Charles University in Prague

Faculty of Science

Ph. D. study program: Modelling of Chemical Properties of Nano- and Biostructures

Mgr. Jiří Hostaš

Accurate Quantum Mechanical Calculations on Noncovalent Interactions: Rationalization of X-ray Crystal Geometries by Quantum Chemistry Tools

Doctoral Thesis

Supervisor:

Prof. Pavel Hobza, DrSc., FRSC, Dr. h. c.

Supervisor − consultant:

Doc. RNDr. Jan Řezáč, Ph.D.

Institute of Organic Chemistry and Biochemistry,
The Czech Academy of Sciences

Praha, 2017

Univerzita Karlova v Praze

Přírodovědecká fakulta

Studijní obor: Modelování chemických vlastností nano- a biostruktur

Mgr. Jiří Hostaš

Přesné Kvantově Mechanické Výpočty Nekovalentních Interakcí: Racionalizace Rentgenových Krystalových Geometrií Aparátem Kvantové Chemie

Disertační práce

Školitel:

Prof. Pavel Hobza, DrSc., FRSC, Dr. h. c.

Školitel − konzultant:

Doc. RNDr. Jan Řezáč, Ph.D.

Ústav organické chemie a biochemie,
Akademie věd České republiky, v.v.i.

Praha, 2017

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, 03.03.2017

Acknowledgement

Creating a PhD thesis is neither personal achievement nor individual experince. Presented work would be not be possible without so many people and I would like to thank them for helping me during my studies.

First of all, I would like to express my sincere appreciation to prof. Pavel Hobza for everlasting guidance, patience and support during my student years in Canon and IOCB building.

I am most grateful to doc. Jan Řezáč for providing countless advise and critical remarks along the way that helped to sharpen my knowledge and look critically at my scientific work. I would like to acknowledge the valuable input of prof. Robert Glaser who introduced me into several projects related to supramolecular chemistry. I would like to thank to my colleages, their friendship is priceless in my academic and personal life. To be more specific, I appreciate stimulating and welcoming academic environment provided by Adam, Ela, Jindra, Tom, Robo and others.

Most importantly, at a personal level, I would like to thank my parents for support during my studies in both Ostrava and Prague. Without their understanding I would not be able to endure this journey. In addition, I wish to extend my warmest thanks to my sister Jana who has always encouraged me. Last but not least, I take this opportunity to record my sincere thanks to those closest to my heart Katka, Miky, Sůňa and to all others who gave me encouragement and mental support in the last weeks.

ABSTRACT

There is a need for reliable rules of thumb for various applications in the area of biochemistry, supramolecular chemistry and material sciences. Simultaneously, the amount of information, which we can gather from X-ray crystal geometries about the nature of recognition processes, is limited. Deeper insight into the noncovalent interactions playing the most important role is needed in order to revise these universal rules governing any recognition process. In this thesis, systematic development and study of the accuracy of the computational chemistry methods followed by their applications in protein•DNA and host•guest systems, are presented.

The non-empirical quantum mechanical tools (DFT-D, MP2.5, CCSD(T) *etc*. methods) were utilized in several projects. We found and confirmed unique low lying interaction energies distinct from the rest of the distributions in several amino acid−base pairs opening a way toward universal rules governing the selective binding of any DNA sequence. Further, the predictions and examination of changes of Gibbs energies (ΔG) and its subcomponents have been made in several cases and carefully compared with experiments. We determined that the choline (Ch+) guest is bound 2.8 kcal/mol stronger (calculated ΔG) than acetylcholine (ACh+) to self-assembled triple helicate rigid cage, corresponding a K(Ch+)/K(ACh+) = 109 that is in fairly good correlation with the experimental value of 20. Finally, excellent correlation between theoretical and experimental ΔG has been reported ($\rho^2 = 0.84$) for cucurbit[*n*]uril (CB[*n*]) host•guest systems. Here, prediction has been made that binding in CB[7]•Diam-4,9-di(NMe_2propanoNH_3) complex could become next world record in the world of noncovalent interactions. This diamantane derivate is now being synthetized. Clearly, these findings demonstrate that the computational chemistry has a solid position as the complementary source of information to the data obtained from the experiments.

ABSTRAKT

Spolehlivá a jednoduše aplikovatelná pravidla jsou potřebná v oblasti biochemie, supramolekulární chemie i materiálových vědách. Zároveň množství informací, které můžeme získat z rentgenových krystalových struktur o povaze rozpoznávacích procesů, je omezené. Lepší pochopení nekovalentních interakcí, které hrají nejdůležitější roli, je potřebné pro přezkoumání univerzálních pravidel, řídících jakékoliv rozpoznávací procesy. V této práci je prezentován systematický vývoj a studium přesnosti výpočetních metod, doplněný aplikacemi na systémech bílkovina•DNA a hostitel•host. Ne-empirické kvantově mechanické nástroje (metody DFT-D, MP2.5, CCSD(T) *atd*.) byly využity v několika projektech. Našli a potvrdili jsme existenci unikátních nízko ležících interakčních energií, vzdálených od zbývajících distribucí v několika párech aminokyselina−báze, které otevírají cestu k univerzálním pravidlům řídícím selektivní navázání jakékoliv sekvence DNA. Dále byly v několika případech provedeny predikce a ověřeny změny Gibbsovy energie (ΔG) a jejich komponentů a nakonec byly pečlivě porovnány s experimenty. Stanovili jsme, že molekula cholinu (Ch+) je vázána o 2.8 kcal/mol silněji (vypočtením ΔG) než acetylcholin (ACh+) v samo-uspořádané tří helikální rigidní kleci, odpovídající K(Ch+)/K(ACh+) = 109, což je v poměrně dobrém souladu s experimentální hodnotou 20. Nakonec byla popsána výborná korelace mezi teoretickou a experimentální ΔG pro systémy hostitel•host s molekulou cucurbit[*n*]urilů (CB[*n*]). Byla provedena predikce, že vazba u CB[7]•Diam-4,9-di(NMe_2propanoNH_3) by se mohla stát novým světovým rekordem v nekovalentní vazbě. Výše zmíněný derivát diamantanu je nyní připravován experimentálně. Tyto výsledky jasně demonstrují pevnou pozici výpočetní chemie jako komplementárního zdroje informací pro experimenty.

