the M062X density functional is one of the most CPU-demanding options used here, mostly due to the time-consuming energy minimization of the snapshots. Therefore, we explored an alternative strategy where the transfer free energies were determined employing a less demanding parametrized semiempirical PM6-D3H4 method. Table 3 presents the differences in the SMD transfer free energies using M062X and PM6.

The estimators which cover multiple conformations seem to be good enough also with the PM6-D3H4 optimization scheme. The *MSAEs* are positive and approximately twice as large as those of M062X. In accordance with the previous section, the results of SMD employing PM6-D3H4 improve when HIV-1 protease inhibitors are excluded. The R^2 of G_3 and G_4 increases to 0.70 and 0.69, respectively, and the *RMSE* decreases to 4.12 kcal/mol and 4.07 kcal/mol, respectively.

3.7 The Convergence of the Estimators

The convergence was examined by a separate evaluation of the first and second halves of the MD simulations. For each half, the set of 50 snapshots underwent the same analysis as the entire set of 100 snapshots. R^2, *MSAE* as well as *RMSE* were almost identical for both halves. The values are provided in the Supplementary Information (Table S2). The largest differences were found for the G_2 estimator while the lowest deviations were found for the G_3 and G_4 estimators.

Typically, the root-mean-square deviation between the transfer free energies obtained from the first and second halves was about 0.35 kcal/mol, lower for G_3 and G_4 estimators, higher for G_2. The number of the molecules having the absolute deviation between the transfer free energies obtained from the first and second halves lower than 0.1 kcal/mol was about 2 for the G_2 estimator and 15 otherwise. The compounds idv, sqv and imp were identified among the least converged. The details for all method-estimator combinations are provided in the Supplementary Information (Table S3).

The root-mean-square deviation between the halves for COSMO-RS was about 0.5 kcal/mol, with the maximal absolute deviation being about 0.98 kcal/mol for imp compound.

3.8 Speed-Accuracy Ratio

In Table 4, we provide the total time needed for the calculation of water–octanol transfer free energies for the initial geometries (i.e. the G_0 estimator). Due to the inherent distinctions within the workflows, the numbers should be considered as the first approximation for the practical effort one has to exert to obtain the transfer free energies of 20 drug molecules. The time was determined for a single-core job run on the processor from the Xeon family.

To put the results into context, there are two aspects which affect the computational demands: the electronic energy part and the solvation energy part. The SMD models seem to be rather demanding, no matter if the semiempirical (thus fast) PM6 or density-functional M062X (slow) optimization is performed. Moreover, when the timing of COSMO-RS and SMD-PM6 is compared, even though they both employ a semiempirical method for internal electronic energy calculation, the former one (used with AM1) is 50 times faster than the latter one.

4 Summary

A series of 20 drug molecules was examined by explicit solvent all-atom classical MD simulations. The output of the water and water-saturated octanol MD simulations – twice 100 snapshots per molecule – was the subject of implicit solvent calculations for COSMO-RS, MST, SMD, PB and two flavors of GB models in order to obtain water-octanol transfer-free energies. Several ensemble averages were proposed as a multi-conformational treatment and compared to the single-conformation results as well as to the experimental values. A direct comparison of the implicit solvation models is extremely difficult, bearing in mind the different parametrization strategies as well as slightly different purposes of the models.

The solvation models underwent parametrization processes of varied complexity. For instance, the SMD model was parametrized against more than 2,800 solvation data including 274 hydration free energies, 206 octanol free energies and 90 water–octanol logP values.[9] Similarly, the COSMO-RS training set contained about 160 hydration free energies and 175 water-octanol logP values.[74] Finally, the MST model was parametrized using a much smaller training set (about 72 data for water and 63 for octanol)[8, 30] The GB models were either parametrized against a series of experimental values (219 neutral compounds)[85] or to represent explicit solvent MD.[34, 35] An attempt was made to calculate effective Born radii cheaply and at the same time accurately with an emphasis on the protein MD simulations. The PB model was subject of various integration schemes while the Born radii were inherited from the previous studies and also targeted to large biomolecular MD. Besides these, the implicit solvent models also differ in the number of parameters.

In the context of the calculations, the overall performance of the models is that the native multi-conformational approach of the COSMO-RS model provides the best results for our set of neutral molecules with a reasonable computer cost. For the entire test set, we observed an increasing accuracy of the calculated transfer free energies in the following order: GB1 ~ PB < GB7 << MST < SMD ~ COSMO-RS, where there was a jump increment identified between MM- and QM-based models. When molecules having fewer than 50 atoms were considered, the accuracy of MST, SMD and COSMO-RS was almost equal. This also corresponds to the results of the previous tests of the methods, e.g. in the SAMPL blind challenge.[93] The dominance of these methods over the GB and PB models is apparent.

For large-scale calculations (e.g. large molecules or a large set of small molecules), it may be, however, advantageous to use methods based on molecular mechanical partial charges. Then, GB7 with the G_1 estimator provided reasonable accuracy for the rigid subset as well as the set excluding large HIV-1 inhibitors while being about 25 times faster than the fastest QM-based method.

Our study indicates that for large, highly flexible molecules, the single-conformation approach failed completely. However, among the multi-conformation approaches tested here, only two models, COSMO-RS and SMD, provided reasonable improvement over the single-conformation approach. The G_1 and G_3 estimators in conjunction with the SMD model provided the most accurate results considering R^2, in case of former estimator, or *RMSE* in case of latter one.

It is of interest that the multi-conformation estimator covering the deformation of conformations, G_2, performed poorly. The reason might be the fortuitous cancellation of errors for the identical conformations in G_3 and G_4 . This cancellation of errors seems to be very sensitive to the level at which the internal energies (thus the deformation energies) of the solute are calculated. In this context, it is hence justified to ignore the deformation contribution.

On the other hand, the traditional single-conformation approach can, to our surprise, provide good enough results as compared to some of the multi-conformation approaches treated here. This holds true especially for rigid molecules, where no attempts to include CPU demanding conformational sampling provided any significant improvement. We thus conclude that in such applications where the solvation free energies and logP values need to be obtained efficiently, a single-conformational approach applied to reliable geometries (e.g. X-ray) provides a rather accurate estimate.

5 Acknowledgement

The authors are grateful to Dr. Milan Kožíšek for valuable discussions on partition coefficients. This work was part of the Research Project RVO: 61388963 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to PH in 2007 is also acknowledged. This work was also supported by the Czech Science Foundation [P208/12/G016] and the operational program Research and Development for Innovations of the European Social Fund (CZ 1.05/2.1.00/03/0058). FJL acknowledges the financial support from the Spanish Ministerio de Economia y Competitividad (SAF2011-27642), Generalitat de Catalunya (SGR2009-298) and XRQTC, and the computational facilities provided by the Centre de Supercomputació de Catalunya (CESCA).

6 Supplementary Information

Table S1 shows the R^2, *MSAE* and *RMSE* of the GB1, GB7 and PB models for optimized and unoptimized geometries. Table S2 shows the R^2, *MSAE* and *RMSE* for the transfer free energies calculated from the first and second halves of the MD simulations. Table S3 summarizes the variations between the transfer free energies calculated from the first and second halves of the MD simulations. Figure S1 depicts the structural formulas of the compounds studied, whereas Figures S2–S6 show the correlation plots for the entire set and the rigid and flexible subsets for the G_0–G_4 estimators. The molecular topology in the Gromacs format is provided for the water-saturated octanol periodic box. 200 snapshots per molecule (water and octanol simulations) are provided in the xyz file format. This material is available free of charge via the Internet at http://pubs.acs.org.

7 Bibliography

1. Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161–2200.
2. Orozco, M.; Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **2000,** *100,* 4187–4226.
3. Feig, M.; Brooks, C. L. Recent Advances in the Developments and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr. Opin. Struc. Biol.* **2004,** *14,* 217–224.
4. Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.
5. Zhou, R.; Berne, B. J. Can a Continuum Solvent Model Reproduce the Free Energy Landscape of a β-hairpin Folding in Water? *Proc. Natl. Sci. Acad.* **2002**, *99*, 12777–12782.
6. Lin, J. H.; Baker, N. A.; McCammon, J. A. Bridging Implicit and Explicit Solvent Approaches for Membrane Electrostatics. *Biophys. J.* **2002**, *83*, 1374–1379.
7. Klamt, A. Conductor-like Solvent Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.
8. Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22,* 1180–1193.
9. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.
10. Mennucci, B. Continuum Solvation Models: What Else Can We Learn from Them? J. *Phys. Chem. Lett.* **2010**, *1*, 1666–1674.
11. Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand Solvation in Molecular Docking. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 4–16.
12. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004,** *47,* 3032–3047.
13. Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Přenosil, O., Konvalinka, J., Hobza, P. A Reliable Docking/Scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.

14. Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Otyepka, M.; Hobza, P. Transferable Scoring Function Based on Semiempirical Quantum Mechanical PM6-DH2 Method: CDK2 with 15 Structurally Diverse Inhibitors. *J. Comput. Aided Mol. Des.* **2011**, *25*, 223–235.
15. Brahmkshatriya, P. S.; Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Paruch, K.; Bronowska, A. K.; Lepšík, M.; Hobza, P. Quantum Mechanical Scoring: Structural and Energetic Insights into Cyclin-dependent Kinase 2 Inhibition by Pyrazolo[1,5-a]pyrimidines. *Curr. Comput. Aided Drug Des.* **2013**, *9*, 118.
16. Kolář, M.; Fanfrlík, J.; Hobza, P. Ligand Conformational and Solvation/Desolvation Free Energy in Protein-Ligand Complex Formation, *J. Phys. Chem. B* **2011**, *115*, 4718–4724.
17. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.
18. Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* **2009,** *98,* 861–893.
19. Leo, A. Calculating logP_{oct} from Structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
20. Buchwald, P.; Bodor, N. Octanol-Water Partition. *Curr. Med. Chem.* **1998**, *5*, 353–380.
21. Viswanadhan, V.N.; Ghose, A.K.; Revankar, G.R.; Robins, R.K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
22. Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculation Partition Coefficients. J. Chem. Inf. Sci. **1997**, *37*, 615–621.
23. Chou, J. T.; Jurs, P. C. Computer-Assisted Computation of Partition Coefficient from Molecular Structure Using Fragment Constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172-178.
24. Suzuki, T.; Kudo, Y. Automated logP Estimation Based on Combined Additive Modeling Methods. *J. Comput.-Aided. Mol. Des.* **1990**, *4*, 155–198.
25. Klopman, G.; Irrof, L. D. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem.* **1981**, *2*, 157–160.
26. Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996,** *96,* 1027–1044.
27. Best, S. A.; Merz, K. M.; Reynolds, C. H. Free Energy Perturbation of Octanol/Water Partition Coefficients: Comparison with Continuum GB/SA Calculations. *J. Phys. Chem. B* **1999**, *103*, 714–726.
28. Lyubartsev, A. P.; Jacobsson, S. P.; Sundholm, G.; Laaksonen, A. Solubility of Organic Compounds in Water/Octanol Systems. A Expanded Ensemble Molecular Dynamics Simulation Study of log P Parameters. *J. Phys. Chem. B* **2001**, *105*, 7775–7782.
29. Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. 1-Octanol/Water Partition Coefficients of n-Alkanes Molecular Dynamics Simulations of Absolute Solvation Free Energies. *J. Chem. Theory Comput.* **2009**, *5*, 2436–2446.
30. Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST model to the IEF formalism: HF and B3LYP parametrizations. *J. Mol. Struc.: THEOCHEM* **2005**, *727*, 29–40.

31. Buggert, M.; Cadena, C.; Mokrushina, L.; Smirnova, I.; Maginn, E. J.; Arlt, W. COSMO-RS Calculations of Partition Coefficients: Different Tools for Conformation Search. *Chem. Eng. Technol.* **2009**, *32*, 977–986.
32. Mokrushina, L.; Yamin, P.; Sponsel, E.; Arlt, W. Prediction of Phase Equilibria in Systems Containing Large Flexible Molecules Using COSMO-RS: State-of-the-problem. *Fluid Phase Equil.* **2012**, *335*, 37–42.
33. Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.
34. Tsui, V.; Case, D.A. Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations. Biopolymers. *Nucl. Acid. Sci.* **2001**, *56*, 275–291.
35. Luo, R.; David, L.; Gilson, M.K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244–1253.
36. Lu, Q.; Luo, R. A Poisson-Boltzmann Dynamics Method with Nonperiodic Boundary Condition. *J. Chem. Phys.* **2003**, *119*, 11035–11047.
37. Hansch, C.; Leo, A.; Hoekman, D. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants, Vol. 2.; American Chemistry Society: Washington, DC, **1995**;
38. Drewe, J.; Gutmann, H.; Fricker, G.; Török, M.; Beglinger, C.; Huwyler, J. HIV Protease Inhibitor Ritonavir: a More Potent Inhibitor of P-glycoprotein than the Cyclosporine Analog SDZ PSC 833. *Biochem. Pharm.* **1999**, *57*, 1147–1152.
39. The Scientific discussion for the approval of Crixivan by European Medicines Agency.
40. Pfizer Canada Inc. Product Monograph - Viracept; Kirkland, Quebec, Canada, **2011**.
41. Tanczos, A. C.; Palmer, R. A.; Potter, B. S.; Saldanha, J. W.; Howlin, B. J. Antagonist Binding in the Rat Muscarinic Receptor - A study by Docking and X-ray Crystallography. *Comput. Biol. Chem.* **2004**, *28*, 375–385.
42. Sundaraligam, M.; Lin, H. Y.; Arora, S. K. Chloramphenicol. ACS Abstr. Papers (Summer) **1971**, 71.
43. McDowell. J. H. H. Crystal and Molecular Structure of Chloropromazine. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1969**, *25*, 2175.
44. Kojic-Prodic. B.; Ruzic-Toros, Z.; Sunjic, V.; Decorte, E.; Moimas, F. Absolute Conformation and Configuration of (2s, 3s)-3-acetoxy-5-(dimethylaminoethyl)-2-(4-methoxyphenyl)-2,3-dihydro-1,5-benzothiazepin-4(5h)-one Chloride (Dilthiazem Hydrochloride). *Helv. Chim. Acta* **1984**, *67*, 916
45. Johnston, A.; Florence, A. J.; Kennedy, A. R. Chlorothiazide-Pyridine (1/3). *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2008**, 64, 1105–1106.
46. McConnell. J. F. *Cryst. Struct. Commun.* **1973**, *2*, 459.
47. Datta, N.; Mondal, P.; Pauling, P. Structure of Haloperidol Hydrobromide [4-[4-(4-chlorophenyl)-4-hydroxypiperidino]-4'-fluorobutyrophenone HBr] *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *35*, 1486–1488.
48. Paulus, E. F. Crystal and Molecular-Structure of 5-(3-dimethylaminopropyl)-10,11-dihydro-5h-dibenzo[b,f]azepine Hydrobromide (Imipramine Hydrobromide) - Comparison with Structure of Imipramine Hydrochloride. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1978**, *34*, 1942–1947.
49. Hanson, A. W.; Banner, D. W. 2-diethylamino-2',6'-acetoxylidide (Lidocaine). *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1974**, *30*, 2486-2488.

50. Peeters, O. M.; Blaton, N. M.; De Ranter, C. J.; Denisoff, O.; Molle, L. 4-amino-n-[2-(diethylamino)ethyl]benzamide Monohydrochloride (Procainamide Hydrochloride), C13H22ClN3O *Cryst. Struct. Commun.* **1980**, *9*, 851–856.

51. Platteau, C.; Lefebvre, J.; Hemon, S.; Baehtz, S.; Danede, F.; Prevost, D. Structure determination of forms I and II of phenobarbital from X-ray powder diffraction. *Acta Crystallogr., Sect. B: Struct. Sci.* **2005**, *61*, 80–88.

52. Bredikhin, A. A.; Savel'ev, D. V.; Bredikhina, Z. A.; Gubaidullinl, A. T.;Litvinov, I. V. Crystallization of Chiral Compounds 2. Propranolol: Free Base and Hydrochloride. *Russ. Chem. Bull.* **2003**, 52, 853–861.

53. Hamaed, H.; Pawlowski, J. M.; Cooper, B. F. T.; Fu, R.; Eichhorn, S. H.; Schurko, R. W. Application of Solid-State ^{35}Cl NMR to the Structural Characterization of Hydrochloride Pharmaceuticals and Their Polymorphs. *J. Am. Chem. Soc.* **2008**, *130*, 11056–11065.

54. Umadevi, B.; Prabakaran, P.; Muthiah, P. T. A Pseudo-Quadruple Hydrogen-Bonding Motif Consisting of Six N-H...O Hydrogen Bonds in Trimethoprim Formate. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **2002**, *58*, 510–512.

55. Liu, F.; Kovalevsky, A. Y.; Tie, Y.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Effect of Flap Mutations on Structure of HIV-1 Protease and Inhibition by Saquinavir and Darunavir *J. Mol. Biol.* **2008**, *381*, 102–115.

56. Munshi, S.; Chen, Z.; Li, Y.; Olsen, D. B.; Fraley, M. E.; Hungate, R. W.; Kuo, L. C. Rapid X-ray Diffraction Analysis of HIV-1 Protease-Inhibitor Complexes: Inhibitor Exchange in Single Crystals of the Bound Enzyme *Acta Crystallogr. D, Biol. Crystallogr.* **1998**, *54*, 1053–1060.

57. Kaldor, S. W.; Kalish, V. J.; Davies II, J. F.; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patick, A. K.; Reich, S. H.; Su, K. S.; Tatlock, J. H. Viracept (Nelfinavir Mesylate, AG1343): A Potent, Orally Bioavailable Inhibitor of HIV-1 Protease. *J. Med. Chem.* **1997**, *40*, 3979–3985.

58. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.

59. Sassi, P.; Paolantoni, M.; Cataliotti, R. S.; Palombo, F.; Morresi, A. Water/Alcohol Mixtures: A Spectroscopic Study of the Water-Saturated 1-Octanol Solution. *J. Phys. Chem. B* **2004**, *108*, 19557−19565.

60. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

61. Wang, J.; Wang, W.; Kolmann, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.

62. Halgren, T. A. The Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827−7843.

63. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

64. DeBolt, S. E.; Kollman, P. A. Investigation of Structure, Dynamics, and Solvation in 1-Octanol and Its Water-Saturated Solution: Molecular Dynamics and Free-Energy Perturbation Studies. *J. Am. Chem. Soc.* **1995**, *117*, 5316−5340.

65. MacCallum, J. L.; Tieleman, D. P. Structures of Neat and Hydrated 1-Octanol from Computer Simulations. *J. Am. Chem. Soc.* **2002**, *124*, 15085−15093.

66. Auffinger, P.; Westhof, E. H-bond Stability in the tRNA$_{\text{Asp}}$ Anticodon Hairpin: 3 ns of Multiple Molecular Dynamics Simulations. *Biophys. J.* **1996**, *71*, 940–954.

67. Caves, L.S.D.; Evanseck, J.D.; Karplus, M. Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin, *Prot. Sci.* **1998**, *7*, 649–666.

68. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

69. Nosé, S. A Molecular Dynamics Method for Simulations in Canonical Ensemble. *Mol. Phys.* **1984**, *52*, 255−268.

70. Hoover, W. G. Canonical Dynamics - Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.

71. Parrinello, M.; Rahman, A. Polymorphic Transitions in Single-Crystals: A New Molecular-Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

72. Kiselev, V. D.; Bolotov, A. V.; Satonin, A.; Shakirova, I.; Kashaeva, H. A.; Konovalov, A. I. Compressibility of Liquids. Rule of Noncrossing V−P Curvatures. *J. Phys. Chem. B* **2008**, *112*, 6674−6682.

73. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

74. Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998,** *102,* 5074−5085.

75. Klamt, A.; Schuurmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, *5*, 799−805.

76. Stewart, J. P. P. Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (**2008**).

77. Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162*, 165–169.

78. Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

79. Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

80. Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *85* ,141−151.

81. Butler, K. T.; Luque, F. J.; Barril, X. Toward Accurate Relative Energy Predictions of the Bioactive Conformations of Drugs. *J. Comput. Chem.* **2009**, *30*, 601−610.

82. Gaussian 09, Revision A.1, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark,

M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT, **2009**.

83. Zhao, Y.; Truhlar, D.G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

84. Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling,C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 11; University of California: San Francisco, **2010**.

85. Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

86. Mongan, J.; Simmerling, C.; A. McCammon, J.; A. Case, D.; Onufriev, A. Generalized Born with a Simple, Robust Molecular Volume Correction. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

87. Kongsted, J.; Soderhjelm, P.; Ryde, U. How Accurate are Continuum Solvation Models for Drug-like Molecules? *J. Comput. Aided Mol. Des.* **2009**, *23*, 395−409.

88. Srinivasan, J.; Cheatham, 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.

89. Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A. & Cheatham, III, T.E Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc Chem. Res.* **2000**, *33*, 889−897.

90. Grimme, S.; Mück-Lichtenfeld, C. Calculation of Conformational Energies and Optical Rotation of the Most Simple Chiral Alkane. *Chirality* **2008**, *20*, 1009−1015.

91. Riley, K. E.; Op't Holt, B. T.; Merz, K. M. Critical Assessment of the Performance of Density Functional Methods for Several Atomic and Molecular Properties. J. *Chem. Theory Comput.* **2007**, *3*, 407−433.

92. Forti, F.; Cavasotto, C. N.; Orozco, M.; Barril, X.; Luque, F. J. A Multilevel Strategy for the Exploration of the Conformational Flexibility of Small Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 1808−1819.

93. Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501−4507.

94. Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds. *J. Phys. Chem. B* **2009**, *113*, 4508−4510.
95. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 4538−4543.
96. Soteras, I.; Forti, F.; Orozco, M.; Luque, F. J. Performance of the IEF-MST Solvation Continuum Model in a Blind Test Prediction of Hydration Free Energies. *J. Phys. Chem. B* **2009**, *113*, 9330−9334.

Table 1: The list of the molecules studied. For each molecule, the number of atoms and rotatable bonds as well as the number and size of the conformational clusters found in molecular dynamics simulations in water and octanol are given.

Commercial name	Abbrev.	No. atoms	No. rotatable bonds	Clusters in water (maxSize)	Clusters in octanol (maxSize)
Atropine	atr	44	5	2 (96)	4 (91)
Chloramphenicol	cam	32	6	4 (67)	2 (89)
Chlorpromazine	cpr	40	4	1	1
Chlorothiazide	cth	23	1	1	1
Diltiazem	dlt	55	7	1	1
Diphenhydramine	dph	40	6	2 (98)	2 (98)
Diazepam	dzp	33	1	1	1
Flufenamic acid	ffa	30	3	2 (92)	1
Haloperidol	hpd	49	6	15 (59)	8 (64)
Indinavir	idv	92	12	59 (15)	22 (40)
Imipramine	imp	45	4	1	1
Lidocaine	ldc	39	5	1	1
Nelfinavir	nfv	85	10	7 (87)	6 (90)
Procainamide	pam	38	6	1	1
Phenobarbital	pbl	29	2	1	1
Propranolol	ppl	40	6	3 (97)	2 (98)
Phenytoin	ptn	31	2	1	1
Saquinavir	sqv	99	13	35 (33)	17 (57)
Tetracaine	tcn	43	8	2 (97)	2
Trimethoprim	tpm	39	5	1	1

Table 2: The absolute error distributions of the studied compound. The molecule with the highest deviation is provided in the *worst* line. The values of the GB and PB models are presented for unoptimized geometries. *For COSMO-RS, the values correspond to the universal multi-conformational estimator.

Model	Bins	G_0	G_1*	G_2	G_3	G_4
C-RS	< 1 kcal/mol	3	5	-	-	-
	1-2 kcal/mol	12	12	-	-	-
	2-3 kcal/mol	2	1	-	-	-
	> 3 kcal/mol	3	2	-	-	-
	worst	cth	idv	-	-	-
MST	< 1 kcal/mol	6	6	4	6	6
	1-2 kcal/mol	3	5	7	6	6
	2-3 kcal/mol	2	2	2	1	1
	> 3 kcal/mol	9	7	7	7	7
	worst	sqv	sqv	sqv	sqv	sqv
SMD	< 1 kcal/mol	10	6	9	12	12
	1-2 kcal/mol	6	8	5	5	4
	2-3 kcal/mol	1	3	5	1	2
	> 3 kcal/mol	3	3	1	2	2
	worst	cth	cth	cth	cth	cth
GB1	< 1 kcal/mol	0	0	8	0	0
	1-2 kcal/mol	0	0	5	0	0
	2-3 kcal/mol	2	2	4	2	2
	> 3 kcal/mol	18	18	3	18	18
	worst	ffa	sqv	sqv	ffa	ffa
GB7	< 1 kcal/mol	0	0	7	0	0
	1-2 kcal/mol	0	0	7	0	0
	2-3 kcal/mol	1	1	4	1	1
	> 3 kcal/mol	19	19	2	19	19
	worst	ffa	sqv	sqv	ffa	ffa
PB	< 1 kcal/mol	0	0	8	0	0
	1-2 kcal/mol	0	0	6	0	0
	2-3 kcal/mol	0	0	4	0	0
	> 3 kcal/mol	20	20	2	20	20
	worst	ffa	sqv	hpd	nfv	idv

Table 3: A comparison of the SMD solvation free energies obtained at two electronic energy levels of theory – density functional M062X and parametrized semiempirical PM6-D3H4. The correlation coefficients R^2, the mean signed absolute errors (*MSAE* in kcal/mol) and the root-mean-square errors (*RMSE* in kcal/mol) are shown.

	R^2		MSAE		RMSE	
	M062X	PM6	M062X	PM6	M062X	PM6
G_0	0.65	0.48	1.25	4.44	2.29	6.26
G_1	0.77	0.60	1.70	4.85	2.29	5.99
G_2	0.62	0.68	1.19	3.37	2.04	4.20
G_3	0.75	0.65	1.14	3.59	1.78	4.42
G_4	0.72	0.55	1.10	3.25	1.79	4.27

Table 4: The time needed for the calculation of water–octanol transfer free energies for the initial (crystal) geometries (i.e. the G_0 estimator).

	time / min
COSMO-RS	26
MST	1038
SMD (M062X)	75056
SMD (PM6)	7906
GB1	< 1
GB7	< 1
PB	8

Figure 1 (one column): The simple (A) and refined (B) thermodynamic cycles describing the gas–water–octanol phase equilibria. See the text for the details.

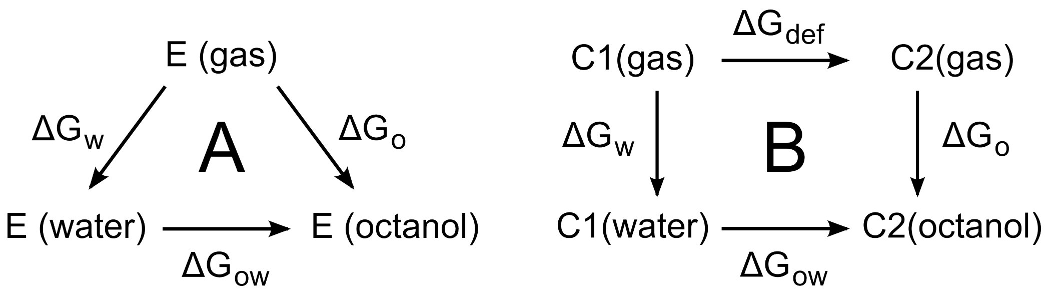


Figure 2 (one column): The mean radii of gyration calculated for 100 snapshots for all the molecules from the molecular dynamics trajectories in explicit water (the upper panel) and water-saturated octanol (the lower panel). The rigid molecules are shown in blue, the flexible molecules are in red (for the definitions of rigid and flexible molecules, see Section 2.4.2). The boxes, representing the standard deviations, are twice larger, for clarity.

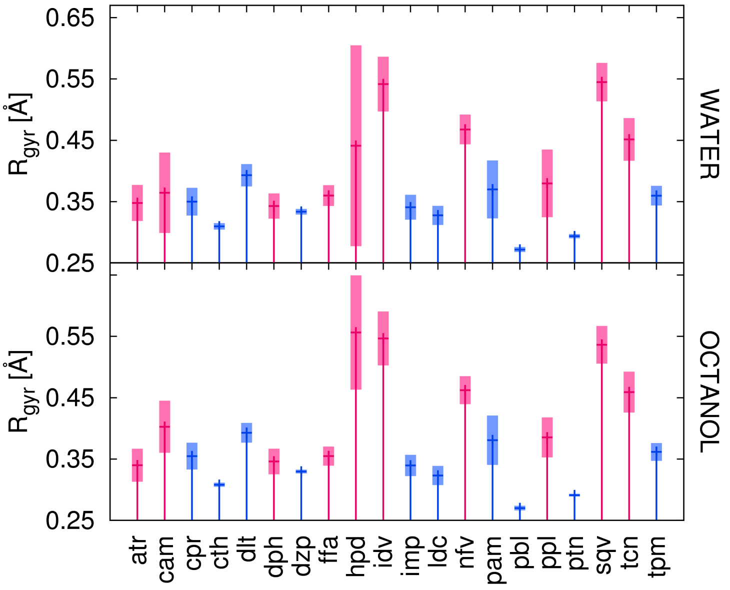


Figure 3 (one column): The compact and extended conformations of haloperidol (left) and procainamide (right) in water together with the alignments of their 100 snapshots. Both molecules contain the same number of *relevant rotatable bonds* (see Methods).

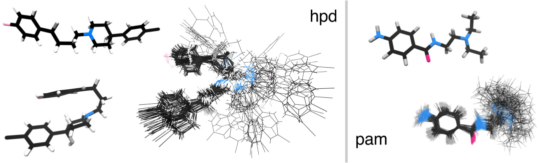


Figure 4 (one column): The correlation coefficient R^2 between the calculated and experimental water–octanol transfer free energies (the upper panel), the root-mean-square error (*RMSE*) (bottom left panel) and the mean signed absolute error (*MSAE*) (the bottom right panel). GB1, GB7 and PB are presented for the optimized geometries. There is only one transfer free energy estimator of the conformational ensemble in the COSMO-RS case (abbrev. C-RS, cross-hatched columns) – see the text for the details.

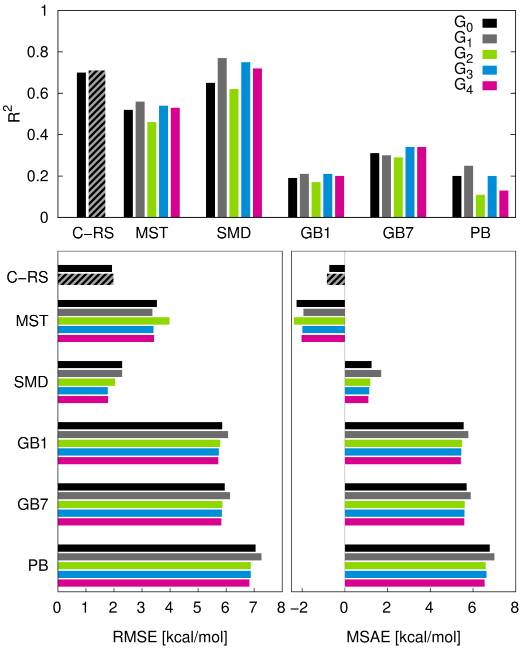


Figure 5 (one colunm): The correlation plots between selected hydration free energies and SMD hydration free energies. The correlation coefficients (*R^2*), the root-mean-square errors (*RMSE*) and the mean signed absolute errors (*MSAE*) are provided as insets.

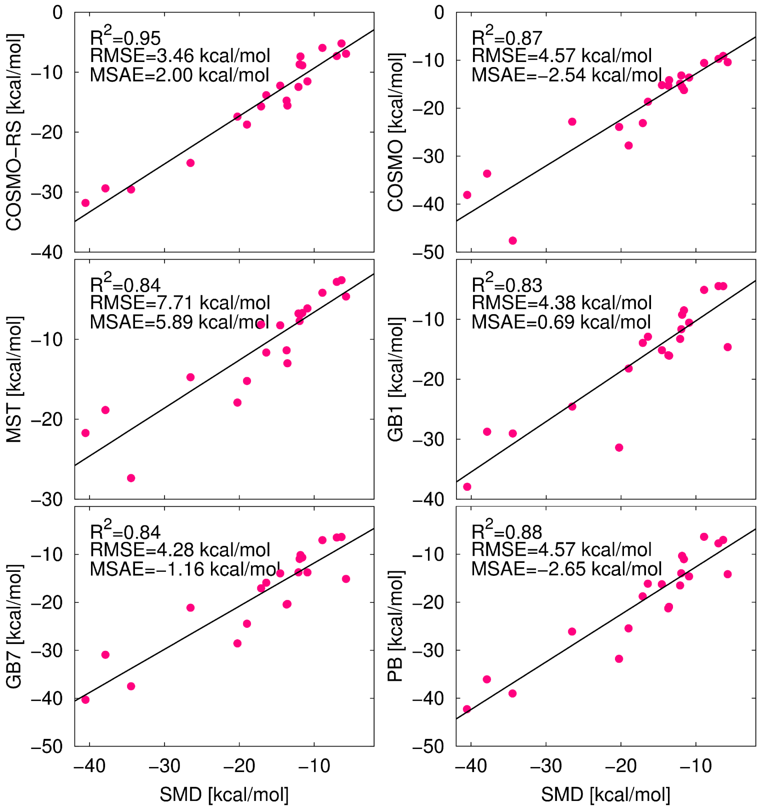


Figure 6 (one column): Left: the Gaussian probability density functions (pdf) (in arbitrary units) corresponding to the mean values and standard deviations of the atropine (atr) transfer free energy estimators using the SMD model. Right: the pdf of the water (wat) and octanol (oct) solvation free energies calculated for 100 snapshots.

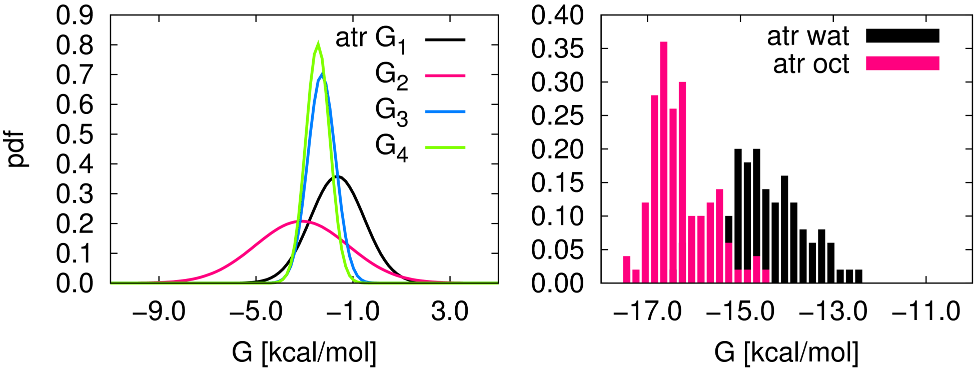
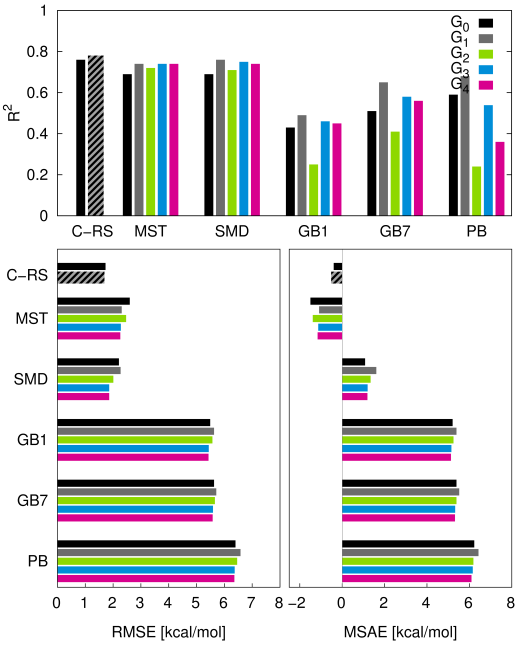


Figure 7 (one column): The correlation coefficient R^2 between the calculated and experimental water–octanol transfer free energies (the upper panel), the root-mean-square error (*RMSE*) (the bottom left panel) and the mean signed absolute error (*MSAE*) (the bottom right panel). GB1, GB7 and PB are presented for optimized geometries. There is only one transfer free energy estimator of conformational ensemble in the COSMO-RS case (abbrev. C-RS, cross-hatched columns) – see the text for the details. The values were calculated for the set of molecules excluding the HIV-1 protease inhibitors (idv, nfv, sqv).



Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches

Supplementary Information

Michal Kolář[a,b], Jindřich Fanfrlík[a], Martin Lepšík[a], Flavio Forti[c], F. Javier Luque[c] and Pavel Hobza[a,d,*]

[a] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 166 10 Prague, Czech Republic
[b] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague, Czech Republic
[c] Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, Campus de l'Alimentació, Santa Coloma de Gramenet, Spain
[d] Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

		R^2		$MSAE$		$RMSE$	
		OPT	SP	OPT	SP	OPT	SP
GB1	G0	0.19	0.17	5.56	5.44	5.86	5.75
	G1	0.21	0.16	5.78	5.96	6.07	6.27
	G2	0.17	0.29	5.49	-0.97	5.79	2.42
	G3	0.21	0.23	5.45	5.44	5.74	5.72
	G4	0.20	0.23	5.43	5.39	5.72	5.67
GB7	G0	0.31	0.27	5.70	5.59	5.95	5.86
	G1	0.30	0.22	5.89	6.10	6.14	6.38
	G2	0.29	0.31	5.61	-0.83	5.87	2.35
	G3	0.34	0.34	5.60	5.61	5.85	5.85
	G4	0.34	0.35	5.59	5.56	5.83	5.81
PB	G0	0.20	0.16	6.78	6.61	7.05	6.89
	G1	0.25	0.24	7.00	6.97	7.26	7.23
	G2	0.11	0.42	6.59	0.05	6.88	1.95
	G3	0.20	0.23	6.63	6.68	6.88	6.93
	G4	0.13	0.22	6.54	6.59	6.83	6.84

Table S 1: Correlation coefficients (R^2), mean signed absolute errors ($MSAE$) and root mean square errors ($RMSE$) calculated for unoptimized (SP) and optimized (OPT) snapshots.

		R^2		$MSAE$		$RMSE$	
		1st	2nd	1st	2nd	1st	2nd
COSMO-RS		0.71	0.72	-0.87	-0.82	1.99	2.03
MST	G1	0.56	0.55	-1.94	-1.94	2.29	2.31
	G2	0.43	0.49	-2.48	-2.30	2.08	2.07
	G3	0.54	0.55	-1.99	-1.99	1.80	1.76
	G4	0.54	0.54	-2.03	-2.02	1.80	1.79
SMD	G1	0.76	0.78	1.68	1.73	3.35	3.40
	G2	0.61	0.64	1.20	1.18	4.14	3.82
	G3	0.74	0.75	1.15	1.13	3.39	3.41
	G4	0.72	0.73	1.09	1.11	3.44	3.41
GB1	G1	0.17	0.15	5.92	6.00	6.22	6.34
	G2	0.25	0.29	-1.27	-0.67	2.81	2.26
	G3	0.23	0.23	5.44	5.44	5.72	5.73
	G4	0.23	0.23	5.39	5.39	5.68	5.67
GB7	G1	0.22	0.20	6.08	6.13	6.35	6.42
	G2	0.27	0.31	-1.11	-0.54	2.71	2.23
	G3	0.34	0.34	5.61	5.61	5.86	5.86
	G4	0.35	0.35	5.57	5.56	5.81	5.81
PB	G1	0.24	0.24	7.00	6.94	7.26	7.21
	G2	0.36	0.41	-0.19	0.28	2.24	1.95
	G3	0.23	0.23	6.68	6.68	6.93	6.92
	G4	0.22	0.22	6.59	6.60	6.84	6.84

Table S 2: Correlation coefficients (R^2), mean signed absolute errors ($MSAE$) and root mean square errors ($RMSE$) calculated for the first and second halves of the snapshot series.

		rmsd	max	worst	no. < 0.1 kcal/mol
COSMO-RS		0.46	0.97	imp	5
MST	G1	0.28	0.95	idv	10
	G2	0.49	1.66	sqv	2
	G3	0.14	0.48	idv	16
	G4	0.11	0.28	ldc	16
SMD	G1	0.35	0.85	sqv	7
	G2	0.31	0.73	ldc	6
	G3	0.15	0.43	nfv	11
	G4	0.14	0.43	ldc	14
GB1	G1	0.47	1.62	sqv	9
	G2	1.57	4.26	imp	2
	G3	0.04	0.13	sqv	19
	G4	0.01	0.03	ppl	21
GB7	G1	0.35	1.21	sqv	10
	G2	1.55	4.29	imp	2
	G3	0.03	0.11	sqv	20
	G4	0.01	0.03	tpm	21
PB	G1	0.33	0.96	idv	10
	G2	1.53	4.22	imp	2
	G3	0.06	0.19	idv	19
	G4	0.03	0.07	ldc	21

Table S 3: The first and second halves deviations in the sets. Max = maximum absolute deviation, worst = compound with max, no. < 0.1 stand for the number of compound which had absolute deviation lower than 0.1 kcal/mol (i.e. number of the most converged compounds).

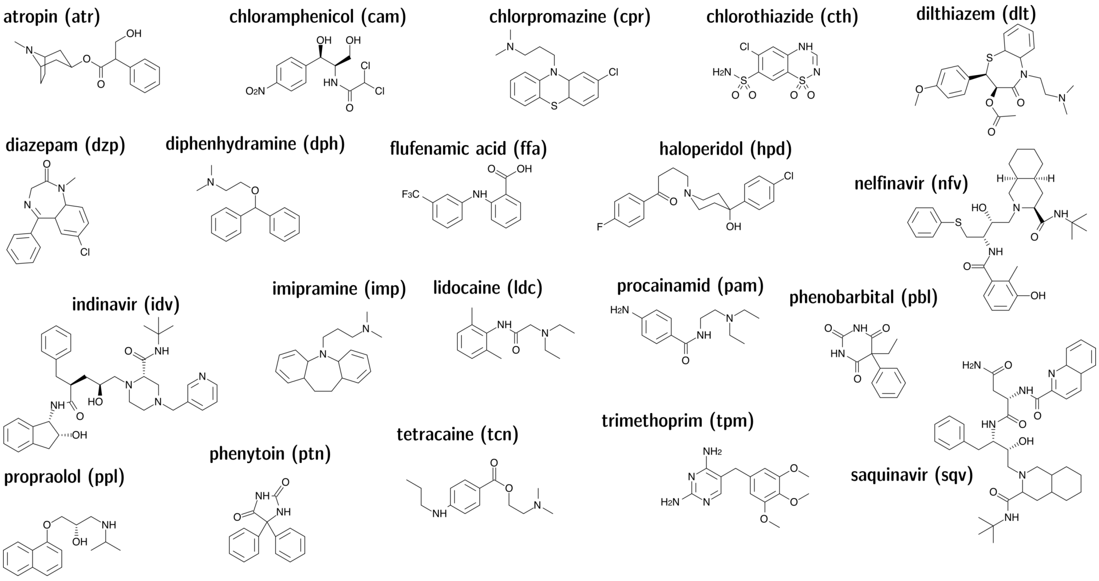


Figure S 1: Structural formulas, names and abbreviations of the molecules investigated

G_0

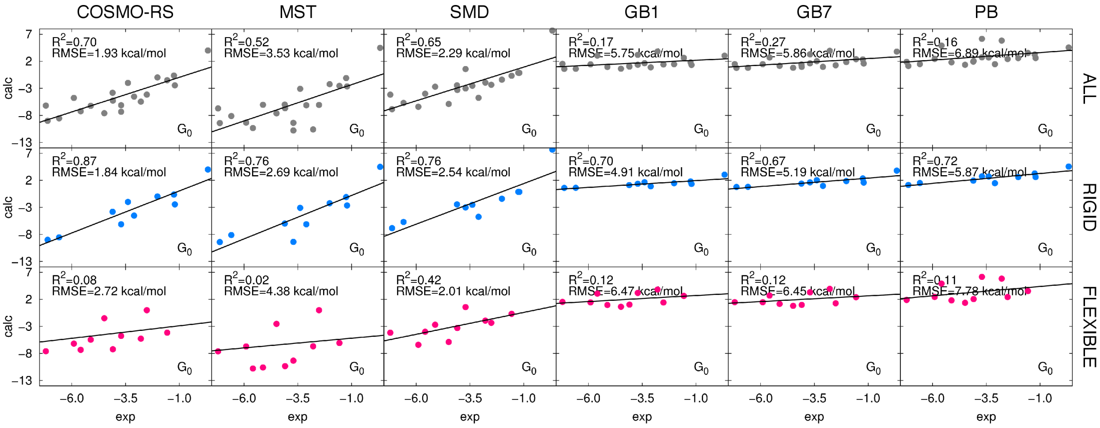


Figure S 2: Correlation plots between calculated and experimental values of water-octanol transfer free energies. The entire drug series (”ALL” in gray), rigid subset (blue) and flexible subset (red) are provided. All values are in kcal/mol. The calculated values correspond to the G_0 estimator.

G_1

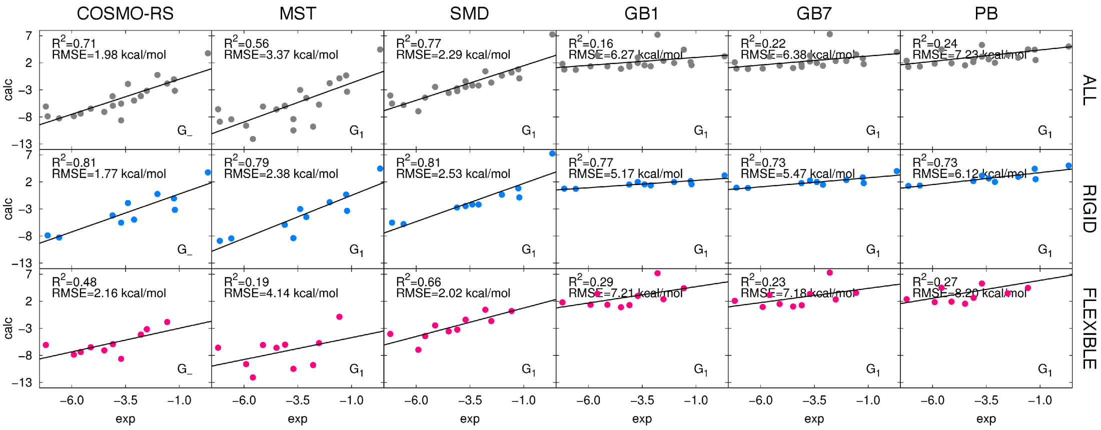


Figure S 3: Correlation plots between calculated and experimental values of water-octanol transfer free energies. The entire drug series (”ALL” in gray), rigid subset (blue) and flexible subset (red) are provided. All values are in kcal/mol. The calculated values correspond to the G_1 estimator.

G_2

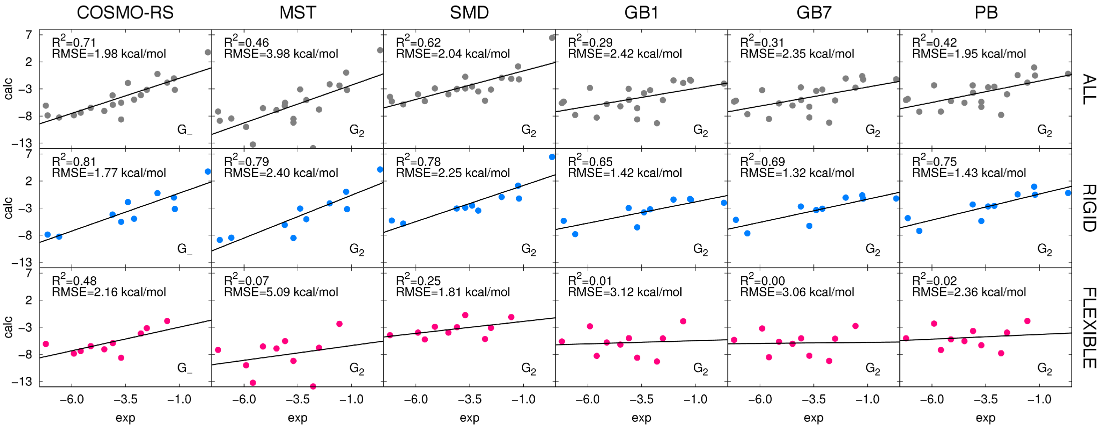


Figure S 4: Correlation plots between calculated and experimental values of water-octanol transfer free energies. The entire drug series (”ALL” in gray), rigid subset (blue) and flexible subset (red) are provided. All values are in kcal/mol. The calculated values correspond to the G_2 estimator.

G_3

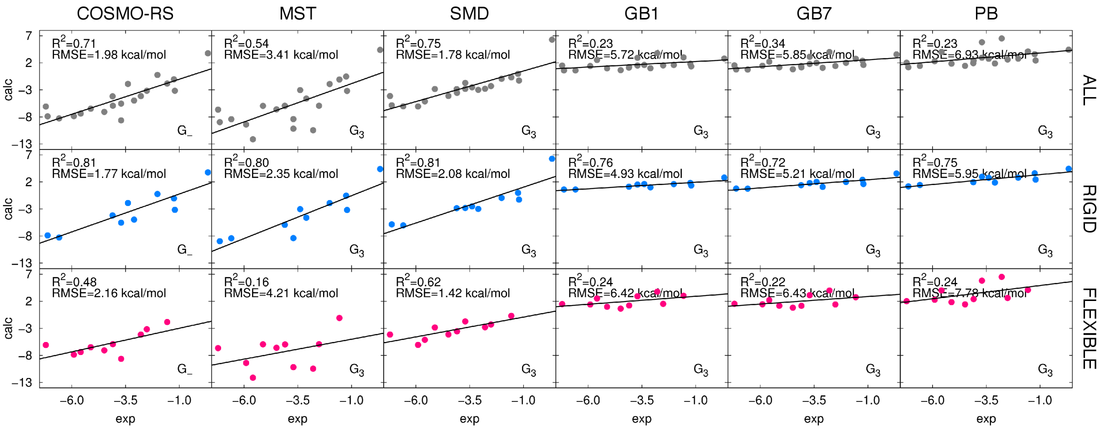


Figure S 5: Correlation plots between calculated and experimental values of water-octanol transfer free energies. The entire drug series (”ALL” in gray), rigid subset (blue) and flexible subset (red) are provided. All values are in kcal/mol. The calculated values correspond to the G_3 estimator.

G_4

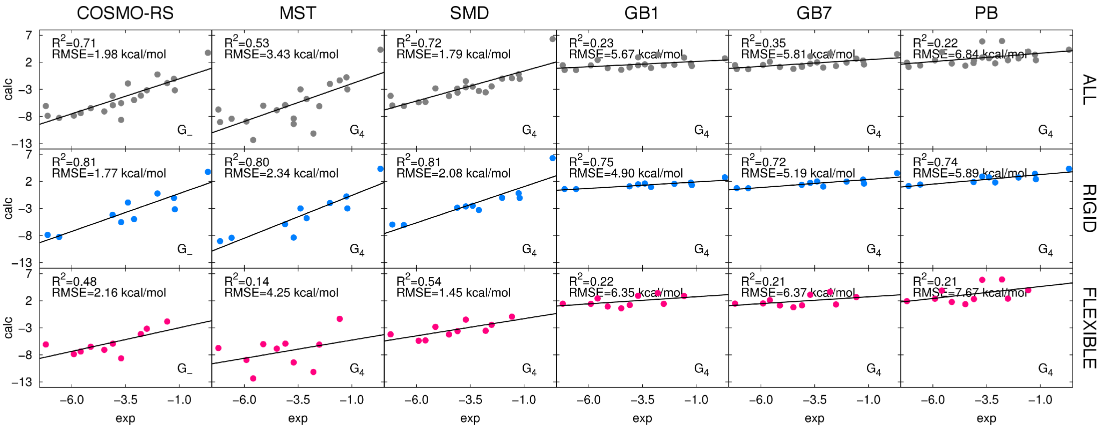


Figure S 6: Correlation plots between calculated and experimental values of water-octanol transfer free energies. The entire drug series (”ALL” in gray), rigid subset (blue) and flexible subset (red) are provided. All values are in kcal/mol. The calculated values correspond to the G_4 estimator.

E

Publication 4 – Hexahalogenbenzene Crystals

The Differences in the Sublimation Energy of Benzene and Hexahalogenbenzenes Are Caused by Dispersion Energy

Jakub Trnka, [a] Robert Sedlak,[a,b] Michal Kolář,[a,b] Pavel Hobza [a,c*]

[a] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo nam 2, 166 10 Prague, Czech Republic

[b] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague, Czech Republic

[c] Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

0 Abstract

The crystals of benzene and hexahalogenbenzenes have been studied by means of the density-functional theory augmented by an empirical dispersion correction term as well as by the symmetry-adapted perturbation theory. In order to elucidate the nature of noncovalent binding, pairwise interactions have been investigated. It has been demonstrated that the structures of dimers with the highest stabilization energy differ notably along the crystals. It has been shown that the differences in the experimental sublimation energies might be attributed to the dispersion interaction. To our surprise, the dihalogen bonding observed in the hexachloro- and hexabromobenzenes plays a rather minor role in energy stabilization, because they are energetically comparable with the other binding motifs. However, the dihalogen bond is by far the most frequent binding motif in hexachloro- and hexabromobenzenes.

0 Keywords

benzene and hexahalogenbenzene crystals, sublimation energy, interaction energy, DFT-SAPT, dihalogen bond

1 Introduction

The benzene dimer is one of the most studied aromatic molecular clusters, which arises among other things from the importance of the stacking $\pi...\pi$ interaction.[1-5] Two dimer structures are supposed to coexist at the respective potential energy surface: the T-shaped, or nearly T-shaped, structure and the parallel-displaced (PD) structure. The parallel C_{2h} structure, which was expected to be the global minimum (because of the maximal overlap), is actually penalized by the quadrupole-quadrupole (Q-Q) electrostatic interaction which is repulsive here.[6] The Q-Q interaction becomes less repulsive or attractive in the case of PD and T-shaped structures, respectively. Evidently, the electrostatic energy plays an important role in the interaction of benzene molecules, and it is thus not surprising that there have been attempts to interpret the sublimation energy of the benzene crystal only in terms of electrostatic quadrupole energy.[7] The resulting sublimation energy of 10.7 kcal/mol agreed exactly with the respective experimental value.[7] When passing to hexahalogenbenzenes, the quadrupole moment remains the first non-zero multipole moment, and it is hence possible to expect that the sublimation energies of hexahalogebenzenes will be determined dominantly also by the electrostatic Q-Q interaction.

Table 1 shows the quadrupole moments, polarizabilities and sublimation energies of benzene and hexahalogenbenzenes; a quick inspection of the quadrupole moments and the respective sublimation energies reveals no correlation between them. The quadrupole moments of hexafluorobenzene (C_6F_6) and benzene have the opposite sign, but their absolute values are similar (the former is slightly larger). With respect to this fact, we could expect the sublimation energy of the C_6F_6 to be slightly larger than that of the benzene, which actually holds true (cf. Table 1). When passing from hexafluorobenzene to hexachlorobenzene (C_6Cl_6), the situation is dramatically changed and the quadrupole moment of the latter molecule is more than order of magnitude smaller. The sublimation energy of C_6Cl_6, however, has increased. Evidently, the assumption that the sublimation energy of hexahalogenbenzenes is determined by electrostatic quadrupole energy is not fulfilled and other energy terms may also have their contribution. In Table 1, we can find a close correlation between the polarizabilities and the sublimation energies, which tells us that the dispersion energy plays an important role in the interaction between hexahalogenbenzenes, because there is a direct connection between the molecular polarizability and dispersion forces.

In the case of benzene dimer (or the crystal), both the electrostatic and dispersion energies are dominant attractive energy terms while the induction term (quadrupole – induced dipole) is much smaller. These two terms are thus responsible for the structure determination, and the relevant dimer structures should be localized in the crystal structure. The situation is exactly the same in the case of hexafluorobenzene. With hexachloro- and hexabromobenzenes, this is no longer valid, because a new interaction motif appears here.

Specifically, the dihalogen bond is formed between two molecules of hexachlorobenzenes or hexabromobenzenes, namely between a halogen, X_1 (Cl, Br, I), which is covalently bound to a less electronegative atom (e.g. carbon), and another halogen, X_2 (C-X_1...X_2).[8,9] This counterintuitive interaction is explained by the fact that a halogen atom is not isotropically negatively charged but it has a region with a positive electrostatic potential located on its top.

This region is usually called a σ-hole;[10] it is depicted in Figure 1 as the blue disc on the halogens in C_6Cl_6 and C_6Br_6. Generally, in a R_1-X_1...X_2-R_2 complex when the $R_1X_1X_2$ and $X_1X_2R_2$ angles are both close to 180 degrees, the interaction of two positive σ-holes is repulsive, resulting from the Coulomb law. However, when one of the respective angles is about 90 degrees while the other remains to be 180 degrees, the positive σ-hole interacts with the negatively charged ring of the atom and the resulting interaction energy is attractive. The strength of the dihalogen bond is expected to increase with the atomic number of the halogens; in other words, the C-Cl...Cl dihalogen bond is weaker than the C-Br...Br or C-I...I bonds. The σ-hole also exists at fluorine covalently bound to carbon, but this is typical only for small inorganic compounds such as NCF and not for aromatic species.[11,12] Consequently, the C-F...F dihalogen bonds between two C_6F_6 are mostly impossible to form. It has to be added that in the case of the dihalogen bond the dominant energy term is dispersion energy followed by electrostatic energy.[13] The important contribution of dispersion energy can be easily explained by the short distance between two heavy halogens possessing high polarizabilities.

The aim of the present study is to examine the nature of noncovalent binding within the crystals of C_6X_6 benzenes (X= H, F, Cl, Br). Specifically, we identify the binding motifs in various dimer structures appearing in these crystals. An attempt is made to correlate the experimental sublimation energy with the total interaction energies calculated for the crystal structures.

2 Methods

2.1 Structure preparation

The X-ray structures of the hexahalogenbenzene crystals were obtained from the Cambridge Structural Database.[14,15] The X-ray structure of the benzene crystal[16] was obtained from the Crystallography Open Database (21000348.cif).[17] Subsequently, it was processed using the JMol program.[18]

Within each crystal, the pairwise interactions were identified in the following manner: a reference molecule was chosen arbitrarily and 20 pairs were created. Each pair contains the reference molecule and one of the 20 nearest neighbors (cf. Figure S1 in the Supplementary Information).

2.2 Computations

The interaction energies for various dimers and for a large cluster, consisting of 21 molecules, were evaluated at the DFT/B3LYP-D3 level using the TZVPP basis set and the empirical pairwise dispersion contribution.[19] No deformation energy nor counterpoise correction were included. The interaction energy (ΔE) for a pair was determined as the difference between the energy of the dimer and the energies of both monomers (Equation 1):

$$\Delta E_{AB} = E(AB) - E(A) - E(B). \qquad (1)$$

The energy of the central reference molecule E(1) and the energy of the cluster containing all but the central molecule E(20) were subtracted from the energy of the entire cluster E(21), providing the *total interaction energy* ΔE_{tot} (Equation 2):

$$\Delta E_{tot} = E(21) - E(1) - E(20). \qquad (2)$$

Finally, the *average interaction energy* (ΔE_{aver}) was evaluated according to Equation 3:

$$\Delta E_{aver} = [\ E(21) - 21{\cdot}E(1)\ ] / 21, \qquad (3)$$

where E(21) stands for the energy of the entire cluster and E(1) is the energy of the central reference molecule. The energy decomposition for all the dimers was done by DFT-SAPT method using the aug-cc-pVDZ basis set.[20,21] The SAPT interaction energy (E_{INT}) was constructed as a sum of electrostatic (ES), induction (IND), dispersion (DISP) and exchange-repulsion (EXCH) terms. The exchange-induction and exchange-dispersion terms were added to the induction and dispersion energies, and, finally, the δ(HF) term was added to the induction energy. More details about the DFT-SAPT method can be found elsewhere.[22]

It is a known fact that using the DFT-SAPT decomposition with an aug-cc-pVDZ basis set provides an unconverged dispersion (DISP) contribution, while the other contributions are converged, indeed, when compared with the complete basis set limit values.[22] Hence, the dispersion contribution was scaled by a factor which was calculated as follows. For the most stable dimers, we performed calculations with the aug-cc-pVTZ and aug-cc-pVDZ basis sets and the scaling coefficients were obtained as the ratio between the dispersion term with the aug-cc-pVTZ and aug-cc-pVDZ basis sets. The coefficients are 1.42, 1.09, 1.09 and 1.12 for benzene, C_6F_6, C_6Cl_6, and C_6Br_6. For more details, cf. Table S1 in the Supplementary Information.

The calculations were carried out with Gaussian,[23] Molpro[24] and Grimme's DFT-D3[19] program packages.

3 Results and Discussion

3.1 Interaction energies

The total DFT-D3 interaction energies of the central reference molecule with the 20 neighboring molecules (cf. Figure S1) evaluated for four molecular crystals are presented in Table 2. Besides the total interaction energies, likewise their DFT and dispersion components are shown. Table 2 also shows the average interaction energies, and also here their DFT and dispersion components are presented.

The total interaction energies of benzene and C_6F_6 are almost equal, and also the DFT and dispersion components are roughly comparable. These results are not surprising regarding the molecular properties (cf. Table 1). However, the relatively large difference between the average interaction energies of C_6H_6 and C_6F_6 is surprising. This discrepancy may arise from the differences in the symmetry of particular crystal structures. This issue will be addressed in more details below.

When passing from C_6F_6 to C_6Cl_6 and C_6Br_6, a significant increase of the total stabilization energy and roughly the same increase of the average stabilization energy were found. In both cases, the dispersion contribution is much larger than in the previous two crystals and it is responsible for the total stabilization energy increase. Let us add that for all four crystals the DFT energy component is repulsive. The decomposition of the total interaction energy presented in Table 2 does not say anything about the nature of the stabilization of the particular pairs.

Table 3 shows the interaction energies for various pairs of benzene and hexahalogenbenzenes. The interaction energy is systematically determined using DFT-D3 and DFT-SAPT approaches, and various pairs are ordered along decreasing stabilization energy; only the pairs with the stabilization energy higher than 1.0 kcal/mol are presented. All pair interaction energies are provided in the Supplementary Information in Table S2. DFT-D3 stabilization energies are in all cases larger than the DFT-SAPT ones. This overestimation is the largest for hexafluorobenzene and benzene (39 and 31 %) while that for hexachloro- and hexabromobenzenes is considerably smaller (5 and 9 %). Evidently, the DFT-SAPT stabilization energies evaluated with the aug-cc-pVTZ basis set are more reliable and will be considered in the following text when analyzing the pair interactions.

Firstly, a quick inspection of the DFT-SAPT energies from Table 3 reveals a feature valid for almost all listed pair interactions. Not surprisingly, all dimers are mainly stabilized by dispersion and electrostatic interactions. Secondly, by comparing the pair interaction energies of C_6H_6 and C_6F_6 with C_6Cl_6 and C_6Br_6, we found an important difference. The stabilization energies for the most and least attractive pairs differ for the former two systems only marginally (by less than 1.2 kcal/mol) while this difference is much more pronounced for the latter two systems (8.1 and 9.4 kcal/mol, respectively). This difference can be documented also with the corresponding relative numbers. The relative increase of interaction from the weakest to the strongest dimer is 83%, 100%, 506% and 448% for C_6H_6, C_6F_6, C_6Cl_6 and C_6Br_6, respectively.

3.2 The relative importance of energy terms

We calculated the ratios of the dispersion and interaction energies (DISP/ E_{INT}) as well as of the electrostatic and interaction energies (ES/E_{INT}). They provide a picture on the balance between the two most important attractive forces. The ES/E_{INT} ratios averaged over the pairs with stabilization higher than 1 kcal/mol are 0.51, 0.54, 0.68 and 0.89 for C_6H_6, C_6F_6, C_6Cl_6 and C_6Br_6, respectively. Clearly, the relative importance of the electrostatic contribution increases with the atomic number of the halogen. However, the value of neither the quadrupole (cf. Table 1) nor the quadrupole-quadrupole electrostatic interaction can interpret these ratios. An important increase of this ratio when passing from C_6H_6 and C_6F_6 to C_6Cl_6 and C_6Br_6 could be connected with the fact that a new binding motif is created in the latter group of crystals. Selected dimers of C_6Cl_6 and C_6Br_6 are stabilized by dihalogen bonds which does not exist in the former two crystals. The value of the ES/E_{INT} ratio for the dihalogen-bonded dimers of the C_6Cl_6 and C_6Br_6 molecules is even more pronounced. The values of 0.70 and 0.94 support our previous statement. Hence, the mere formation of dihalogen bonds in selected dimers of C_6Cl_6 and C_6Br_6 can explain the increase of the ES/E_{INT} ratios for the C_6Cl_6 and C_6Br_6 dimers. The different electrostatic potential of C_6Cl_6 and C_6Br_6 with respect to the other two molecules, which is the reason for the formation of dihalogen-bond structures, may potentially be responsible for the increased value of the ES/E_{INT} ratio. A more detailed view on the electrostatic potentials of all four molecules will be presented below. In Table 3, other relatively interesting features can be observed. The PD structure is either the most stable or one of the most stable dimer structures. When investigating the ES DFT-SAPT energies for this structure, we found its dramatic increase for hexachloro- and hexabromobenzenes, which contradicts the decrease of the quadrupole moment when passing from C_6H_6 and C_6F_6 to C_6Cl_6 and C_6Br_6. Visualizing the PD structures of all crystals (Figure 2), we found that monomers in C_6Cl_6 and C_6Br_6 PD dimers are much closer to each other than in the C_6F_6 dimer; the distance between the centers of mass of the C_6F_6, C_6Cl_6 and C_6Br_6 crystals amounts to 5.76, 3.76 and 3.95 Å, respectively. A closer contact in the C_6Cl_6 and C_6Br_6 PD structures (which contradicts the larger vdW radii of Cl and Br than of F) is clearly due to very large dispersion energy (cf. Table 3), which pushes monomers together. The penetration energy, defined as a difference between SAPT electrostatic energy and multipole-multipole electrostatic energy, is negligible at the distances larger than equilibrium and becomes important (attractive) at shorter distances. Large SAPT electrostatic energies for the C_6Cl_6 and C_6Br_6 dimers are thus due to attractive penetration energies and have no connection with quadrupole-quadrupole electrostatic energy.

The DISP/E_{INT} ratios averaged over the pairs with stabilization higher than 1 kcal/mol are 1.56, 1.69, 2.00 and 1.99 for C_6H_6, C_6F_6, C_6Cl_6 and C_6Br_6, respectively. Unsurprisingly, the relative importance of the dispersion contribution is the lowest for C_6H_6 and the highest for C_6Cl_6 and C_6Br_6.

3.3 Structural analysis

The binding motifs between the central and neighboring molecules in our cluster models as well as the geometrical parameters of the individual dimers are discussed in the following subsection.

The differences in the binding motifs themselves, along with the different energetic degeneracy for all four molecular crystals, reveals that the relative arrangement of the molecules in the cluster models is different (cf. Table 3).

The highest degree of the energetic as well as binding motif degeneracy is exhibited by the benzene crystal. The twelve molecules which surround the central molecule are grouped into three structural motifs, each including four dimers (cf. the first part of Table 3 and Figure S1). Several structural motifs can be recognized: T-shape, distorted T-shape and L-shape.

The crystal of C_6F_6 possesses the lowest degree of structural motif and energetic degeneracy. The eleven neighboring molecules are divided into eight groups (cf. the second part of Table 3 and Figure S1). The three the most stable dimers correspond to the PD structures. The structures of the remaining eight dimers can be classified as T-shape or distorted T-shape structures. The least stable dimer (with stabilization < 1 kcal/mol) with the planar molecular structure is rare in the cluster model.

The crystals of C_6Cl_6 and C_6Br_6 are almost identical, hence possessing similar energetic and structural characteristics. The fourteen neighboring molecules are divided into five groups. The most stable are two PD structures followed by two planar structures with two dihalogen bonds. As already mentioned above, dimers with dihalogen bonds are considerably less stable than the PD structures. Another two dimers represent a distant PD structure. The eight least stable dimers were included in the category of distorted halogen-bonded structures. However, they represent two distinct stabilization levels (cf. the third and fourth parts of Table 3 and Figures 3 and S1).

One could expect that the similarity or the dissimilarity in the mutual arrangement of the neighboring molecules in the molecular crystals can be predicted for different chemical species based on the values of molecular properties, such as permanent multipole moments, polarizabilities etc. However, the crystal structure analysis showed that such an assumption would lead to wrong interpretations. The structural differences between the crystals of C_6H_6 and C_6F_6 are remarkable while the opposite is true when the crystals of C_6Cl_6 and C_6Br_6 are compared. Nevertheless, in the first example the values of molecular properties are very similar, whereas in the second there are significant differences (cf. Table 1). This leads us to the statement that more sophisticated approaches are necessary for the interpretation of the structural motif among noncovalently bound clusters.

In the next paragraphs, the geometrical parameters of individual dimers will be discussed. The most attractive pair of C_6H_6 is represented by the T-shape structure while the distorted T-shape and L-shaped structures are considerably less stable (by 29 and 43 %, respectively). The situation with the remaining three hexahalogenebenzenes is different, and here the most attractive pairs correspond to the PD structures. However, while the stabilization of the PD structure of C_6F_6 is comparable to that of the remaining structures, in the case of the other halogenbenzenes the difference is significant. A comparison of the PD structures of hexahalogenbenzenes brings quantitative differences. For C_6F_6, the distance between the centers of mass is 5.8 Å (Figure 2). On the other hand, in the case of chloro- and bromoderivates, the equivalent distance ranges between 3.8 and 4.0 Å, respectively. Hence, in the case of C_6F_6, the electrostatic and dispersion terms are much smaller. While for C_6Cl_6 and C_6Br_6 the PD structure is significantly more stable than the other structures, the situation for C_6F_6 is different and here the stability of PD and other structures (see below) differs only marginally. A further comparison of the most attractive PD structure for the three studied halogenbenzenes leads to the electrostatic term being larger for C_6Cl_6 and C_6Br_6 (than for C_6F_6) by 4.6 and 6.9 kcal/mol, respectively. This difference is, however, significantly larger (by 14.6 and 17 kcal/mol) for the dispersion contribution. Consequently, it is mostly the dispersion energy for the PD structures that makes the total stabilization energy of C_6Cl_6 and C_6Br_6 much larger than that of C_6F_6 (cf. the polarizabilities of hexahalogenbenzenes presented in Table 1).

Investigating other less stable pairs, we again found more pronounced differences between C_6H_6, C_6F_6, C_6Cl_6 and C_6Br_6. The three most stable structures of the second crystal possess a PD structure while all the others have a T-shaped structure.

The crystals of hexachloro- and hexabromobenzenes differ from the crystals of benzene and hexafluorobenzene by the presence of structures possessing dihalogen bonds (cf. Figure 3). There are two structures with two ("cyclic") dihalogen bonds for each crystal with stabilization energies of 2.1 and 2.9 kcal/mol for C_6Cl_6 and C_6Br_6, respectively. The $C_1X_1X_2$ angle (α) in these structures is, as it should be, almost linear (171 and 173 degrees for C_6Cl_6 and C_6Br_6, respectively), and the $X_1...X_2$ distance is 3.7 and 3.8 Å. The $X_1X_2C_2$ angle (β) is 123 degrees for C_6Cl_6 and C_6Br_6 (cf. Figure 3). Other dimer structures, named distorted dihalogen bonds, are not planar. One molecule is distorted from the imaginary plane (cf. Figure 3), hence the structure contains only one dihalogen bond. The arrangement of the CX_1X_2 atoms is almost identical as in the case of structures with two ("cyclic") dihalogen bonds. Originally, we expected that due to this rather short distance between heavy halogens the stabilization energy of the structures with dihalogen bonds will be significantly higher. From the Table 3 it is, however, evident that these stabilization energies are only slightly larger than the stabilization energies of other structures.

Investigating different structures of hexafluorobenzene, whose stabilization energy exceeds 1 kcal/mol, we found neither planar nor distorted structures with a difluoro noncovalent bond. This is caused by the fact that fluorine covalently bound to an aromatic ring usually does not exhibit a σ-hole, which is a prerequisite for the existence of halogen bonding (cf. Figure 1). This significant difference between the electrostatic potential of C_6F_6 and C_6Cl_6 (together with C_6Br_6) crystals can be seen as the reason for the significant differences in the crystal structures (cf. Figure 1). The region of the positive electrostatic potential (σ-hole), present at the top of each chlorine and bromine atom in a hexahalogenbenzene molecule (cf. Figure 1), is the moiety via which the intermolecular interaction is realized (cf. Figure 3).

Nevertheless, the stabilization energies of various hexafluorobenzene structures mostly having the T-shaped structure without a direct X...X interaction are comparable to the stabilization energies of the structures possessing dihalogen bonds.

3.4 Discussion

Similar total interaction energies (1+20) of benzene and hexafluorobenzene agree with similar sublimation energies of these two crystals, and the much larger total interaction energy of hexachlorobenzene again agrees with its much larger sublimation energy. The relatively large difference in the average interaction energy of C_6H_6 and C_6F_6 (of as much as 75%) can be interpreted as a consequence of a different spatial orientation of the pairs within the clusters considered. Comparing the entire cluster model of the C_6H_6 and C_6F_6 (cf. Figure S1) crystals, it is evident that the former one is spherically less symmetric than the latter one, which means that the molecules around the central one are ordered less compactly. Hence, the average stabilization energy of the C_6H_6 molecule is substantially smaller.

The significant differences between the binding motif of the most stable dimer of C_6H_6 and C_6F_6 crystals (T-shape and PD structure) can be seen as a consequence of a subtle difference in the electrostatic potential. In the case of the C_6H_6 molecule, where the hydrogen-atom regions are represented by a continuously increased positive potential (cf. Figure 1), the T-shape conformer is energetically more preferred. On the other hand, the electrostatic potential of C_6F_6 in the regions of fluorine atoms does not show the same properties. Even though the fluorine atoms are surrounded by a negative region of the potential, an increase of the potential on top of each fluorine can be observed. This is a consequence of a mutual electron repulsion, hence the T-shape structure is not as preferred as the PD structure.

4 Conclusions

i) Both the total and the average interaction energies increase when passing from benzene through hexafluorobenzene over hexachlorobenzene, and this increase is proportional to the increase of sublimation energy.

ii) The most favorable pair structure with benzene corresponds to the T-shaped structure while that for hexahalogenbenzenes corresponds systematically to the PD structure. Because of the much higher polarizability of the hexachloro- and hexabromobenzene, the dispersion energy in this structure is also much higher than that in the hexafluorobenzene. The significant increase in the total interaction energy as well as in the experimental sublimation energy when passing from hexafluorobenzene to hexachloro- and hexabromobenzene is thus mainly caused by the increase in dispersion energy. Indeed, the DFT-SAPT decomposition shows that the dominant part of the interaction energy originates in the dispersion energy. Nevertheless, the relative importance of the electrostatic contribution increases when passing to heavier halogens, in case of hexabromobenzene it is at the expense of the dispersion term.

iii) The new structural type, found in the crystals of hexachloro- and hexabromobenzenes, is stabilized by dihalogen bonds. However, the stabilization energies of these structures do not differ much from the stabilization energies of other, mainly T-shaped, structures of hexahalogenzenzene. The existence of the structures with dihalogen bonds thus cannot be responsible for the higher total interaction and sublimation energies of hexachloro- and hexabromobenzene. However, the presence of dihalogen bonds in hexachloro- and hexabromobenzenes has a crucial role for the determination of geometries of their crystals.

5 Acknowledgement

This work was part of the Research Project RVO: 61388963 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. This work was also supported by the Czech Science Foundation [P208/12/G016] and the operational program Research and Development for Innovations of the European Social Fund (CZ 1.05/2.1.00/03/0058).

6 Supplementary Information

The xyz coordinates of the clusters, the details on the DFT-SAPT dispersion contribution, the interaction energies and decompositions for all pairs and the depiction of the cluster models are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

7 Bibliography

1 Watson, J. D.; Crick, H. C. D. *Nature* **1953**, 171, 737-738.

2 Lerman, L. S. *J. Mol. Biol.* **1961**, 3 (1), 18-20.

3 Burley, S. K., Petsko, G. A. *Science* **1985**, 229 (4708), 23-28.

4 Hunter, C. A.; Singh, J.; Thornton, J. M. *J. Mol. Biol.* **1991**, 218, 838-846.

5 McGaughey, G. B.; Gagne, M.; Rappe, A. K. *J. Biol. Chem.* **1998**, 273, 15458-15463.

6 Battaglia, M. R.; Buckingham, A. D.; Williams, J. H. *Chem. Phys. Lett.* **1981**, 78, 421–423.

7 Sen, P; Basu, S. *J. Chem. Phys.* **1968**, 48, 4075-4076.

8 Politzer, P.; Murray, J. S.; Concha, M. C. *J. Mol. Model.* **2008**, 14, 659-665.

9 Politzer, P.; Riley, K. E.; Bulat, F. A.; Murray, J. S. *J. Comput. Theor. Chem.* **2012**, 998, 2-8.

10 Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2007**, *13*, 291-296.

11 Politzer, P.; Lane, P.; Concha, M.; Ma, Y.; Murray, J. *J. Mol. Modeling* **2007**, 13, 305-311.

12 Metrangolo, P.; Murray, J. S.; Pilati, T.; Politzer, P.; Resnati, G.; Terraneo, G. *Crystal Growth & Design* **2011**, *11*, 4238-4246.

13 Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2008**, 4, 232-242.

14 Boden, N.; Davis, P. P.; Stam, C. H.; Wesselink, G. A. *Mol. Phys.* **1973** 25, 81-86.

15 Reddy, C. M.; Kirchner, M. T.; Gundakaram, R. C.; Padmanabhan, K. A.; Desiraju, G. R. *Chem.-Eur. J.* **2006** , 12, 2222-2234.

16 Budzianowski A.; Katrusiak A. *Acta Cryst.* **2006**, B62, 94-101.

17 webpage: http://www.crystallography.net

18 Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

19 Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, 132, 154104-154120.

20 Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, 94, 1887–1930.

21 Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, 105, 646-659.

22 A. Haselmann, G. Jansen, and M. Schütz, *J. Chem. Phys.* **2005**, 122, 014103-014119.

23 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.;

Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, *Revision A.1*; Gaussian, Inc.: Wallingford, CT, **2009**.

24 Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Schuetz, M.; Celani, P.; Knizia, G.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; O. Deegan, M. J.; Dobbyn, A. J.; Eckert, F.; Goll, F.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.;Koeppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflueger, K.; Pitzer, K.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. MOLPRO, version 2010.1, a package of ab initio programs; molpro, **2010**.

Table 1: The quadrupole moments (Q, a. u.), polarizabilities (α, Å^3) and sublimation energies (E^{sub}, kcal/mol) of the C_6X_6 (X=H, F, Cl, Br) systems

	Q	α	E^{sub}
C_6H_6	–6.59	56.23	10.7
C_6F_6	7.89	57.24	11.8
C_6Cl_6	0.25	120.63	23.8
C_6Br_6	–4.72	152.93	-

Table 2: The interaction energies (in kcal/mol) of the central molecule with the 20 neighboring molecules (the Total columns) and the average interaction energies (the Average columns) for the clusters are shown. Its DFT and dispersion components are provided.

	Total			Average		
	$\Delta E_{DFT+Disp}$	ΔE_{DFT}	ΔE_{Disp}	$\Delta E_{DFT+Disp}$	ΔE_{DFT}	ΔE_{Disp}
C_6H_6	–27.6	4.6	–32.2	–6.0	1.0	–7.0
C_6F_6	–27.9	1.0	–28.8	–10.5	–2.5	–8.0
C_6Cl_6	–45.6	19.3	–64.9	–13.5	5.5	–19.0
C_6Br_6	–61.6	24.3	–85.9	–17.6	7.3	–24.9

Table 3: The DFT-D3 and DFT-SAPT pair interaction energies (ΔE and E_{INT}) for the energetically most favorable pairs. The numbers in parentheses refer to the absolute value of the dispersion component of the DFT-D3 energy, and besides the total DFT-SAPT interaction energy (E_{INT}) also its components, electrostatic (ES), induction (IND) and dispersion (DISP) are presented; all energies are listed in kcal/mol. The table shows only pairs with the DFT-SAPT stabilization energy larger than 1.0 kcal/mol; deg stands for the degeneracy level of the particular structure (deg. = n means that n+1 identical structures exist).

			DFT-D3	DFT-SAPT			
molecule	bind. motif	deg.	–ΔE	–ES	–IND	–DISP	$–E_{INT}$
C_6H_6	T-shape	3	2.8 (3.3)	1.2	0.1	3.3	2.2
	distorted T-shape	3	2.0 (2.4)	1.0	0.0	2.4	1.5
	L-shape	3	1.6 (2.0)	0.4	0.0	1.9	1.2
C_6F_6	PD	1	3.3 (2.9)	1.4	0.1	3.7	2.4
	distant PD	0	3.3 (2.7)	1.7	0.1	3.7	2.4
	distorted T-shape	0	3.0 (2.8)	1.0	0.1	3.6	2.3
	distorted T-shape	0	2.7 (2.9)	1.0	0.1	3.7	1.9
	distorted T-shape	0	2.5 (2.4)	1.1	0.1	3.2	1.8
	distorted T-shape	0	2.4 (2.2)	1.1	0.1	2.9	1.7
	distorted T-shape	1	2.0 (1.7)	0.5	0.0	2.0	1.6
	distorted T-shape	1	1.9 (2.1)	0.7	0.0	2.8	1.2
C_6Cl_6	PD	1	11.5 (16.6)	6.0	0.4	19.6	9.7
	dihalogen bonded	1	2.0 (2.6)	1.2	0.1	3.8	2.1
	distorted T-shape	3	1.9 (2.4)	1.2	0.1	3.4	1.9
	distant PD	1	1.9 (2.9)	1.1	0.0	3.7	1.7
	distorted T-shape	3	1.6 (2.3)	1.3	0.1	3.4	1.6
C_6Br_6	PD	1	14.1 (20.9)	8.3	0.5	22.8	11.5
	dihalogen bonded	1	3.0 (3.8)	2.3	0.3	5.3	2.9
	distorted T-shape	3	2.7 (3.4)	2.3	0.3	4.8	2.7
	distorted T-shape	3	2.1 (3.1)	2.4	0.3	4.8	2.2
	distant PD	1	2.6 (4.1)	1.8	0.1	4.6	2.1

Figure 1 (two columns)

The electrostatic potential mapped on the 0.001 e/bohr^3 electron isodensity surface of C_6X_6 (X=H, F, Cl, Br). The maps were determined at the B3LYP/TZVPP level for the central reference molecule of the crystal model and for B3LYP/6-311+G* optimized monomers. The color scale is in a. u.

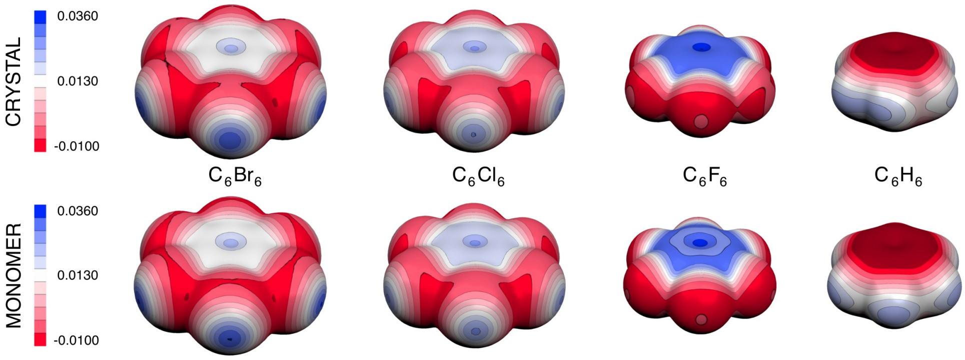


Figure 2 (one column)

The most stable pair structures for benzene, hexafluorobenzene, hexachlorobenzene and hexabrombenzene; the colors: silver = C, white = H, pink = F, orange = Cl and green =Br; A/ the side view; B/ the perspective view.

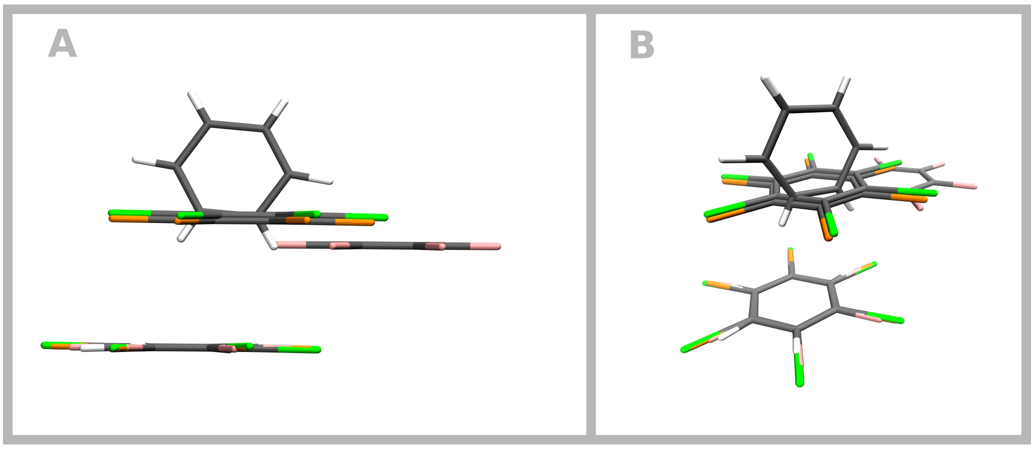
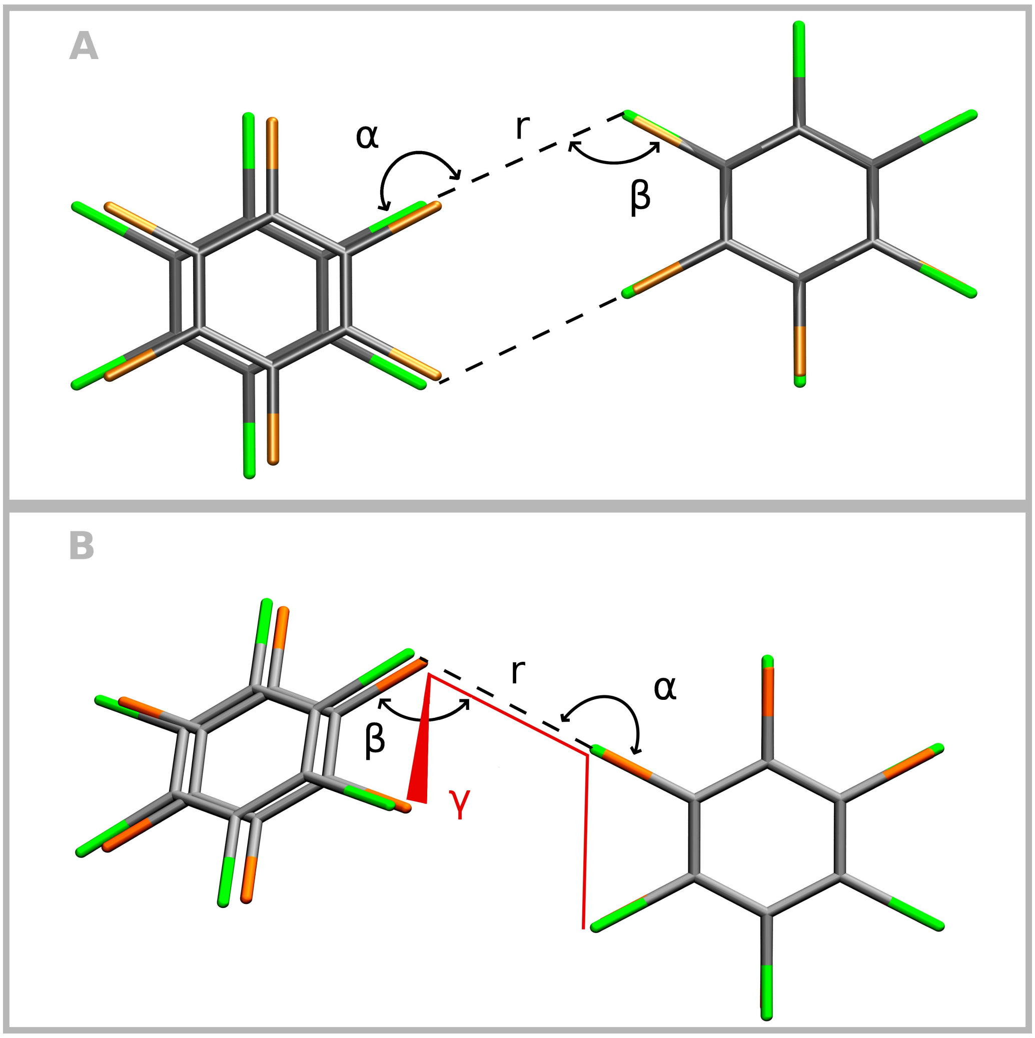


Figure 3 (one column)

A/ the structures of the planar dihalogen-bonded dimer of hexachloro- and hexabromobenzene, with two (“cyclic”) dihalogen bonds: α = 171°, β = 123°, r = 3.7 Å and α = 173°, β = 123°, r = 3.8 Å for C_6Cl_6 and C_6Br_6, respectively; the top view; B/ the structures of the distorted dihalogen-bonded dimer of hexachloro- and hexabromobenzene, with one dihalogen bond: α = 175°, β = 117°, γ = 35°, r = 3.4 Å and α = 174°, β = 115°, γ = 35°, r = 3.5 Å for C_6Cl_6 and C_6Br_6, respectively; the perspective view; the colors: silver = C, orange = Cl and green =Br.



The Differences in the Sublimation Energy of Benzene and Hexahalogenbenzenes Are Caused by Dispersion Energy

Supplementary Information

Jakub Trnka, [a] Robert Sedlak,[a,b] Michal Kolář,[a,b] Pavel Hobza [a,c*]

[a] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 166 10 Prague, Czech Republic

[b] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague, Czech Republic

[c] **Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic**

Table S1

The DFT-SAPT dispersion (DISP) terms of the "stacked" dimers for all 4 molecules investigated. The calculation preformed with two consistent Dunning basis sets of increasing size: aug-cc-pVDZ (aDZ) and aug-cc-pVTZ (aTZ). The scaling coefficient (the last column) calculated as the ratio between the values DISP[aTZ] and DISP[aDZ]; all the energies are listed in kcal/mol.

molecule	pair	–DISP[aDZ]	–DISP[aTZ]	coefficient
benzene	1-5	2.29	3.26	1.42
hexafluorobenzene	1-40	3.35	3.65	1.09
hexachlorobenzene	1-75	17.97	19.61	1.09
hexabromobenzene	1-14	20.32	22.82	1.12

Table S2

The DFT-D3 and DFT-SAPT interaction energies (in kcal/mol) for all of the pairs.

benzene

	DFT-SAPT					DFT-D3		
pair	ES	IND	DISP	Exch	E_{INT}	DFT	D3	E_{Tot}
1-5	-1.18	-0.27	-3.26	2.53	-2.17	0.44	-3.28	-2.84
1-4	-1.18	-0.27	-3.26	2.53	-2.17	0.44	-3.28	-2.84
1-10	-1.18	-0.27	-3.26	2.53	-2.17	0.44	-3.28	-2.84
1-19	-1.18	-0.27	-3.26	2.53	-2.17	0.44	-3.28	-2.84
1-6	-0.99	-0.22	-2.42	2.09	-1.54	0.34	-2.37	-2.03
1-7	-0.99	-0.22	-2.42	2.09	-1.54	0.34	-2.37	-2.03
1-17	-0.99	-0.22	-2.42	2.09	-1.54	0.34	-2.37	-2.03
1-20	-0.99	-0.22	-2.42	2.09	-1.54	0.34	-2.37	-2.03
1-8	-0.40	-0.12	-1.85	1.22	-1.15	0.38	-2.00	-1.62
1-9	-0.40	-0.12	-1.85	1.22	-1.15	0.38	-2.00	-1.62
1-14	-0.40	-0.12	-1.85	1.22	-1.15	0.38	-2.00	-1.62
1-21	-0.40	-0.12	-1.85	1.22	-1.15	0.38	-2.00	-1.62
1-2	-0.09	-0.01	-0.38	0.02	-0.46	-0.05	-0.45	-0.50
1-3	-0.09	-0.01	-0.38	0.02	-0.46	-0.05	-0.45	-0.50
1-13	-0.06	-0.00	-0.20	0.00	-0.26	-0.04	-0.25	-0.29
1-18	-0.06	-0.00	-0.20	0.00	-0.26	-0.04	-0.25	-0.29
1-11	-0.02	-0.00	-0.08	0.00	-0.09	-0.01	-0.09	-0.10
1-15	-0.02	-0.00	-0.08	0.00	-0.09	-0.01	-0.09	-0.10
1-12	-0.00	-0.00	-0.04	-0.00	-0.04	-0.00	-0.04	-0.04
1-16	-0.00	-0.00	-0.04	-0.00	-0.04	-0.00	-0.04	-0.04

hexafluorobenzene

	DFT-SAPT					DFT-D3		
pair	ES	IND	DISP	Exch	E_{INT}	DFT	D3	E_{Tot}
41-40	-1.44	-0.25	-4.76	2.97	-3.48	-0.45	-2.85	-3.30
41-42	-1.45	-0.25	-4.76	2.97	-3.49	-0.45	-2.85	-3.30
41-95	-1.69	-0.18	-4.82	3.21	-3.49	-0.60	-2.74	-3.34
41-128	-0.98	-0.17	-4.64	2.41	-3.38	-0.20	-2.82	-3.02
41-125	-1.04	-0.18	-4.83	3.05	-3.00	0.14	-2.86	-2.72

41-162	-1.14	-0.18	-4.18	2.74	-2.77	-0.15	-2.39	-2.54
41-163	-1.07	-0.14	-3.83	2.42	-2.62	-0.29	-2.15	-2.44
41-17	-0.46	-0.11	-2.57	0.94	-2.20	-0.26	-1.70	-1.96
41-14	-0.46	-0.11	-2.57	0.94	-2.20	-0.26	-1.70	-1.96
41-13	-0.71	-0.10	-3.63	2.36	-2.09	0.12	-2.05	-1.93
41-16	-0.72	-0.10	-3.63	2.36	-2.09	0.12	-2.05	-1.93
41-96	-0.08	-0.03	-1.82	0.85	-1.08	0.20	-0.95	-0.75
41-129	-0.27	-0.04	-1.73	1.22	-0.82	0.22	-0.88	-0.66
41-124	-0.01	-0.01	-0.82	0.21	-0.62	0.03	-0.43	-0.40
41-126	0.06	-0.00	-0.26	0.00	-0.21	0.06	-0.21	-0.15
41-127	-0.01	-0.00	-0.17	0.00	-0.19	-0.01	-0.13	-0.14
41-94	-0.05	-0.00	-0.12	-0.00	-0.18	-0.03	-0.10	-0.13
41-59	-0.01	-0.00	-0.11	0.00	-0.12	0.02	-0.07	-0.05
41-69	-0.01	-0.00	-0.11	-0.00	-0.12	0.02	-0.07	-0.05
41-56	-0.01	0.00	-0.10	0.00	-0.11	0.01	-0.07	-0.06
41-161	0.01	-0.00	-0.08	0.00	-0.08	0.01	-0.06	-0.05
41-18	-0.01	-0.00	-0.07	-0.00	-0.08	0.02	-0.05	-0.03
41-44	-0.01	-0.00	-0.07	0.00	-0.08	0.02	-0.06	-0.04
41-60	-0.01	-0.00	-0.06	0.00	-0.07	0.02	-0.04	-0.02

hexachlorobenzene

	DFT-SAPT					DFT-D3		
pair	ES	IND	DISP	Exch	E_{INT}	DFT	D3	E_{Tot}
78-75	-5.99	-1.11	-25.52	17.02	-15.60	5.19	-16.64	-11.45
78-81	-5.98	-1.11	-25.51	17.00	-15.60	5.19	-16.64	-11.45
78-66	-1.23	-0.28	-4.87	3.21	-3.17	0.60	-2.61	-2.01
78-90	-1.23	-0.28	-4.87	3.21	-3.17	0.60	-2.61	-2.01
78-22	-1.23	-0.29	-4.45	3.10	-2.88	0.51	-2.38	-1.87
78-26	-1.23	-0.30	-4.44	3.09	-2.88	0.50	-2.38	-1.88
78-39	-1.23	-0.30	-4.45	3.09	-2.88	0.50	-2.38	-1.88
78-43	-1.23	-0.29	-4.44	3.09	-2.87	0.50	-2.38	-1.88
78-69	-1.09	-0.16	-4.75	3.23	-2.76	1.02	-2.93	-1.91

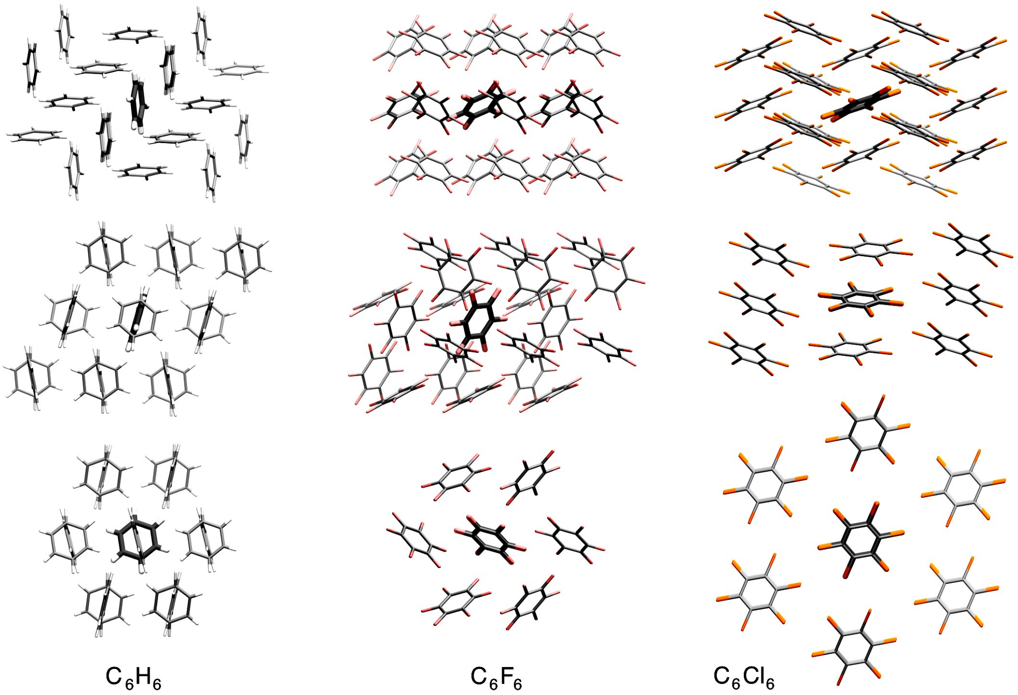
78-87	-1.09	-0.16	-4.75	3.23	-2.76	1.02	-2.93	-1.91
78-23	-1.26	-0.29	-4.44	3.37	-2.62	0.71	-2.30	-1.59
78-27	-1.26	-0.29	-4.44	3.37	-2.62	0.71	-2.30	-1.59
78-38	-1.26	-0.29	-4.44	3.37	-2.62	0.72	-2.30	-1.58
78-42	-1.26	-0.29	-4.44	3.38	-2.62	0.72	-2.30	-1.58

hexabromobenzene

	DFT-SAPT					DFT-D3		
pair	ES	IND	DISP	Exch	E_{INT}	DFT	D3	E_{Tot}
14-11	-8.29	-1.29	-28.85	20.94	-17.48	6.77	-20.90	-14.13
14-17	-8.29	-1.29	-28.85	20.94	-17.48	6.77	-20.90	-14.13
14-26	-2.25	-0.65	-6.65	5.28	-4.28	0.81	-3.78	-2.97
14-2	-2.25	-0.66	-6.65	5.28	-4.28	0.81	-3.78	-2.97
14-49	-2.33	-0.68	-6.12	5.19	-3.93	0.72	-3.37	-2.65
14-53	-2.33	-0.68	-6.12	5.19	-3.93	0.72	-3.37	-2.65
14-70	-2.33	-0.67	-6.12	5.19	-3.93	0.72	-3.38	-2.66
14-66	-2.33	-0.68	-6.05	5.18	-3.87	0.72	-3.38	-2.66
14-50	-2.39	-0.69	-6.02	5.64	-3.46	1.06	-3.13	-2.07
14-54	-2.40	-0.69	-6.03	5.64	-3.47	1.06	-3.13	-2.07
14-65	-2.39	-0.69	-6.02	5.63	-3.47	1.06	-3.13	-2.07
14-69	-2.39	-0.68	-6.02	5.63	-3.46	1.06	-3.13	-2.07
14-5	-1.77	-0.27	-5.79	4.48	-3.34	1.52	-4.09	-2.57
14-23	-1.77	-0.27	-5.79	4.48	-3.34	1.52	-4.09	-2.57

Figure S1

Various views of the benzene (C_6H_6, left column), hexafluorobenzene (C_6F_6, middle column) and hexachlorobenzene (C_6Cl_6, right column) crystal models containing the central molecule (bolder) and the 20 nearest neighbors. The model for hexabromobenzene is identical to the hexachlorobezene.



F

Publication 5 – Explicit σ-hole

JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds

Michal Kolář[†,‡] and Pavel Hobza*[,†,§,∥]

[†]Institute of Organic Chemistry and Biochemistry and Gilead Science Research Center, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, 166 10 Prague 6, The Czech Republic

[‡]Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague 2, The Czech Republic

[§]Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, Olomouc, 771 46 Olomouc, The Czech Republic

[∥]Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Republic of Korea

Supporting Information

ABSTRACT: Until recently, the description of halogen bonding by standard molecular mechanics has been poor, owing to the lack of the so-called σ hole localized at the halogen. This region of positive electrostatic potential located on top of a halogen atom explains the counterintuitive attraction of halogenated compounds interacting with Lewis bases. In molecular mechanics, the σ hole is modeled by a massless point charge attached to the halogen atom and referred to as an explicit σ hole (ESH). Here, we introduce and compare three methods of ESH construction, which differ in the complexity of the input needed. The molecular mechanical dissociation curves of three model complexes containing bromine are compared with accurate CCSD(T)/CBS data. Furthermore, the performance of the Amber force field enhanced by the ESH on geometry characteristics is tested on the casein kinase 2 protein complex with seven brominated inhibitors. It is shown how various schemes depend on the selection of the ESH parameters and to what extent the energies and geometries are reliable. The charge of 0.2*e* placed 1.5 Å from the bromine atomic center is suggested as a universal model for the ESH.

1. INTRODUCTION

The halogen bond, a type of noncovalent interaction between a halogen atom and a Lewis base, has already been extensively studied and reviewed.[1−3] Despite the fact that halogens have higher electronegativity than carbon, which creates a negative partial charge on halogens in organic molecules, halogens favorably interact with a Lewis base atom, such as oxygen or nitrogen with a lone electron pair. This counterintuitive attraction is commonly explained by the existence of a region of positive electrostatic potential (ESP), located on top of the halogen atom.[4] This region, usually referred to as the σ hole, is an inherent feature of compounds containing covalently bound halogens; i.e., it is not induced by the interacting partner in a complex.

The halogen bond motif has been found in various crystalline materials as well as in biomolecular complexes and thus attracts the attention of current science.[5,6] It plays an especially important role in the design of novel drugs, and as many as about 40% of newly introduced drugs contain halogens.[7] It is believed that halogen bonding is at least partially responsible for the high biological action of these drugs. The basic problem which has triggered our interest in this field is that molecular mechanics (MM), which is almost exclusively used in *in silico* drug design, fails to describe halogen bonding (see below). The strength of the halogen bond reaches several kilocalories per mole and increases with the atomic number of the halogen—it is rather weak for chlorine and strongest for iodine. A fluorine atom covalently bound in organic molecules usually does not contain a σ hole, which turns into an inability to create a halogen bond in such systems.[3] When, however, fluorine is bound to a more electronegative atom than carbon, for instance to another fluorine, the σ hole is again formed.[8] Also, inorganic halogen-bonded complexes containing fluorine covalently bound to carbon were recognized.[9] Besides the electrostatic component of the interaction energy, dispersion was also shown to be essential, mainly owing to the close contacts of two atoms with high polarizability (C or N and halogen).[10]

Until recently, the description of halogen bonds with common biomolecular empirical force fields (e.g., the Amber family of force fields)[11,12] has been poor. Our observation of the faulty behavior of the General Amber Force Field (GAFF)[12] in the description of the protein−ligand interaction is certainly not rare.[13] Generally, the nonbonded interaction between a halogen atom and any other atom (as between any two atoms) within the empirical force field is characterized by a partial charge centered on the halogen and two Lennard-Jones (LJ) parameters standing effectively for Pauli repulsion and dispersion attraction. The polarization effects are either included implicitly in the prepolarized charge and LJ parameters[14,15] or explicitly via an additional polarizability

Received: November 22, 2011
Published: March 10, 2012

parameter. The anisotropy of the ESP around the halogen atom (i.e., σ hole), which is of quantum origin, is missing completely.

Very recently, a novel approach was suggested by Ibrahim,[16] who modeled a σ hole explicitly as a massless point charge placed on top of the halogen atom. He applied the explicit σ hole (called "extra point" in ref 16) to calculate the interaction energies and solvation energies as well as to run a short molecular dynamics of a protein−ligand complex. The procedure will be described and discussed below. In the past, the idea of an extra point charge (negative) in the force field was utilized to mimic a lone pair of Lewis bases.[17,18] While the use with halogens appears to be promising, a deeper insight into the construction of the σ hole as well as a revision of the comparison with the benchmark data seem to be needed.

The aim of this study is to provide several schemes of the explicit σ hole (ESH) construction and to compare the ability of the Amber empirical force field with and without the ESH to describe the energetics and geometrical features of halogen bonding. Recently,[19] we applied one of the schemes for an advanced scoring study of aldose reductase inhibitors, one of which contains halogen, with compelling results. Here, we have limited the complexes studied to the brominated ones, and besides the model systems we have also investigated protein−ligand complexes. It should be mentioned that the halogen bond in these systems contributes significantly to their biological action. The treatment of brominated compounds is the least problematic among halogens. The strength of the bromine halogen bond is significantly higher as compared to chlorine,[3,4] and the results are less biased by the eventual relativistic effects than they might be in the iodine case.

2. METHODS

2.1. ESH Construction. We studied three different ways to include the anisotropy of the ESP around the bromine atom described by the GAFF. Because of the electrostatic character of the σ hole, we have not applied any changes to the LJ parameters. We admit that a further reparameterization of LJ parameters might improve the results, but the design of the new halogen parameters does not seem to be as conspicuous as in the case of the σ hole.

In the first scheme, we calculated the molecular mechanical charges by means of the RESP methodology[20] and substituted the bromine point charge placed on the atomic center with two charges—the first representing a σ hole placed at a fixed distance from the bromine atomic center and the second representing a bromine atom. The σ-hole charge and bromine charge were chosen in such a way that the sum of them was equal to the bromine value obtained by the usual RESP fit. In other words, we subtract the ESH charge from the bromine charge. All of the other atomic partial charges were kept intact. In fact, we replaced the point charge of the bromine with a dipole moment, the size of which is parametrically dependent. The first approach, abbreviated here "nF" (as "no fit"), contains two parameters—the charge of the ESH and its distance from the atomic center of bromine. It should be noted that the atomic partial charges of all of the atoms have to be known prior to the construction of the σ hole. On the other hand, no further *ab initio* calculation is needed, which saves computational time significantly.

Contrary to the first approach, where no charges were modified during the construction of the ESH except for the charge of bromine, in the second scheme we chose the charge of the ESH and its distance and adjusted the partial charges of the rest of the molecule employing the RESP methodology. Two parameters had to be attributed to the ESH (i.e., charge and distance) in this approach, called "rF" (abbreviated as "rest fit"). The effect of the ESH is more delocalized across the molecule, and no charges need to be known prior to the construction of the ESH. More likely, the *ab initio* electrostatic potential grid has to be known in order to perform the RESP fit.

The third approach, abbreviated as "aF" ("all fit"), is identical to that introduced by Ibrahim. Fundamentally, only one parameter of ESH is needed here, i.e., the distance of the ESH from the bromine atomic center. The charge of the ESH was calculated by means of RESP. In other words, before the calculation of the partial charges using RESP, an additional fitting position was placed on top of the bromine atom. Undoubtedly, the charge of the ESH was as physically correct as possible (within the validity of the RESP technique) in this case. In ref 16, the ESH was allowed to move on the sphere of bromine providing the bond and angle force constants of Br−ESH and C−Br−ESH, respectively. We rather kept the position of the ESH fixed in order to be consistent with the previous two approaches and to reduce the number of degrees of freedom in the parametrization. Indeed, the origin and significance of the force constants in ref 16 comes from the previous studies of oxygen lone pairs.[17,18] We claim that the ESH should be constructed within the bromime van der Waals diameter. In the case of GAFF, the repulsion LJ parameter σ (not to be confused with the σ hole) is 1.8 Å (giving the r_{min} = 2.02 Å).[12] Since in the original paper, the ESH was placed out of this region, a repulsion LJ parameter was necessary to maintain the numerical stability of the calculation/simulation. We applied the σ LJ parameter of 1.00 Å as provided in ref 16.

The approaches are summarized in Table 1. The complexity of the input data differs across the schemes. While the nF scheme needs partial charges of the atoms (low complexity data), the rF and aF schemes require the *ab initio* grid of the electrostatic potential (highly complex data). Thus, the ratio of the accuracy/computer demands has to be considered as well.

Table 1. A Summary of the ESH Construction Schemes

		nF (no fit)	rF (rest fit)	aF (all fit)
parameters		charge, distance	charge, distance	distance
range	charge	0.05−0.50*e*	0.05 − 0.50*e*	
	distance	0.8−1.6 Å	0.8−1.6 Å	0.8−2.6 Å[a]
step	charge	0.05*e*	0.05*e*	
	distance	0.1 Å	0.1 Å	0.2 Å
charges needed a priori		yes	no	no
ab initio ESP grid needed		no	yes	yes

[a]The Br−ESH distance range was higher in the aF case to make the results comparable with the results from ref 15.

2.2. Gas-Phase Interaction Energies. To clarify the importance of particular ESH parameters, we compared the *ab initio* gas-phase dissociation curves calculated on the CCSD(T)/CBS level with the MM dissociation curves calculated with and without the ESH. As a reference method, we used the CCSD(T)/CBS technique, which, as the only QM method, describes the various motives of noncovalent interactions, including halogen bonding, with chemical (about 1 kcal/mol) or even subchemical (about 0.1 kcal/mol) accuracy.[21]

Table 2. The Location of the Energy Minima of the Dissociation Curves[a]

complex		Br2F_O	Br_O	Br_N
CCSD(T)/CBS	E_{min} [kcal/mol]	−2.43	−2.96	−3.62
	d_{min} [Å]	3.1	3.1	2.9
no ESH	E_{min} [kcal/mol]	−0.70	−0.38	−0.73
	d_{min} [Å]	3.5	3.5	3.5
nF	E_{min} [kcal/mol]	−3.14 (1.6, 0.10)	−2.28 (1.5, 0.10)	−3.83 (1.6, 0.20)
	d_{min} [Å]	3.1 (1.6, 0.10)	3.3 (1.5, 0.10)	3.3 (1.6, 0.20)
rF	E_{min} [kcal/mol]	−2.80 (1.5, 0.15)	−2.33 (1.5, 0.15)	−3.93(1.6, 0.30)
	d_{min} [Å]	3.1 (1.5, 0.15)	3.3(1.5, 0.15)	3.3 (1.6, 0.30)
aF	E_{min} [kcal/mol]	−2.90 (2.2)	−2.48 (2.0)	−2.18 (2.4)
	d_{min} [Å]	3.1 (2.2)	3.1 (2.0)	3.3 (2.4)

[a] Scheme without ESH (abbrev. "no ESH") as well as with ESH are shown for B3LYP charge sets. Where relevant, the ESH parameters are provided in the parentheses. MM values of E_{min} and d_{min} correspond to the lowest mean unsigned absolute error shown in Table 5.

Generally, the reference CCSD(T)/CBS data are accurate and adequately describe the dispersion interaction, unlike the DFT/B3LYP treatment or basis-set-superposition-error (BSSE) uncorrected MP2 treatment used in ref 16 as the benchmark. Since the dispersion energy contributes significantly[10] and sometimes dominantly to the halogen-bond stabilization, its nonadequate treatment can strongly affect the parametrization and/or verification of the ESH. Moreover, Lu and co-workers showed on the set of halogen bonded complexes that B3LYP performs rather poorly when compared with other DFT functionals.[22] The B3LYP average absolute error in interaction energy was as much as 0.86 kcal/mol, which was tens percent of the total interaction energies.[22]

In this study, the complete basis set values were estimated by the extrapolation of the aug-cc-pVDZ and aug-cc-pVTZ BSSE corrected values. The BSSE correction was done employing the counterpoise scheme of Boys and Bernardi.[23] For the CCSD(T) calculation, the Molpro program suite was used.[24] The studied complexes were bromobenzene···acetone (Br_O), bromobenzene···trimethylammonia (Br_N), and 1-bromo-3,5-difluorobenzene···acetone (Br2F_O).

The structures were prepared as follows: We started with an idealized halogen bond of the Br_O complex (C−Br···O angle = 180°, Br···O=C angle = 120°) and performed the full gradient optimization at the MP2/cc-pVTZ level. The final C−Br···O angle was 178.9°, but the same procedure in the Br2F_O case led to a structure significantly distorted by secondary (nonhalogen bonding) interactions. Since our major interest was in the halogen bond itself rather than in the overall interaction between the two molecules, we decided to keep the nearly ideal halogen bond distance and C−Br···O angle of the Br_O complex fixed for the Br2F_O case and optimize the rest. The Br_N complex was fully optimized at the MP2/cc-pVTZ level starting from the 180° C−Br···N angle. From the optimized structures, we generated a series of dissociative geometries by varying the intermolecular distance, while keeping all other geometrical parameters fixed. These geometries, which were not reoptimized, were used for CCSD(T) calculations of dissociation curves. Hence, these curves do not represent the true molecular separation from the minimum but rather a genuine halogen bond dissociation.

The geometric and energetic features of the complexes are summarized in Table 2. The structures with intermolecular distances corresponding to the lowest energy are shown in Figure 1. All of the geometries are available in xyz file format as the Supporting Information. The CCSD(T)/CBS interaction energies are shown in Table S1.

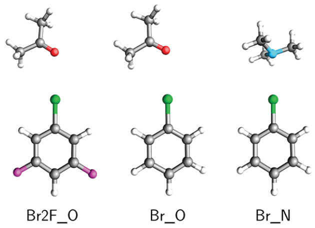


Figure 1. The structures of the complexes investigated in the gas phase. Br2F_O stands for the 1-bromo-3,5-difluorobenzene···acetone complex. Br_O stands for the bromobenzene···acetone complex, and Br_N stands for the bromobenzene···trimethylammonia complex.

For MM calculations, the monomers were optimized at the B3LYP/cc-pVTZ level followed by the calculation of the ESP grid points around the molecule. The grid was constructed with eight layers with a density of three points per unit area (see the Gaussian manual).[25] This option increases the statistical accuracy of the RESP procedure especially in the area around bromine, and as shown below it has a dramatic impact on the quality of the partial charges. The comparison of the MM and CCSD(T) gas-phase energies is not straightforward and might be questioned.[26−28] The reason is that the common biomolecular force fields were originally designed for the condensed phase. For instance, the charges obtained by the RESP fit onto the ESP grid points calculated at the HF/6-31G* level are usually higher in magnitude by about 15% than the real vacuum charges, which should, as proposed by Cornell et al.,[15] intentionally compensate for the missing polarization in the empirical force field.

Recently, a study about the polarization of σ holes was published.[29] On the set of hydrogen bonded complexes, the authors showed that polarization effects might have substantial effect on the extent of the σ hole. We assume that the extent of missing polarization in the force field is of similar magnitude as in the case of σ holes. Thus, to be consistent, the ESP grid points here were calculated at the HF/6-31G* level, as suggested by the developers of the General Amber Force Field, and also at the B3LYP/cc-pVTZ level. Using the DFT method with a significantly larger basis set provides vacuum charges which are not prepolarized and suitably represent the gas-phase electrostatics.[28] For each complex and each ESH variant, two charge sets were thus prepared.

It should be stressed that the MM gas phase calculations were performed to help build up an idea of how the energetics

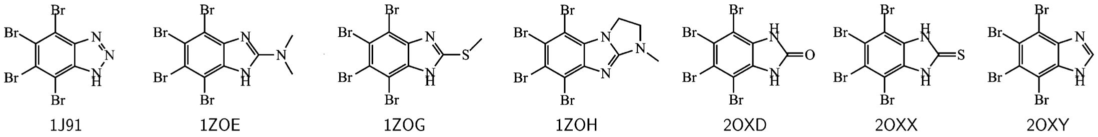


Figure 2. The structures of the tetrabrominated CK2 inhibitors and the corresponding PDB codes.

of halogen bonding is affected by various ESH parameters, while using liquid-phase parameters (i.e., LJ and bonding). Thus, for instance, we abandoned performing gas phase geometry optimizations, which would be difficult to interpret, indeed. Instead, we investigated the role of ESH on molecular geometries in the protein−ligand case (see bellow).

We performed a two-dimensional scan of the parameters in the nF and rF approaches and a one-dimensional scan for the aF approach. For each point (q, d) (charge, distance) or (d) (distance), we calculated the dissociation curve of all of the complexes. The calculations of the MM interaction energies were performed without cutoffs. The Gromacs program suite[29] was used for the MM calculations. The ESH was represented by a so-called virtual site algorithm available in the program as described by Berendsen and van Gunsteren.[31,32] This algorithm keeps the ESH position fixed with respect to the real atoms and redistributes the forces acting on ESH properly.

2.3. Optimization of Protein−Ligand Complexes. The effect of the ESH parameters on the geometry of a protein−ligand complex was tested on the set of casein kinase 2 (CK2) complexes, which we investigated recently.[13] The protein is inhibited by polyhalogenated ligands,[33−35] from which seven tetrabrominated ones were chosen. Their structural formulas are shown in Figure 2. High-quality X-ray structures are available in the Protein Data Bank under the codes 1J91, 1ZOE, 1ZOG, 1ZOH, 2OXD, 2OXX, and 2OXY.[36−38] All of the complexes were energy minimized, and the position of the ligand in the active site was evaluated.

The protein was described using the Amber parm03 force field[11,39] and the ligands using GAFF.[12] We used HF charge sets consistently with the force-field definitions. All of the bromines were enhanced by the ESH using all of the approaches mentioned (i.e., nF, rF, and aF). Such a system was energy minimized in the implicit Generalized Born model until the maximum force was lower than 2.4×10^{-5} kcal/mol/Å. The L-BFGS algorithm together with no cutoffs for interatomic interactions were used. To decrease the complexity of the problem, all of the heavy atoms except for the active site were under the position restraints of 12 kcal/mol/Å^2. The active site was chosen on the basis of visual inspection[40] and is shown in Figure 3. The same setup was used for minimizations without the ESH.

The root-mean-square deviation (RMSD) of the heavy atoms was calculated with respect to the X-ray structure. At first, the backbone coordinates of the minimized structure were aligned onto the X-ray coordinates, followed by the separate calculations of ligand and active-site amino acid RMSDs. The number of oxygen atoms within a 3.5 Å vicinity of the bromine atoms was calculated. The value stands for the approximate number of the halogen bonds between the ligand and protein. This was done for all of the ESH parameters summarized in Table 1.

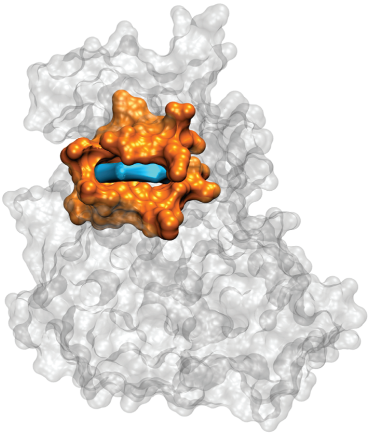


Figure 3. The overall shape of the CK2 protein. The selection of the active-site amino acids (orange) and the ligand (blue) were allowed to move freely during the optimization. The remaining amino acids (gray) were subject to position restraints of 12 kcal/mol/Å^2.

3. RESULTS AND DISCUSSION

3.1. Gas-Phase Interaction Energies. The introduction of the ESH led to the variation of the bromine and other atomic partial charges. The representative values of the charges are shown in Table 3. The aF charges are considered to be more physically sound than those of the nF and rF. The charges of the ESH and the closest atoms vary with the Br−ESH distance, and these variations are similarly pronounced for the rF and aF schemes. In contrast, the nF scheme by definition varies the ESH and Br charges differently, and the other charges are kept intact (see Methods). When the aF ESH charges in bromobenzene and 1-bromo-3,5-difluorobenzene are compared, the latter contains a less positively charged ESH. However, the magnitude of the σ hole should be higher in fluorinated bromobenzene, as discussed in ref 41. The lower ESH charge in the case of 1-bromo-3,5-difluorobenzene is thus quite counterintuitive. It seems that the counterintuitive ESH charges are a consequence of the RESP fitting scheme we used, where we did not employ the default Gaussian program setup for ESP grid calculation (4 layers, density 1 point/Å) but enhanced ESP grid (8 layers, density 3 points/Å). Interestingly, we found that when the default ESP grid is used for RESP fitting, the resulting ESH charges are intuitive (i.e., ESH is more positive for 1-bromo-3,5-difluorobenzene compared with bromobenzene). We investigated a set of halogenated molecules[42] with different ESP grids. We conclude that for denser ESP grids, we obtained a better RESP fit in terms of relative ESP root-mean-square error (not shown), although the dipole moments were not always improved. The unlikely performance of RESP fitting procedure might be attributed to the complicated ESP shape around the halogen atom.

Table 3. The B3LYP Atomic Partial Charges of the Bromobenzene and 1-Bromo-3,5-Difluorobenzene Calculated Using Various Schemes[a]

			bromobenzene		
	d/Å	no ESH	nF	rF	aF
ESH	0.8		0.30	0.30	0.29447
	1.2		0.15	0.15	0.15965
	1.6		0.10	0.10	0.09641
Br	0.8	−0.07145	−0.37145	−0.58701	−0.57727
	1.2	−0.07145	−0.22145	−0.38922	−0.41041
	1.6	−0.07145	−0.17145	−0.32663	−0.31709
C	0.8	−0.13085	−0.13085	0.31959	0.30966
	1.2	−0.13085	−0.13085	0.23242	0.26086
	1.6	−0.13085	−0.13085	0.22171	0.20644
			1-bromo-3,5-difluorobenzene		
	d/Å	no ESH	nF	rF	aF
ESH	0.8		0.20	0.20	0.20462
	1.2		0.10	0.10	0.11694
	1.6		0.10	0.10	0.07350
Br	0.8	−0.0802	−0.21802	−0.36148	−0.36957
	1.2	−0.0802	−0.11802	−0.23042	−0.26715
	1.6	−0.0802	−0.11802	−0.27764	−0.20684
C	0.8	−0.17183	−0.17183	0.13140	0.13948
	1.2	−0.17183	−0.17183	0.07823	0.12575
	1.6	−0.17183	−0.17183	0.21757	0.10058

[a] Only the ESH, bromine and carbon covalently bound to bromine are shown. The values for several Br-ESH distances *d* are shown. Note that in the nF and rF cases, the ESH charge is an arbitrary parameter while in the aF case it is calculated from the *ab initio* data. The charges from the unmodified force field are provided in "no ESH" column. For details about the schemes, see the Methods section.

In Figure 4, the ESP of the σ hole calculated at the B3LYP/cc-pVTZ and various MM schemes is shown. Qualitatively wrong results of the model lacking ESH are apparent (second row plots). No region with positive ESH results from the original force field in contrast with ESH schemes. Indeed, the electrostatic potentials of the benzenes are well behaved in all of the ESH schemes, and Br2F experiences a region which is more positively charged, in agreement with the *ab initio* data.[41] The MM electrostatic potential plots were calculated with the charges derived from the denser ESH grid (i.e., 8 layers, 3 points/Å) calculated at B3LYP/cc-pVTZ. This is, no doubt, an important finding, showing that the simple MM treatment enhanced by ESH is able to describe effectively a complicated induction and the polarization effects of fluorines on bromine.

Table 4 shows the dipole moments for various ESH models and relative root-mean-square errors (RRMS) of the MM ESP with respect to the QM ESP. Standard RESP fit without any ESH yields quite good dipole moments differing by less that 0.2 D. When ESH was introduced by the simplest nF scheme, the dipole moments worsened notably, differing by about 0.4 D. Both "fitting" schemes (i.e., rF and aF) perform much better, providing better dipole moments than the scheme without ESH. The quality of the RESP fit in terms of RRMS is also improved when ESH is included by rF or aF schemes. The simplest nF scheme is not based on ESP generation, hence no RRMS values are provided.

The magnitude of the charges has a direct effect on the dissociation curves. The representative dissociation curves are plotted in Figure 5. For all of the dissociation curves, we have calculated the mean unsigned absolute error (MUAE) and

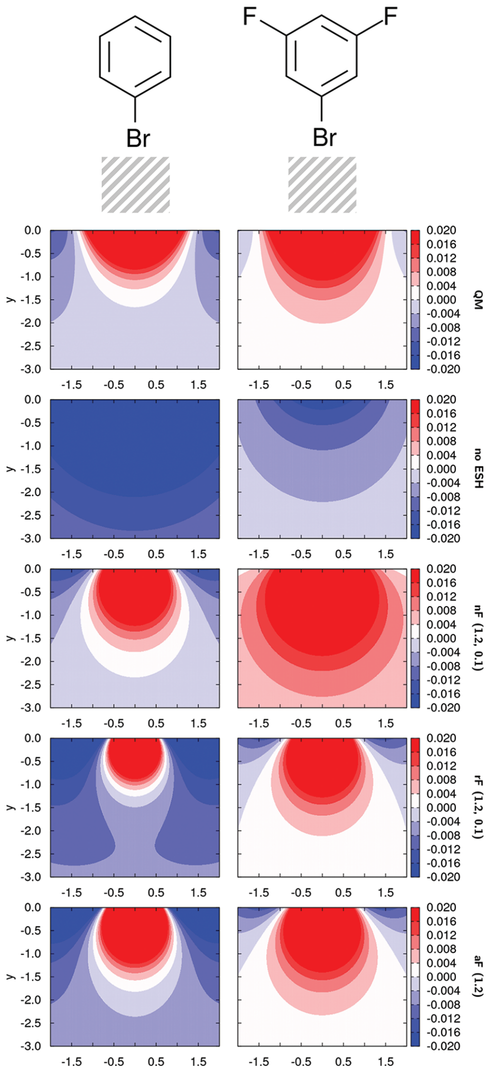


Figure 4. The electrostatic potential maps of the σ hole for bromobenzene (left) and 1-bromo-3,5-difluorobenzene (right). The negative values are in blue, the positive in white and red. The ESP in the hatched areas was calculated at the B3LYP/cc-pVTZ level (abbrev. QM), with the standard MM (abbrev. "no ESH") and with three ESH schemes. The ESH is located at x = 0.0, y = 0.0. The bromine atom is located at x = 0.0, y = 1.2. The charge parameter of the nF and rF schemes was chosen to be 0.10e.

mean unsigned relative error (MURE) according to eqs 1 and 2.

$$\mathrm{MUAE} = \frac{1}{N}\sum_i^N |E_i(\mathrm{MM}) - E_i(\mathrm{QM})| \tag{1}$$

$$\mathrm{MURE} = \frac{1}{N}\sum_i^N \left|\frac{E_i(\mathrm{MM}) - E_i(\mathrm{QM})}{E_i(\mathrm{QM})}\right| \tag{2}$$

Table 4. RESP Fit Characteristics[a]

molecule	bromobenzene	1-bromo-3,5-difluorobenzene
μ [D] (QM)	1.83	0.11
μ [D] (no ESH)	1.99	0.16
μ [D] (nF: 1.2, 0.1)	1.42	0.40
μ [D] (rF: 1.2, 0.1)	1.95	0.13
μ [D] (aF: 1.2)	1.93	0.12
RRMS [%] (no ESH)	18.8	23.0
RRMS [%] (nF: 1.2, 0.1))		
RRMS [%] (rF: 1.2, 0.1)	12.5	14.5
RRMS [%] (aF: 1.2)	10.6	13.9

[a]Dipole moments for various ESH models and relative root mean square errors (RRMS) of the MM ESP with respect to the QM ESP. Only results for B3LYP charge set are shown.

Table 5. The Lowest Mean Unsigned Absolute Error (MUAE) for $R > R_{eq}$[a]

	B3LYP			
	no ESH	nF	rF	aF
Br2F_O	1.17	0.17 (1.6, 0.10)	0.09 (1.5, 0.15)	0.05 (2.2)
Br_O	1.16	0.14 (1.5, 0.10)	0.08 (1.5, 0.15)	0.04 (2.0)
Br_N	2.93	0.83 (1.6, 0.20)	0.73 (1.6, 0.30)	1.27 (2.4)
	HF			
	no ESH	nF	rF	aF
Br2F_O	1.18	0.16 (1.4, 0.10)	0.09 (1.4, 0.15)	0.03 (2.0)
Br_O	1.29	0.13 (1.5, 0.10)	0.06 (1.3, 0.20)	0.10 (2.0)
Br_N	3.13	0.84 (1.6, 0.15)	0.72 (1.6, 0.25)	0.98 (2.4)

[a]The ESH parameters (d, q) for the nF and rF schemes or (d) for the aF scheme are provided in parentheses. The results of the unmodified force field not containing the ESH are shown in the "no ESH" column. The MUAE values are in kcal/mol, the distance in Å, and the charge in e.

Table 6. The Lowest Mean Unsigned Relative Error (MURE) for $R > R_{eq}$[a]

	B3LYP			
	no ESH	nF	rF	aF
Br2F_O	75.1	20.0 (1.6, 0.05)	6.2 (1.4, 0.15)	8.6 (2.2)
Br_O	306.2	31.2 (1.6, 0.05)	13.0 (1.0, 0.30)	42.5 (2.0)
Br_N	106.4	57.5 (1.6, 0.10)	41.9 (1.6, 0.20)	38.0 (2.4)
	HF			
	no ESH	nF	rF	aF
Br2F_O	75.6	19.0 (1.6, 0.05)	4.1 (1.6, 0.10)	4.0 (2.0)
Br_O	417.0	28.1 (1.2, 0.10)	19.1 (0.9, 0.45)	122.2 (2.0)
Br_N	122.2	51.7 (1.6, 0.10)	35.8 (1.6, 0.20)	32.6 (2.4)

[a]The ESH parameters (d, q) for the nF and rF schemes or (d) for the aF scheme are provided in parentheses. The results of the unmodified force field not containing the ESH are shown in the "no ESH" column. The MURE values are in %, the distance in Å, and the charge in e.

where N is the number of points of the dissociation curves and E(MM) and E(QM) are the interaction energies calculated with a force field or on an *ab initio* level, respectively. Both quantities express how well the dissociation curves are represented with respect to the references data. While the MUAE presented in kilocalories per mole shows the absolute difference between the curves, the MURE presented in percentage (%) describes the relative difference. No information about the overestimation or underestimation of the interaction energies is provided. Owing to the enormous errors of the repulsion parts of the curves (not shown), only the points which are farther than the minima of the CCSD(T)/CBS curves were considered. For the B3LYP charge sets, the lowest MUAE and MURE are summarized in Tables 5 and 6. The plots MUAE and MURE as well as the dissociation curves for all of the ESH parameters are provided as Supporting Information.

The B3LYP charge sets are discussed first and the HF charge sets are mentioned below. First, a complete failure of the unmodified force field is apparent. The MUAE reaches about 1 kcal/mol for complexes of acetone and almost 3 kcal/mol for a complex of trimethylammonia, which is comparable with the absolute values of the interaction energies. This fact is well reflected by the high values of the MURE, reaching as much as 300% in the bromobenzene···acetone case. The inclusion of the ESH by any scheme improves the results greatly. The improvement of the acetone complex results is more pronounced than that of trimethylammonia. We claim that this is probably because of the higher electrostatic nature of the interaction in the acetone cases. The electrostatic contribution, originally not covered by the force field well, is corrected for by

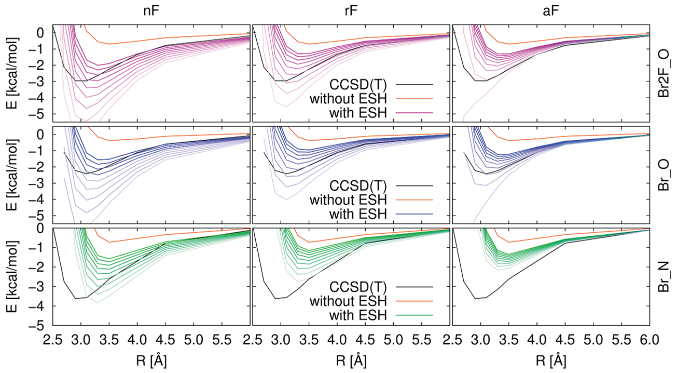


Figure 5. The dependence of the dissociation curves on the Br-ESH distance. The results for a charge of 0.20e in the nF and rF cases are shown. The charge of the aF is calculated exactly (see Methods). The lighter the curve is, the larger the Br—ESH distance used. The dissociation curve calculated with the force field lacking the ESH is plotted in orange; the reference *ab initio* data are in black.

the ESH. Other drawbacks of the force field, such as the unreliable repulsion part, which is largely pronounced in the Br_N case, are not directly connected with the ESH concept and hence cannot really be corrected for by ESH.

Both rF and aF perform better than nF. It should not be surprising, because the atomic charge set in the rF and aF schemes represents the true *ab initio* electrostatic potential better than nF. The nF electrostatic potential is slightly overestimated as shown in Figure 4, third row, as compared with rF and aF (Figure 4, fourth and fifth rows) for both halogenated benzenes. The MUAE of the acetone complexes for the rF and aF schemes is lower than 0.1 kcal/mol and slightly higher (about 0.15 kcal/mol) for the nF scheme. For the trimethylammonia case, the two-parameter models (nF and rF) perform better than one-parameter models (aF), providing MUAEs of 0.83, 0.73, and 1.27, respectively, still much better than the original force field without the ESH. In this context, it should be noted that the additional degree of freedom, i.e., the charge, in the nF and rF cases, unlike with aF, might compensate for the worse repulsion in the Br_N case, yielding better results for nF and rF than aF.

Generally, for two parameter models, the lowest MUAE and MURE were obtained with rather higher Br−ESH distances (above 1.4 Å) and lower ESH charges (below 0.20*e*). However, the lowest MUAE of all was calculated by the one-parameter aF model. The Br−ESH distance of 2.0 Å in this case is questionable owing to the need for an ESH repulsion parameter.

Table 2 shows the optimum distances, absolute interaction energies, and ESH parameters for the curves with the lowest MUAE. The numbers suggest that the repulsion parameter of bromine should be addressed in the future since almost all of the lowest MUAE curves' minima lie in somewhat too high distances.

3.2. Different Charge Sets.

The lowest MUAE and MURE for the HF charge sets are presented in Tables 5 and 6. Due to the larger magnitude of the charges as compared to B3LYP charges, the differences between the various *d* and *q* or *d* parameters were slightly more pronounced. The best MUAE and MURE values are fully comparable with the B3LYP charge sets, and the overall behavior of the parameter dependence is also similar, as shown in Figures S1, S2, and S3.

3.3. Optimization of Protein−Ligand Complexes.

We investigated the effect of the *d* and *q* or *d* parameter selection on the quality of the optimized protein−ligand geometries. The representative root-mean-square deviations (RMSDs) of the ligand with respect to the X-ray structure are depicted in Figure 6 in red, and those of the entire active site are depicted in blue. Only the results for the 1ZOE complex are shown. The RMSD plots for all of the ligands are provided in Supporting Information, Figure S4. The projections of the 3D plots represent the nF and rF schemes (Figure 6, first and second rows); the 2D plot is for the aF approach (Figure 6, third row). In the 3D plots, the darker the color is and the higher the RMSD the optimized structure has, the worse the result it represents. Note the different ranges of the colors. In orange, the number of the protein oxygen atoms which are located closer than 3.5 Å to the ligand bromine atoms is shown (abbreviated as "xbs"). The numbers of "xbs" halogen−oxygen contacts for all of the ligands are provided in the Supporting Information, Figure S6.

Two-dimensional models (nF, rF) provide better results for higher Br−ESH distances and higher ESH charges (the bright areas in Figures 6 and S4). The closer contact of the ESH with protein oxygen atoms tends to stabilize the correct geometry of the protein−ligand complex. The RMSDs of the entire active site are noisier, but a trend similar to ligand RMSDs is apparent. Moreover, the range of the values is narrower as compared to the ligand values. It should be noted that despite the fact that in the nF and rF cases the charges of all four bromine atoms were chosen to be identical, the bromine atoms do not behave identically in the calculations. The vicinity of bromines creates a unique electrostatic potential around each of the bromine atoms, thus assuring the requirement of the distinguishability between them in accordance with chemical intuition.

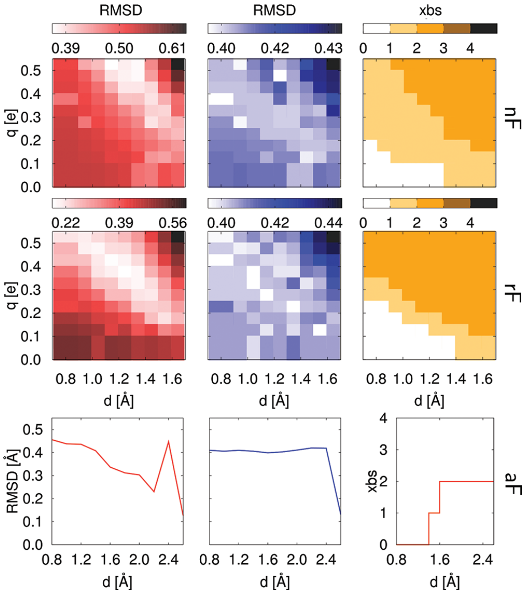


Figure 6. The optimization results of the 1ZOE complex. Two-dimensional scans of the *d* and *q* parameters were performed for the nF and rF schemes (the first and second row); a one-dimensional scan was conducted for the aF scheme (the third row). The RMSDs of the ligand with respect to the active side are depicted in red; the RMSDs of the entire active site are depicted in blue. The numbers of oxygen atoms "xbs" located within 3.5 Å of the bromine atoms are in orange. For comparison, the X-ray structure contains two bromine−oxygen contacts (i.e., xbs = 2), and the force field lacking the ESH provided a structure without any bromine−oxygen contact (i.e., xbs = 0).

The X-ray bromine−oxygen contacts vary between zero (1J91 complex) and three (1ZOH complex). As shown in Figure S6, the regions of the (*d*, *q*) space which provide the correct number of bromine−oxygen contacts are similar to those with lower ligand RMSDs (i.e., larger Br−ESH distance and higher charge). The results of the one-parameter aF scheme are fully comparable with two-parameter schemes.

The summary of the lowest ligand RMSDs is shown in Table 7. We defined a quality "success" as a percentage number of parameter combinations (*d*, *q*) which provided a lower RMSD than the unmodified force field without the ESH. For the aF approach, the quantity is defined in a similar one-dimensional manner. The higher the success is, the less parameter-dependent the scheme appears to be. The success values are provided also in Table 7.

Table 7. The Lowest RMSDs of the Ligands with Respect to the X-Ray Structures[a]

PDB code	no ESH	nF [Å]	rF [Å]	aF [Å]
1J91	1.31	0.57 (90%)	0.34 (100%)	0.56 (100%)
1ZOE	0.45	0.37 (59%)	0.13 (92%)	0.12 (100%)
1ZOG	0.58	0.30 (100%)	0.33 (100%)	0.07 (100%)
1ZOH	0.48	0.49 (0%)	0.49 (0%)	0.26 (10%)
2OXD	0.72	0.37 (100%)	0.27 (71%)	0.03 (100%)
2OXX	0.60	0.23 (91%)	0.43 (96%)	0.06 (90%)
2OXY	1.94	0.22 (100%)	0.17 (96%)	0.06 (100%)

[a] The success values are provided in parentheses. For comparison, also the results of the unmodified force field are shown in the "no ESH" column.

The unmodified force field yielded the lowest ligand RMSD for the 1ZOH case (0.48 Å) but the highest RMSD for 2OXY (1.94 Å). Considering the size of the ligand, all of the values higher than 1.0 Å might be considered as a significant failure. The RMSDs for the nF case are quite similar or better than the unmodified force field, and a high percentage of ESH (d, q) combinations provides an improvement of the ligand geometry within the active site of the protein. An exception is the 1ZOH case, where no (d, q) combination led to a lower RMSD in both the nF and rF schemes. The reasons are 2-fold—first, the unmodified force field itself already provides quite good agreement with the X-ray data, and second, the higher RMSD is mostly caused by the undesirable movement of the five-member ring of the ligand upon minimization. However, the number of bromine−oxygen contacts in the 1ZOH case was indeed improved by adding the ESH into the force field (see Figure S6).

The rF results are very similar to those of nF. The 1ZOH problem is still pronounced, but the success values are slightly better than for the other ligands. Two-parameter schemes are thus quite comparable in spite of the different quality of the partial charges. The one-parameter aF scheme provided much lower RMSDs as compared to the unmodified force field. In four cases, values lower than 0.1 Å were obtained, which is certainly a remarkable result. However, like in the gas phase calculations, the best results were obtained for rather high Br−ESH distances (2.2 Å or more). When one keeps in mind that the common halogen-bond length (i.e., the distance between the halogen and Lewis base) is about 3.2 Å, then the position of the ESH in the best performing aF force-field modifications is only about 1 Å. This might be undesirable for force-field calculations as well as for MD because of the high probability of a collision of the respective atoms. The physical correctness of this charge alignment is still questionable. High success values across the schemes suggest that any force-field enhancement by the ESH is promising and needed.

4. SUMMARY

We have presented and compared three schemes for the description of the anisotropy of the charge distribution of halogenated ligands in molecular mechanics. The molecular mechanical explicit σ hole (ESH) was constructed as a massless point charge. The one-parameter model aF provided excellent results in both the gas-phase calculations and protein−ligand geometry optimization, but only for very high Br−ESH distances (more than 2.0 Å) beyond the bromine vdW radius. The charge of the ESH fitted to the electrostatic-potential grid seems to be somewhat too small to ensure significant improvement when placed within the bromine van der Waals radius. When placed further outside, the ESH performs much better, but probably for wrong reasons. The ESH then is very close (even less than 1 Å) to the electron-donor atom.

Both two-parameter models nF and rF performed slightly worse as compared to the aF scheme. The rF scheme gas-phase interaction energies were better than the nF owing to a more physical description of the electrostatic potentials of the halogenated molecules. The results of the protein−ligand geometry optimizations were very similar. Generally, all of the calculations with the ESH surpass those without ESH.

The practical aspects were considered with the following conclusions: Placing the ESH outside the van der Waals radius might cause numerical instabilities of MD simulations. Thus, a modest overestimation of the ESH charge makes it possible to reach sufficiently short Br−ESH distances. For the adjustment of the other atomic charges, an electrostatic-potential grid based on *ab initio* calculation is needed. This might be a complication for large molecules or for a high number of molecules studied (e.g., drug-design docking/scoring studies). In that case, the nF scheme is an acceptable alternative because the time-consuming ESP grid generation is not necessary.

We showed that ESP is well represented by a low charged ESH placed about 1.2 Å from the bromine mass center. Perhaps, due to the poor repulsion part of the force field, particularly bromine, the best results (in terms of energy as well as protein−ligand geometry) were obtained with ESH at a larger distance form the halogen. According to our results, we suggest a Br−ESH distance of 1.5 Å and an ESH charge of 0.20e as competent parameters for brominated molecules. Similar results might be expected when a higher charge is attached closer to the bromine atomic center, although with a better stability of the simulations. Indeed, these universal parameters might be improved by a careful parametrization targeted to the specific molecules/problems, but we believe that for the majority of the problems solved by molecular mechanics the parameters are sufficiently reliable.

To prove the ESH concept completely, an application to molecular dynamics is also needed. Certainly, it is beyond the scope of this paper, but quantities such as liquid densities or molecular hydration energies have to be addressed.

■ ASSOCIATED CONTENT

S Supporting Information

The geometries of the complexes used for the dissociation-curve calculations, the CCSD(T)/CBS interaction energies, mean unsigned absolute errors (MUAE) and mean unsigned relative errors (MURE) for the B3LYP and HF charge sets, the root-mean-square deviations of the CK2 inhibitors and CK2 active sites, and the numbers of inhibitor bromines and protein oxygen atoms. This material is available free of charge via Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: (+420) 220 410311. E-mail: pavel.hobza@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to Dr. Kevin E. Riley for providing the reference dissociation curves and for valuable discussions and to

Dr. Tomáš Kubař for the discussions of the RESP fitting. This work was supported by the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic [Z40550506], the Czech Science Foundation [P208/12/G016], and Korea Science and Engineering Foundation [World Class Univ. program: R32-2008-000-10180-0]. This work was also supported by the Operational Program Research and Development for Innovations—European Science Fund (CZ.1.05/2.1.00/03.0058). The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged.

■ REFERENCES

(1) Lommerse, J. P. M; Stone, A. J.; Taylor, R.; Allen, F. H. *J. Am. Chem. Soc.* **1996**, *118*, 3108−3116.

(2) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16789−16794.

(3) Politzer, P.; Lane, P.; Concha, M.; Ma, Y.; Murray, J. *J. Mol. Model.* **2007**, *13*, 305−311.

(4) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2007**, *13*, 291−296.

(5) Metrangolo, P.; Neukirch, H.; Pilati, T.; Resnati, G. *Acc. Chem. Res.* **2005**, *38*, 386−395.

(6) Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. *J. Med. Chem.* **2009**, *52*, 2854−2862.

(7) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. *Chem. Soc. Rev.* **2011**, *40*, 2267−2278.

(8) Munusamy, E.; Sedlák, R.; Hobza, P. *ChemPhysChem* **2011**, *12*, 3253−3261.

(9) Lu, Y. X.; Zou, J. W.; Yu, Q. S.; Jiang, Y. J.; Zhao, W. N. *Chem. Phys. Lett.* **2007**, *449*, 6−10.

(10) Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 232−242.

(11) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(12) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(13) Dobeš., P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P. *J. Phys. Chem. B* **2011**, *115*, 8581−8589.

(14) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(15) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620−9631.

(16) Ibrahim, M. A. A. *J. Comput. Chem.* **2011**, *32*, 2564−2574.

(17) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632−1646.

(18) Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048−1057.

(19) Fanfrlík, J.; Kolář, M.; Lepšík, M; Musuramy, E.; Řezáč, J.; Hobza, P. In preparation.

(20) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(21) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023−5063.

(22) Lu, Y. X.; Zou, J. W.; Fan, J. C.; Zhao, W. N.; Jiang, Y. J.; Yu, Q. S. *J. Comput. Chem.* **2009**, *30*, 725−732.

(23) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.

(24) Werner, H. J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; *MOLPRO*, version 2010.1; Cardiff University: Cardiff, Wales; Universität Stuttgart: Stuttgart, Germany, 2010.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.1; Gaussian, Inc.: Wallingford, CT, 2009.

(26) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944−955.

(27) Kolář, M.; Berka, K.; Jurečka, P.; Hobza, P. *ChemPhysChem* **2010**, *11*, 2399−2408.

(28) Zgarbová, M.; Otyepka, M.; Šponer, J.; Hobza, P.; Jurečka, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10476−10493.

(29) Hennemann, M.; Murray, J. S.; Politzer, P.; Riley, K. E.; Clark, T. *J. Mol. Model.* **2011**, DOI: 10.1007/s00894-011-1263-5.

(30) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(31) Berendsen, H. J. C.; van Gunsteren, W. F. Molecular dynamics simulations: Techniques and approaches. In *Molecular Liquids-Dynamics and Interactions*; Barnej, A. J., Orville-Thomas, W. J., Yarwood, J., Eds.; D. Reidel: Dordrecht, The Netherlands, 1984; NATO ASI C 135, pp 475−500.

(32) van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C. Interaction function and force field. *Gromacs User Manual*; version 4.5.4; University of Uppsala: Uppsala, Sweden, 2010.

(33) Pagano, M. A.; Bain, J.; Kazimierczuk, Z.; Sarno, S.; Ruzzene, M.; Di Maira, G.; Elliott, M.; Orzeszko, A.; Cozza, G.; Meggio, F.; Pinna, L. A. *Biochem. J.* **2008**, *415*, 353−365.

(34) Gianoncelli, A.; Cozza, G.; Orzeszko, A.; Meggio, F.; Kazimierczuk, Z.; Pinna, L. A. *Bioorg. Med. Chem.* **2009**, *17*, 7281−7289.

(35) Cozza, G.; Bortolato, A.; Moro, S. *Med. Res. Rev.* **2010**, *30*, 419−462.

(36) De Moliner, E.; Brown, N. R.; Johnson, L. N. *Eur. J. Biochem.* **2003**, *270*, 3174−3181.

(37) Battistutta, R.; Mazzorana, M.; Sarno, S.; Kazimierczuk, Z.; Zanotti, G.; Pinna, L. A. *Chem. Biol.* **2005**, *12*, 1211−1219.

(38) Battistutta, R.; Mazzorana, M.; Cendron, L.; Bortolato, A.; Sarno, S.; Kazimierczuk, Z.; Zanotti, G.; Moro, S.; Pinna, L. A. *ChemBioChem* **2007**, *8*, 1804−1809.

(39) Sorin, E.; Pande, V. S. *Biophys. J.* **2005**, *88*, 2472−2493.

(40) Contains the following amino acids: Val45, Gly46, Arg47, Ser51, Glu52, Val53, Ile66, Ile67, Lys68, Glu81, Val95, Lys96, Phe113, Glu114, Tyr115, Val116, Asn118, His160, Asn161, Met163, Arg172, Ile174, Asp175, and Gly177.

(41) Riley, K.; Murray, J.; Fanfrlík, J.; Řezáč, J.; Solá, R.; Concha, M.; Ramos, F.; Politzer, P. *J. Mol. Model.* **2011**, *17*, 3309−3318.

(42) Bromobenzene, chlorobenzene, 1-bromo-3,5-difluorobenzene, bromoethane, chloroethane, bromomethane, chloromethane, 3-bromo-pro-1-en, 3-chloro-prop-2-en, 4-bromotoluene, and 2-chlorotoluene.

On extension of the current biomolecular empirical force field for the description of halogen bonds

Michal Kolář[1,2] and Pavel Hobza[1,3,4]*

[1]Institute of Organic Chemistry and Biochemistry and Gilead Science Research Center, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, 166 10 Prague 6, The Czech Republic, tel.: (+420) 220 410311, email: *pavel.hobza@uochb.cas.cz*

[2]Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 128 43 Prague 2

[3]Regional Center of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, Olomouc, 771 46 Olomouc, The Czech Republic

[4]Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Republic of Korea

Supplementary Information

Table S1: CCSD(T)/CBS interaction energies E_{int} of the complexes.

Filename	Br...Y distance [Å]	E_{int} [kcal/mol]
br2f_o_25.xyz	2.5	0.45
br2f_o_27.xyz	2.7	–2.13
br2f_o_29.xyz	2.9	–2.95
br2f_o_31.xyz	3.1	–2.96
br2f_o_33.xyz	3.3	–2.65
br2f_o_35.xyz	3.5	–2.24
br2f_o_40.xyz	4.0	–1.34
br2f_o_45.xyz	4.5	–0.78
br2f_o_60.xyz	6.0	–0.19
br2f_o_70.xyz	7.0	–0.09
br_o_27.xyz	2.7	–1.10
br_o_29.xyz	2.9	–2.22
br_o_31.xyz	3.1	–2.42
br_o_33.xyz	3.3	–2.23
br_o_35.xyz	3.5	–1.91
br_o_40.xyz	4.0	–1.10
br_o_45.xyz	4.5	–0.59
br_o_60.xyz	6.0	–0.07
br_o_70.xyz	7.0	0.00
br_n_25.xyz	2.5	0.11
br_n_27.xyz	2.7	–2.74
br_n_29.xyz	2.9	–3.62
br_n_31.xyz	3.1	–3.57
br_n_33.xyz	3.3	–3.15
br_n_35.xyz	3.5	–2.62
br_n_45.xyz	4.5	–0.78
br_n_60.xyz	6.0	–0.12

Figure S1: Error analysis of the gas phase dissociation curves of 1-bromo-3,5-difluorobenzene...acetone complex (Br2F_O). Mean unsigned absolute errors (MUAE) are plotted in green and mean unsigned relative errors (MURE) are in grey. Two charge sets are compared: RESP charges based on B3LYP/cc-ptvz electrostatic potential grid and on HF/6-31G* electrostatic potential grid. Three explicit sigma-hole construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Note the different color ranges of the plots.

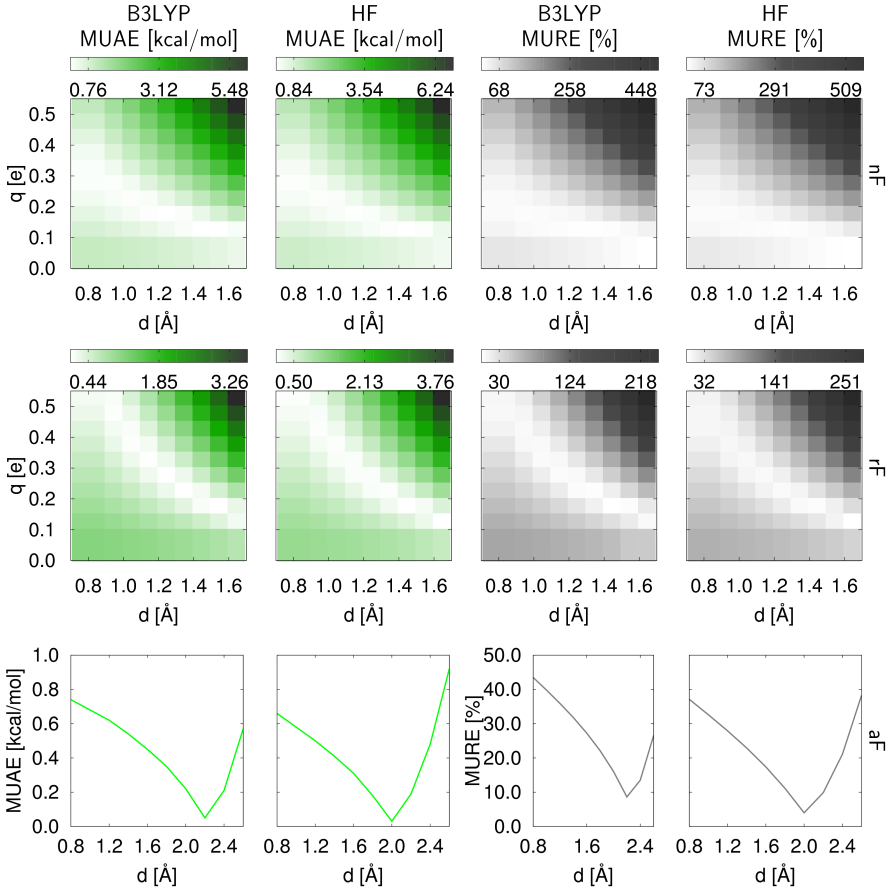


Figure S2: Error analysis of the gas phase dissociation curves of bromobenzene...acetone complex (Br_O). Mean unsigned absolute errors (MUAE) are plotted in green and mean unsigned relative errors (MURE) are in grey. Two charge sets are compared: RESP charges based on B3LYP/cc-ptvz electrostatic potential grid and on HF/6-31G* electrostatic potential grid. Three explicit sigma-hole construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Note the different color ranges of the plots.

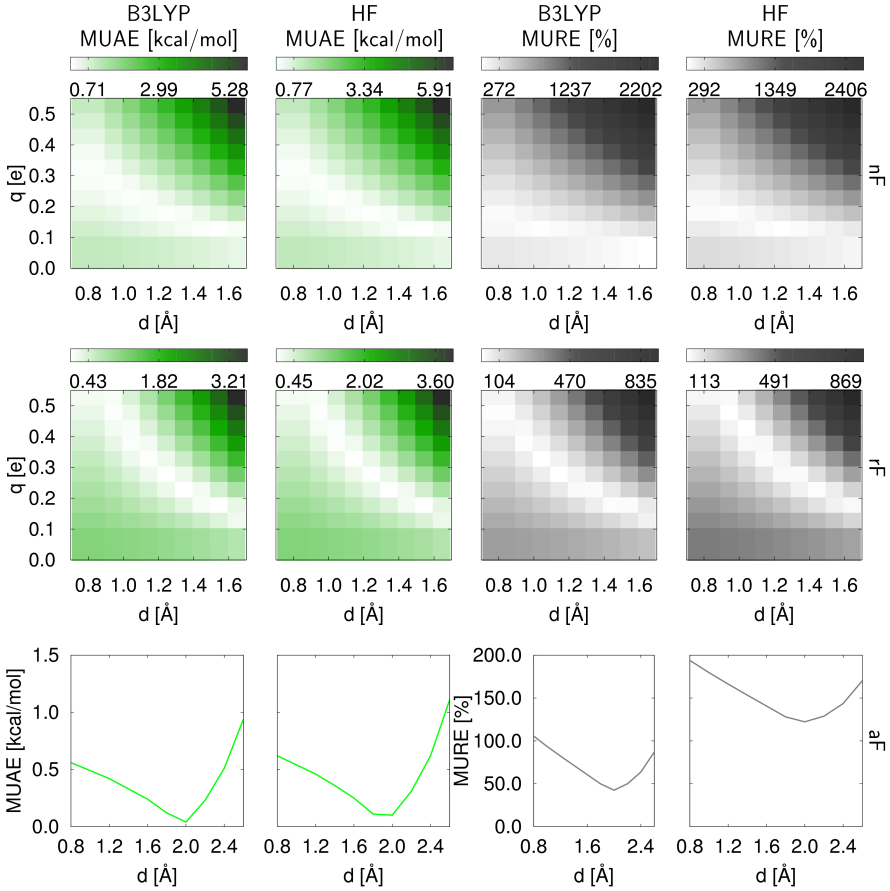


Figure S3: Error analysis of the gas phase dissociation curves of bromobenzene...trimethylammonia complex (Br_N). Mean unsigned absolute errors (MUAE) are plotted in green and mean unsigned relative errors (MURE) are in grey. Two charge sets are compared: RESP charges based on B3LYP/cc-ptvz electrostatic potential grid and on HF/6-31G* electrostatic potential grid. Three explicit sigma-hole construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Note the different color ranges of the plots.

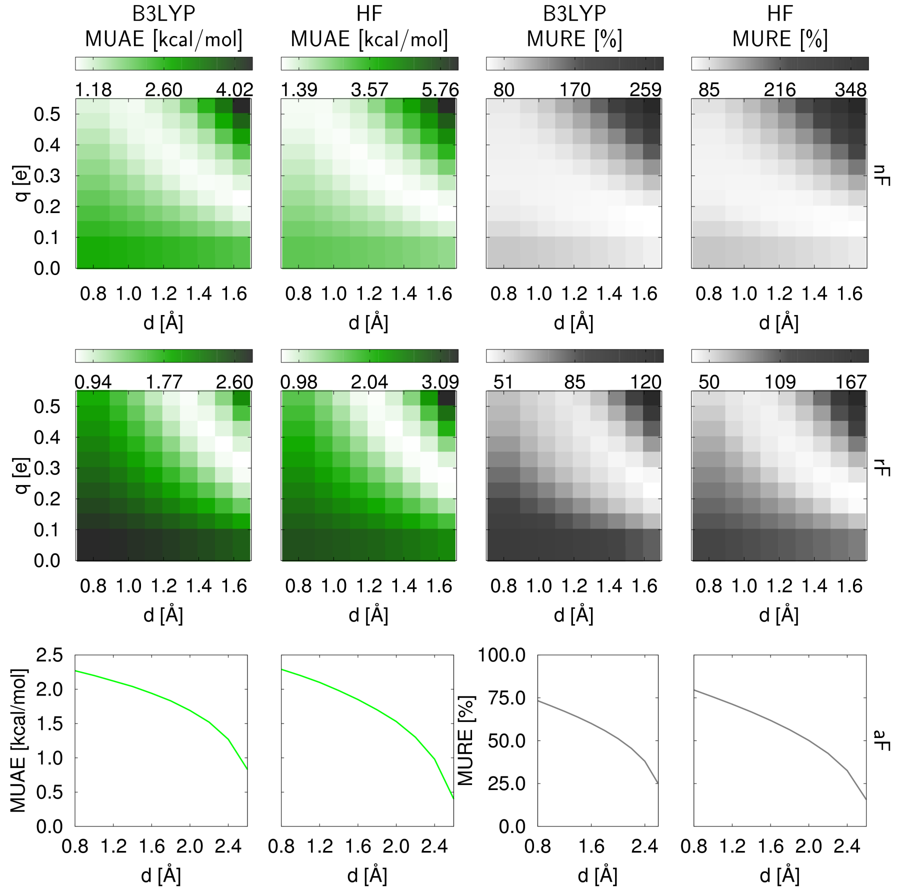


Figure S4: Root mean square deviations (RMSD) with respect to the X-ray geometry. PDB codes were calculated for seven casein kinase 2 inhibitors. Three explicit sigma-hole construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Note the different color ranges of the plots.

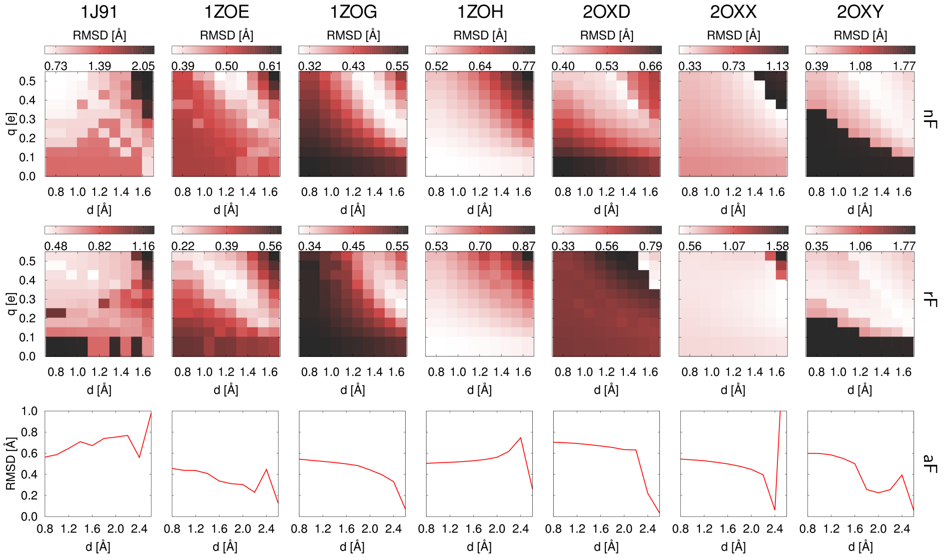


Figure S5: Root mean square deviations (RMSD) with respect to the X-ray geometry. PDB codes were calculated for the active sites of seven casein kinase 2 complexes. Three explicit sigma-hole construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Note the different color ranges of the plots.

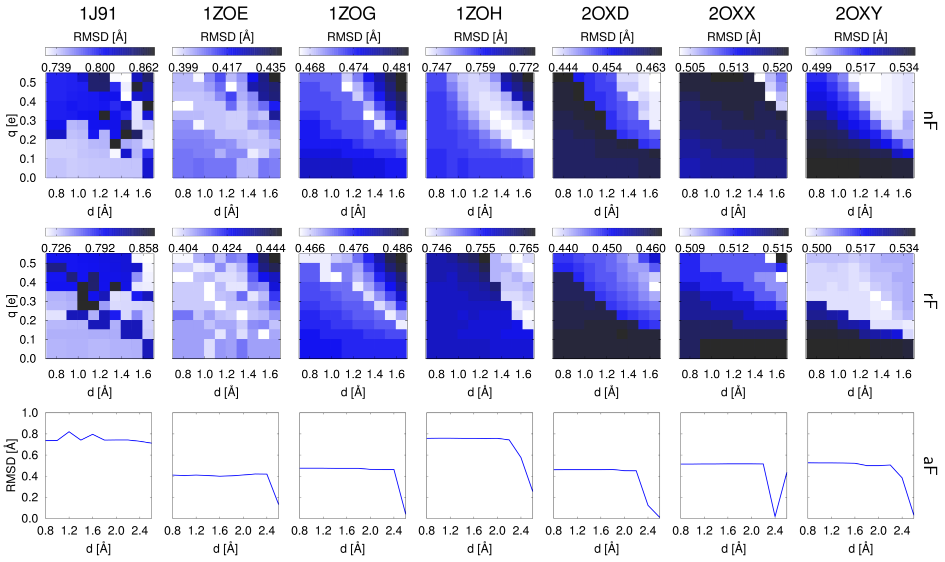
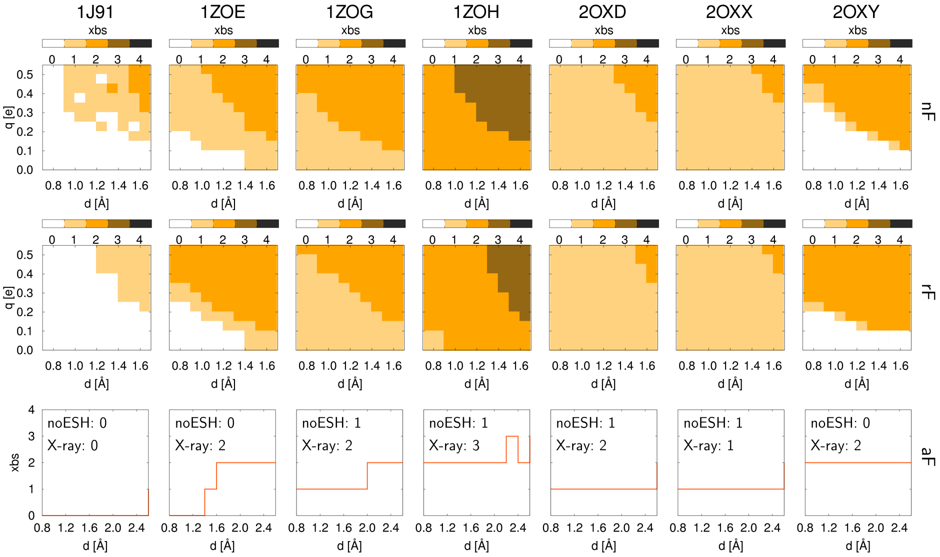


Figure S6: Number of contacts between inhibitor bromines and protein oxygen atoms (xbs). The results for three explicit sigma-hole (ESH) construction schemes are shown: two-parameter models nF and rF, and one-parameter model aF. Fro comparison, the number of bromine-oxygen contacts in experimental X-ray structure (X-ray) and force field without ESH (noESH) are provided as the insets of the last row plots.



G

Publication 6 – Docking with Explicit σ-hole

Chemical Communications, vol. 49, issue 10, pp. 981–983, available online at `http://dx.doi.org/10.1039/C2CC37584B` – Reproduced by permission of The Royal Society of Chemistry.

ChemComm

RSCPublishing

COMMUNICATION

View Article Online
View Journal | View Issue

Plugging the explicit σ-holes in molecular docking†

Michal Kolář,[ab] Pavel Hobza[ac] and Agnieszka K. Bronowska*[de]

Cite this: *Chem. Commun.*, 2013, **49**, 981

Received 18th October 2012, Accepted 12th December 2012

DOI: 10.1039/c2cc37584b

www.rsc.org/chemcomm

A novel approach in molecular docking was successfully used to reproduce protein–ligand experimental geometries. When dealing with halogenated compounds the correct description of halogen bonds between the ligand and the protein is shown to be essential. Applying a simple molecular mechanistic model for halogen bonds improved the protein–ligand geometries as well as halogen bond features, which makes it a promising tool for future computer-aided drug development.

A promising tool for computer-aided molecular design is presented. By means of molecular docking we calculated the binding poses of a series of 92 halogenated inhibitors and successfully reproduced the crystallographic data in 90 out of 92 instances. For the first time we incorporate into a docking program suite a molecular-mechanical approach that correctly describes halogen bonding. The approach is based on a massless positive point charge included in addition to the halogen atoms, which mimics the anisotropy of the charge density around the halogen atom, known as the σ-hole. We show that this description of halogen bonding considerably improves the reliability of the protein–ligand geometries determined by a docking process, especially in those cases where more than one halogen bond is established between the ligand and the active site of the protein.

The cost of a drug being developed by a major pharmaceutical company is at least $4 billion, and it can be as much as $11 billion.[1] The time required for the drug development may vary, but typically it takes from 7 to 12 years.[2] Considering that a major reason for drug failing is lack of efficacy,[3] new methods for describing the drug–target binding are actively being sought.

Many drugs available on the market and new bioactive chemical entities are halogenated compounds. The halogen atoms are introduced to increase the membrane permeability hence improving oral absorption, to fill hydrophobic cavities in the protein binding site, to facilitate the blood–brain barrier crossing, and to prolong the lifetime of the drug.[3] Apart from those non-specific effects, halogens were recognised as being able to participate in a highly specific, directional, non-covalent interaction, known as halogen bond.[4,5] According to the most recent ''provisional recommendation'' by IUPAC, it is an attractive interaction occurring between an electrophilic region of a halogen atom and a nucleophilic region of another atom or a molecular fragment such as a carbonyl oxygen. The strength of the interaction increases with the atomic number of the halogen reaching several kcal mol^{-1}.[6] Typical binding geometry is depicted in Fig. 1a. The nature of the attraction lies, in large part, in a so-called σ-hole.[7–9] Quantum chemical calculations revealed that the charge distribution around the halogen atom is highly anisotropic, creating a region with positive electrostatic potential located on top of the halogen atom.

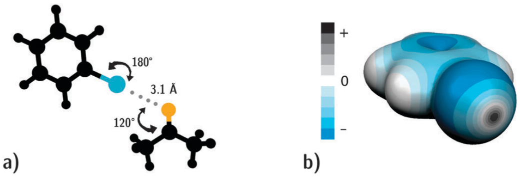


Fig. 1 (a) Geometrical features of a typical halogen bond between bromobenzene and acetone. (b) The charge distribution around the bromobenzene molecule. The regions of negative electrostatic potential are in blue, positive regions in grey. The grey disc in the forefront is called σ-hole.

[a] *Institute of Organic Chemistry and Biochemistry and Gilead Science Research Center, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, 16610 Prague 6, Czech Republic*
[b] *Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 12843 Prague 2, Czech Republic*
[c] *Department of Physical Chemistry, Palacký University, Olomouc, 77146 Olomouc, Czech Republic*
[d] *Faculty of Chemistry, University of Heidelberg, Im Neuenheimer Feld, D-69115 Heidelberg, Germany*
[e] *Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany. E-mail: agnieszka.bronowska@h-its.org; Fax: +49 6221 533298; Tel: +49 6221 533510*

† Electronic supplementary information (ESI) available: The list of PDB codes of protein–ligand complexes studied, the charge fitting and ESH fitting procedures, and the description of the molecular docking protocol. See DOI: 10.1039/c2cc37584b

This positive region, the σ-hole, attracts the negative lone-pair of the Lewis base (Fig. 1b). This poses a serious challenge to current modelling approaches, which treat halogen atoms as having all-negative electrostatic potential, thus failing to correctly describe the halogen-bonded systems, such as protein–ligand complexes. It should be emphasized, though, that the role of halogen-bonding in tuning the intermolecular interactions is not limited to medicinal and pharmaceutical chemistry. There is a growing recognition of this type of interactions among inorganic and supramolecular chemists in the applications of halogen-bonding in liquid crystals, light-induced surface patterning of supramolecular polymers and crystal engineering.[10,11]

Halogen bonding is described well at Hartree–Fock or DFT levels of theory, providing at least a double-zeta basis set is used. Semi-empirical methods fail to describe halogen bonds as well as standard molecular mechanics. In the case of computer-aided drug design the use of computationally cheap methods is inevitable. Since a correct description of halogen bonding is of such a fundamental importance, molecular mechanical approaches correctly describing σ-holes were introduced by several laboratories.[12–15] The essential component of all these approaches was a positively charged, optionally massless, dummy-atom, representing the σ-hole. However, all these corrections were applied to the molecular-mechanical force fields, which require at least preliminary structural data and which are designed to study the dynamic behaviour of systems of known structures. So far, no improvements have been implemented in the suites, which are designed to predict the structure of macromolecular complexes.

Herein we applied such a concept for the first time to the molecular docking scheme. The entity, denoted explicit σ-hole (ESH), was used in conjunction with the UCSF DOCK molecular docking suite.[16] The performance of the improved docking has been tested on 92 protein–ligand complexes for which the crystallographic data are available. The geometries calculated with and without the ESH concept were compared with the experimental geometries (Fig. 2). It should be emphasised that the faithful geometrical representation of the protein–ligand complexes is the essential prerequisite for any further computational investigation not only in drug design studies.

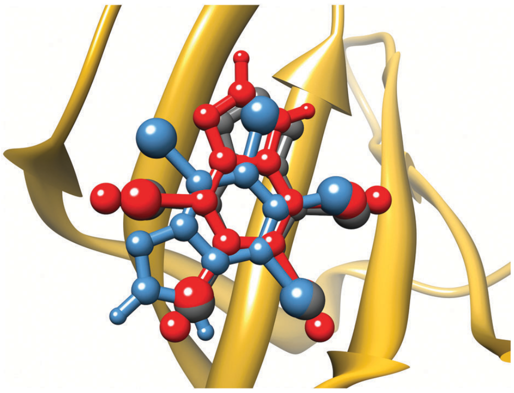


Fig. 2 The overlay of the predicted binding poses of the K17 inhibitor of casein kinase 2 (PDB code 2OXY) with (red) and without (blue) explicit σ-holes (ESH) and comparison with the crystal structure (grey).

Four pharmaceutically attractive protein targets were chosen, namely aldose reductase (ALDR), cyclin-dependent kinase 2 (CDK2), casein kinase 2 (CK2), and human immunodeficiency virus 1 reverse transcriptase (HIVRT), since they are known to be effectively inhibited by halogenated ligands. From the Protein Data Bank[17] a set of protein–ligand X-ray geometries was collected (see ESI†) and their analysis revealed the following facts: the set contained 55 chlorinated, 38 brominated and one iodinated ligand and about 55% of ligands contained more than one possible halogen-bond donor (*i.e.* Cl, Br or I). The set comprised both halogen bond complexes (about 57%) as well as the complexes without any significant halogen–Lewis base contact (43%). In some instances, mostly in the CK2 case, two or three halogen bonds were identified. About 85% of halogen bonds were established with protein backbone carbonyl oxygens. No nitrogen was involved in halogen bonds which reflects the low abundance of nitrogen acceptors in the protein structures contrary to *e.g.* advanced crystalline materials.[11] All the ligands were subject to the docking procedure, which is described in detail in the ESI† section. The outcome of the docking was a set of geometries of the protein–ligand complexes. For each ligand 25 highest-ranked geometries were analysed. Although the ranking is based on electrostatic and van der Waals interactions, and therefore quite simplistic, it can consistently filter out non-physical ligand orientations. The root-mean-square deviation (RMSD) of the heavy atoms of the ligand was calculated with respect to the X-ray geometry.

The RMSDs are summarised in Table 1. The lowest RMSD, the highest RMSD and the average RMSD over all ligands and all their docked orientations were calculated. By visual inspection also the correct binding poses (*i.e.* "native orientations") were distinguished. Evidently, all the RMSD descriptors are improved by inclusion of ESH. In other words, the geometries predicted by including ESH outperformed those without ESH. The most striking distinctions appear in the case of CK2, where more than one halogen bond contributes to the binding arrangement.

Table 1 Root mean square deviations (RMSDs) of heavy atoms of the ligand calculated with respect to the X-ray experimental geometries. Natives stands for the number of correctly identified binding poses

	Lowest RMSD [Å]	Highest RMSD [Å]	Average RMSD [Å]	Natives detected
ALDR				
No ESH	0.11	3.82	1.74	7/7
ESH	0.08	3.67	1.21	7/7
CDK2				
No ESH	0.41	11.46	6.25	26/32
ESH	0.32	8.89	4.16	32/32
CK2				
No ESH	0.83	10.17	5.76	11/16
ESH	0.09	5.22	3.21	16/16
HIVRT				
No ESH	0.45	15.86	7.63	29/37
ESH	0.17	9.52	3.59	35/37

Table 2 The lengths of halogen–acceptor contacts averaged over all protein–ligand complexes. Items in the first column, *e.g.* V47(O), stand for the average distance between valine 47 oxygen (acceptor atom) and the closest ligand halogen. All distances are in Å

	No ESH	ESH	X-ray
ALDR			
V47(O)	3.53	3.38	3.04
T113(OG1)	3.31	3.32	3.32
CDK2			
I10(O)	4.17	3.67	4.15
E12(N)	5.59	3.58	3.56
F80(ring COM)	3.91	3.85	3.43
L83(O)	7.96	3.01	2.9
Q131(O)	4.93	3.64	3.62
D145(O)	7.55	3.61	4.06
CK2			
E108(O)	3.46	3.39	3.41
V110(O)	3.36	3.18	2.97
HIVRT			
K101(O)	6.99	3.88	4.06
Y188(O)	4.51	4.09	3.18
F227(ring COM)	6.39	4.76	4.40
L234(O)	5.48	4.00	4.12

The analysis of the lengths of the halogen bonds in protein–ligand complexes is presented in Table 2. The effect of ESH is emphasised by the halogen–acceptor distances, where ESH typically provides shorter halogen–acceptor contacts than those predicted in the absence of ESH, the shorter distances agreeing better with the experimental geometries. Also the number of halogen bonds established between the pose and protein is affected by the ESH presence. In CK2 complexes, which contain more than one halogen bond, the docking with ESH was able to reproduce 7 of 10 complexes with the correct halogen bonds pattern (*i.e.* all amino acids involved agreed with the X-ray data) compared to 3 of 10 without ESH.

To summarise, we performed a docking study of halogenated enzyme inhibitors. By including a molecular mechanistic model for σ-hole description we obtained generally better protein–ligand geometries than those accessed so far. It has to be noted that due to the simplicity of the ESH model, the improvement was reached without significant additional computational cost which makes it promising for all future docking studies involving halogenated compounds.

This work was supported by the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic [RVO: 61388963], and the Czech Science Foundation [P208/12/G016]. This work was also supported by the Operational Program Research and Development for Innovations – European Science Fund (CZ.1.05/2.1.00/03.0058). The support of *Praemium Academiae* of the Academy of Sciences of the Czech Republic awarded to P. H. in 2007 is acknowledged. Gratitude is also expressed towards Interdisciplinary Centre for Scientific Computing (IWR) and HGS MathComp Graduate School, University of Heidelberg, Germany, for support of this study.

Notes and references

1 B. H. Munos, InnoThink Center For Research In Biomedical Innovation, *Thomson Reuters Fundamentals via FactSet Research Systems*, 2012.
2 A. Merino, A. K. Bronowska, D. B. Jackson and D. J. Cahill, *Drug Discovery Today*, 2010, **15**, 749.
3 M. Z. Hernandez, S. M. Cavalcanti, D. R. Moreira, W. F. de Azevedo Jr and A. C. Leite, *Curr. Drug Targets*, 2010, **11**, 303.
4 Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang and W. Zhu, *J. Med. Chem.*, 1999, **52**, 2854.
5 P. Auffinger, F. A. Hays, E. Westhof and P. S. Ho, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 16489.
6 P. Politzer, K. E. Riley, F. A. Bulat and J. S. Murray, *Comput. Theor. Chem.*, 2012, **998**, 2.
7 P. Politzer, P. Lane, M. Concha, Y. Ma and J. Murray, *J. Mol. Model.*, 2007, **13**, 305.
8 T. Clark, M. Hennemann, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2007, **13**, 291.
9 P. Politzer, J. S. Murray and M. Concha, *J. Mol. Model.*, 2008, **14**, 659.
10 E. Corradi, S. V. Meille, M. T. Messina, P. Metrangolo and G. Resnati, *Angew. Chem., Int. Ed.*, 2000, **39**, 1782.
11 P. Metrangolo, H. Neukirch, T. Pilati and G. Resnati, *Acc. Chem. Res.*, 2005, **38**, 386.
12 M. A. A. Ibrahim, *J. Comput. Chem.*, 2011, **32**, 2564.
13 S. Pieraccini, A. Forni and M. Sironi, *Phys. Chem. Chem. Phys.*, 2011, **13**, 19508.
14 M. Kolář and P. Hobza, *J. Chem. Theory Comput.*, 2012, **8**, 1325.
15 W. L. Jorgensen and P. Schyman, *J. Chem. Theory Comput.*, 2012, **8**(10), 3895.
16 P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James and I. D. Kuntz, *RNA*, 2009, **15**, 1219.
17 www.pdb.org.

Electronic Supplementary Information (ESI) for Chemical Communications

Plugging the explicit σ-holes in molecular docking

Michal Kolář,[a,b] Pavel Hobza[a,c] and Agnieszka K. Bronowska*[d,e]

[a] Institute of Organic Chemistry and Biochemistry and Gilead Science Research Center, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, 16610 Prague 6, Czech Republic.
[b] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Albertov 6, 12843 Prague 2, Czech Republic.
[c] Department of Physical Chemistry, Palacký University, Olomouc 77146 Olomouc, Czech Republic.
[d] Faculty of Chemistry, University of Heidelberg, Im Neuenheimer Feld, D-69115 Heidelberg, Germany.
[e] Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany. Fax: +49 6221 533298; Tel: +49 6221 533510; E-mail: agnieszka.bronowska@h-its.org

Materials and Methods

In this work, the following protein-ligand complexes were studied: 1IEI, 1US0, 1Z89, 1Z8A, 2IKI, 2IKJ, 2PFH, 2R3F, 1H07, 2VU3, 2R3K, 1WCC, 1PXI, 1FVT, 1H1R, 1H08, 1H01, 2R3J , 2J9M, 2V22, 2R3Q, 2I40, 2C68, 1P5E, 3MY5, 2VTJ, 2R3R, 2R3P, 1Y8Y, 1YKR, 3UNK, 2R3L, 2IW6, 2C69, 3LFS, 2VTR, 2B54, 2BHE, 3LE6, 3KXH, 2OXX, 1ZOH, 3PVG, 3KXG, 1ZOG, 3KXN, 1J91, 2PVK, 2OXY, 3KXM, 2QC6, 2OXD, 1ZOE, 3NGA, 3RPS, 1HNV, 1HNI, 3DYA, 3DLE, 3C6U, 2VG6, 1TKZ, 3MEC, 2VG5, 1FK9, 1TL1, 1RT5, 3C6T, 1VRU, 1RT6, 3I0R, 3FFI, 3DI6, 2RKI, 1TL3, 1RT7, 3E01, 2RF2, 1EP4, 3T19, 3DRP, 2VG7, 1TKT, 3I0S, 3DLG, 1DTT, 3R8D, 3QIN, 3HYF, 1JLG, 2YKM, 1IKX.

The charges of the ligands were assigned by the UCSF Chimera program suite[1] at the AM1-BCC level in a standard manner.[2] Then, the ESH was constructed as the nF model[3] as described in Kolář and Hobza:[4] the dummy atom with a desired positive charge was added to the halogen and the charge of the halogen was lowered by the same value. Hence, the net charge of the ESH-halogen pair remained identical as the initial halogen atom charge. None of the other atoms was modified. This model is well suited for high-throughput calculations since it does not require any additional quantum chemical calculation, once the atomic partial charges are known. On the other hand the effect of sigma-hole is reduced only to the vicinity of the halogen. Nevertheless, its performance on interaction energies was proven to be sufficient.[4]

The ESH was added to all halogen atoms except fluorine, which is known not to create halogen bonds in organic drug-like molecules. [5] The ESH parameters (charge, ESH-halogen distance) were chosen as follows and were not subject of any further optimization: (0.1 e, 1.0 Å) for chlorine, (0.2 e, 1.3 Å) for bromine, and (0.3 e, 1.6 Å) for iodine. These parameters follow the recommendation in Ref. 14 and also the known features of halogens, where iodine exhibits the largest σ-hole and chlorine the smallest. The ESH-halogen distance was, however slightly shortened when compared with Ref. 14 (i.e. 1.3 Å vs. 1.5 Å for bromine). Large ESH-halogen distance caused problems with the docking algorithm. Consequently, the change in improved scoring arises mainly from the improved electrostatics and also from the shape complementarity between ligand and receptor (distances being corrected for the presence of dummy atom mimicking the sigma-hole).

Molecular docking was performed using UCSF DOCK6.5 suite, [6] using a grid scoring, in an implicit solvent. The grid spacing was 0.25 Å, and the grid box included 12 Å beyond the ligand binding site. The energy score has been regarded as a sum of electrostatic and Van der Waals contributions. In the course of the docking procedure, the ligand was subjected to 2500 cycles of molecular-mechanical energy minimization. The number of maximum orientations was 5000.

References:

1 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605.

2 A. Jakalian, D. B. Jack, C. I. Bayly *J. Comput. Chem*., 2002, **23**, 1623.

3 The script used to introduce ESH into MOL2 files is available upon request.

4 M. Kolář, P. Hobza, *J. Chem. Theory Comp*., 2012, **8**, 1325.

5 P. Politzer, K. E. Riley, F. A. Bulat, J. S. Murray, *Comput. Theor. Chem.*, 2012, **998**, 2.

6 P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, I. D. Kuntz, *RNA*, 2009, **15**, 1219.

H

Publication 7 – Halogen Bonds in Drug Development (in Czech)

Reproduced by permission from Vesmír s. r. o., Vesmír 91, pp. 552-553, 2012.

stvo u „plazů“, neboť dinosaurům příbuzní krokodýli a ptáci o nakladená vajíčka a vyklubaná mláďata pečují. Jakýkoliv přímý doklad znaků a chování u vymřelých zvířat je samozřejmě užitečný, hlavně když se nám mapování evoluce znaků u žijících skupin nedaří. Občas se díky vzácným okolnostem zachovaly stopy dinosauřích stád nebo bohatých společenstev savců a ptáků a nedávno se podařilo objevit ve Spojených arabských emirátech 14 sloních tras dlouhých 200–300 metrů. Tyto fosilizované stopy jsou datovány do svrchního miocénu, kdy se tu vyskytovalo hned několik druhů chobotnatců z různých rodů. Počtem nalezených jedinců a vazbou na otevřenější biotopy by nejpravděpodobnějším „pachatelem“ stop mohl být *Stegotetrabelodon syrticus*. Třináct ze čtrnácti tras směřovalo jedním směrem, což v kombinaci s různou velikostí stop naznačuje skupinový způsob života tehdejších chobotnatců. Tyto trasy křižovala jedna trasa od mnohem většího zvířete, což by mohl být samec-samotář (a pak by sociální struktura plně odpovídala dnešním slonům), ale mohlo jít i o jiný druh chobotnatce. Protože sirény a damani netvoří sociální skupiny typu slonů, údaje o socialitě slonů získané z fosilních dat se jistě hodí. (Biology Letters, doi: 10.1098/rsbl.2011.1185, 2012)

Jan Robovský, PřF JU

Halogenová vazba

aneb Popletené náboje novou nadějí pro medicínu

Říká se, že protiklady se přitahují. I chemik, studující stavbu biomolekul, často použije tuto jednoduchou poučku, týkající se ovšem v jeho případě nabitých atomů. Od základní školy platí, že dva souhlasné náboje se odpuzují, pokud ovšem nejde o zvláštní druh nekovalentní interakce – halogenovou vazbu. Ta totiž ono tvrzení na první pohled zcela popírá.

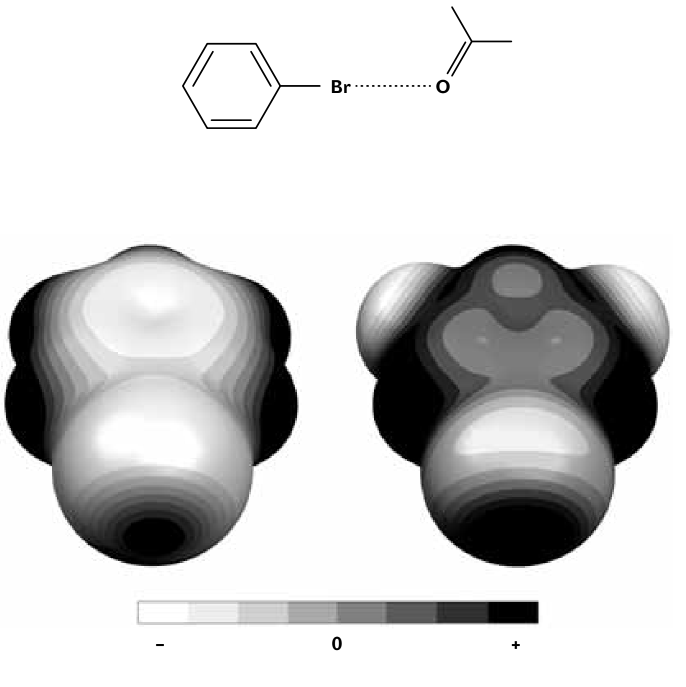
Halogenová vazba byla objevena relativně nedávno, když chemikové studovali atomární strukturu krystalů halogenovaných sloučenin. Pomocí rentgeno-strukturní analýzy vědci objevili, že halogen vázaný na jednu molekulu v krystalu se nachází velmi blízko kyslíku na vedlejší molekule. Tento motiv se opakoval s periodickou přesností, ale něco zde nehrálo.

Při pohledu do periodické tabulky zjistíme, že halogeny (např. chlor) mají vyšší elektronegativitu než uhlík, což znamená, že halogen bude mít tendenci odtáhnout elektrony z uhlíku blíž k sobě, čímž ovšem získá částečný záporný náboj. Není tajemstvím, že kyslík je taktéž elektronegativnější než uhlík, a proto má částečný záporný náboj podobně jako halogen. Jak je tedy možné, že se v krystalické látce halogen a kyslík vyskytují tak blízko sebe? Naše úvahy nás přece vedou k závěru, že by se tyto záporně nabité atomy měly podle Coulombova zákona odpuzovat.

Právě analýza krystalů dokázala, že tomu tak není a molekuly nejenže se neodpuzují, ony se naopak směle přitahují. Důvodem je tzv. σ-díra. Přesné kvantově-chemické výpočty ukázaly, že halogen sice díky vyšší elektronegativitě elektrony přitáhne, ale zjednodušeně řečeno je nerozmístí symetricky kolem sebe, nýbrž do jakéhosi prstýnku (viz obrázek, spodní část). Kromě prstýnku tak vznikne i místo s kladným nábojem na špičce halogenu a tento kladný náboj si už se záporným nábojem blízkého kyslíku zcela rozumí (viz obrázek, horní část).

Toto vysvětlení přišlo na začátku 21. století, ale přineslo nový pohled i na studie sto let staré. Vždyť první zmínka o nekovalentní vazbě mezi halogenem a elektronegativním atomem se objevila již v roce 1863 v práci jistého Fredericka Guthrie, který studoval komplexy amoniaku s jódem. V druhé polovině 20. století se začaly objevovat krystalické materiály, k jejichž struktuře halogenová vazba významně přispívala. Mimo krystalické materiály byl tento typ nekovalentní interakce objeven i v několika desítkách protein-ligandových komplexů. Například receptor jodovaného hormonu thyroxinu využívá hned několika halogenových vazeb k tomu, aby hormon do svého aktivního místa efektivně navázal. Současná medicinální chemie proto k halogenové vazbě upírá značnou pozornost.

Za povšimnutí stojí několik jejích pozoruhodných vlastností. Především: halogenová vazba je velmi směrová (obdobně jako vodíková vazba) a její síla roste s atomovým číslem příslušného halogenu, tedy v řadě Cl<Br<I. Fluor na organických molekulách nevykazuje σ-díru a halogenovou vazbu proto obvykle netvoří. Mimo to je kvalita halogenové vazby do jisté míry „laditelná“. Velikost σ-díry, tj. jejího kladného náboje, je totiž nastavitelná chemickým okolím halogenu, především pozicí dalších elektronegativních atomů, např. fluoru (viz obrázek, spodní část). Ne náhodou obsahuje



Nahoře bromobenzen nekovalentně vázaný s acetonem. Energie potřebná k jejich odtržení činí přibližně 8 kJ/mol (pro srovnání: energie potřebná k roztržení vodíkové vazby dvou molekul vody je asi 20 kJ/mol). Dole rozložení náboje okolo molekuly bromobenzenu (vlevo) a 1-bromo-3,5-difluorobenzenu (vpravo). Tmavě jsou zobrazeny kladně nabité oblasti, světle záporně nabité oblasti. Tmavý terčík v popředí se nazývá sigma-díra.

Kormorán subtropický (*Phalacrocorax brasilianus*), snímek © Stanislav Vaněk.

až 40 % v současnosti studovaných léků některý z halogenů a existuje opodstatněná naděje, že účinnost těchto léků lze hrou s σ-dírou ještě zvýšit.

O aktuálnosti tématu halogenové vazby svědčí fakt, že členové asociace IUPAC vedou od roku 2010 jednání o přesné definici této nekovalentní interakce. To, že stále nedospěli k uspokojivému závěru, může naznačovat, že náboje popletly hlavu i jim. **Michal Kolář, ÚOChB AV ČR a PřF UK**

Jak moc zvláštní je kormorán galapážský?

Galapážské ostrovy jsou proslavenou přírodní laboratoří, která značnou měrou přispěla k formulaci evoluční teorie Charlese Darwina. Hostí vskutku pozoruhodné tvory – třeba obrovité želvy, mořského leguána, „pěnkavy" a jiné pěvce snad se všemi možnými variacemi zobáků a velikostí, tučňáka (Galapágy leží na rovníku, ale až sem zasahuje chladný Humboldtův/Peruánský proud) nebo nelétavého kormorána. Poslední jmenovaný není zrovna typickým zástupcem své skupiny, neboť má kratičká křídla, což se v kombinaci s větší váhou promítlo ve ztrátě jeho letových schopností. Z kormoránů vybočuje i sekvenční polyandrií, což znamená, že samice mezi sebou bojují o samce, s nimiž zakládají postupně několik snůšek. Každý z několika samců se o vajíčka stará a posléze vychová i vylíhnutá mláďata. Dodejme, že jde o druh poměrně vzácný s asi 1400 jedinci (odhad z r. 2006). Pro jeho odlišnosti mu byl některými zoology vymezován samostatný rod *Nannopterum*, jiní v něm spatřovali jen výrazně změněného kormorána r. *Phalacrocorax*, do kterého patří třeba evropští kormorán velký a k. malý. Molekulárně-fylogenetické zhodnocení kormorána galapážského potvrdilo druhý pohled, neboť je blízce příbuzný americkým kormoránům r. *Phalacrocorax*, kormoránu ušatému (*P. auritus*) a k. subtropickému (*P. brasilianus*). Od nich se odštěpil před asi 2 mil. let, je tedy podobně starý jako předek proslavených galapážských „pěnkav" (2,3 milionu let), mladší než tučňák galapážský (4 miliony let) a výrazně starší než káně galapážská (300 tisíc let). Tato datace otevírá zajímavou otázku, kam prakormorán dorazil, neboť ačkoli dnes žije na ostrovech Isabela a Fernandina, Isabela se objevila před 500–800 tisíci lety a Fernandina teprve před 70 tisíci let. Možná na západní části ostrova Santa Cruz, který se zformoval asi před 2,2–2,3 milionu let... (Molecular Phylogenetics and Evolution 53, 94–98, 2009)

Jan Robovský, PřF UK

Neobvyklý metabolismus

Olavius algarvensis je máloštětinatec čili živočich ze stejné skupiny jako třeba žížala nebo nítěnka. Žije ve dně pod porosty mořských trav nedaleko italského ostrova Elba. Je to prostředí velmi chudé na živiny. Olavius nemá trávicí soustavu, proto hostí pod pokožkou svého těla pět druhů symbiotických bakterií, které ho *de facto* živí.

Společenstvo máloštětinatého červa spolu s bakteriemi má velmi neobvyklý metabolismus a několik dalších unikátních vlastností, které odhalujeme teprve v současné době. Ústředním systémem je dosud nepopsaná metabolická dráha. Živiny získávají bakterie z odpadních pro-