Table of Contents

List of Abbreviations

MM	molecular mechanics
QM	quantum mechanical
PES	potential enery surface
ΔG	changes of Gibbs energies
NMR	nuclear magnetic resonance
CYPs	cytochrome P450 enzymes
FCI	full configurational interaction
ΔZPVE	change of zero-point vibration energy
RMSE	the root-mean-square error
RMSD	root-mean-square deviation
MSE	mean signed error
rRMSE	RMSE relative to the average absolute interaction energy
rMSE	MSE relative to the average absolute interaction energy
BOA	Born–Oppenheimer approximation
HF	Hartree-Fock
CBS	complete basis set
MPP	Møller-Plesset perturbation
SCS	spin component scaled
CC	coupled cluster
CCSD	coupled cluster singles and doubles
CCSD(T)	CCSD(T) method augmented by non-iterative connected triples
DFT	density functional theory

SQM	semiempirical QM methods
RESP	restrained electrostatic potential
RRHO	rigid rotor harmonic oscillator
KEM	kernel energy method
EDA	energy decomposition analysis
COSMO	conductor-like screening model
SMD	density-based solvation model
BSSE	basis set superposition error
GGA	generalized gradient approximation
BJ	rational damping to finite values according to Becke and Johnson
zero	standard "zero-damping" function
Ch+	choline
ACh+	acetylcholine
TMA+	tetramethylamonium
CB[n]	cucurbit[n]uril
FN-DMC	fixed node-diffusion Monte Carlo
NBO	natural bond orbital
WM	WaterMap software
PC	packing coefficients
MD	molecular dynamics

List of Figures

List of Tables

1 INTRODUCTION

The understanding of the rules regulating the recognition processes in the area of biochemistry, supramolecular chemistry, material sciences or molecular biology, is far from being complete.[1-3] In recent years, the continually increasing amount of structural data has opened the space for studies combining the techniques of X-ray crystallography, bioinformatics and computational chemistry.[4] Despite enormous endeavor, no unanimous recognition code applicable to all types of bindings in the realm of biochemistry has been described to date, though *e.g.* the conclusions about zinc finger domains and transcription activator-like effector proteins have already found their use in the genetic engineering.[5] This demonstrates that the understanding of binding selectivity principles has a great application potential in biotechnology, medicine and related areas.

In spite of the ongoing efforts, the number of new molecular entities that are approved for use as medicines per year is decreasing. The companies leading the development toward new pharmaceutical drugs have to deal with both, the growing cost of the clinical trials and the luck which seem to be at the basis of the discovery of the most drugs. The newly reported technologies, such as the DNA-linked Inhibitor Antibody Assay can speed up the screening significantly, however it still requires the drug molecule being synthesized and purified.[6] Fortunately, recent studies show that the trial and error is not the only options in the further development.[7] The usage of computational and experimental tools in the so-called "rational drug design" has the potential to dramatically reduce the number of trial compounds which have to be prepared, purified and tested in the process of a new drug discovery. It has been successfully demonstrated that binders of the Y220C mutant of the p53 tumor suppressor can be found with virtual screening methods.[8,9] Broad spectrum of similar applications requires thorough understanding of the nature of the noncovalent interactions and phenomena, such as hydrophobicity, high-energy water molecules and many-body effects. Research in these

areas is mandatory and has to be performed in hand with the development and the identification of efficient, accurate and suitable methods.

Here, the main focus is being put on the description and behavior of protein, DNA and various biomimetic complexes both, isolated and in its natural environments. Nowadays, the specific types of interactions which manifest metal-coordination or surface adsorption play crucial role in the large variety of applications.[10] Although we do believe that these complex systems will have a large impact in the near future, the accurate description of the biopolymers represents the first essential goal which is still to be achieved in the large systems containing thousands of atoms.[11] While the experiments successfully assess the strength of the binding in many instances, they mostly provide only limited information about the nature of the noncovalent interactions involved in the binding processes. Here, computational results can complement, provide insight or even predict the information received from experiment.[12]

For a very long time, the main focus had been put on the description of the dominant hydrogen bonding interactions.[13,14] Later, it was shown that the wide range of interaction motifs exist together with the hydrogen bonding: London dispersion, electrostatic, σ-hole interactions *etc.* and they need to be described adequately and in a balanced way.[15,16] Dispersion energy in particular plays an important role in biomolecular complexes and its accurate evaluation in systems of larger scale is of prime importance.[17] Further, charge transfer can play a key role in stabilization in charged systems; however, especially this last contribution is at the molecular mechanics (MM) level very often being described poorly or not at all. Nonetheless that is, up to now, the most widely used methodology in the majority of the cutting edge applications. These striking facts support the interest of our laboratory that lies in the deeper understanding of non-dynamic properties of molecules and their interactions. For the studies where binding may be primarily enthalpy-driven, the quantum mechanical (QM) methods bring many advantages in description of many body effects, electron and proton transfers, formation and dissociation of a covalent bond *etc*. However sometimes it is not straightforward to find out the role of enthalpy and entropy contributions *a priori*.[18] Enthapic contribution can now be treated, even at QM level, however the determination of entropic effects is still challenging for most of the

applications.[19] A various approaches are available at MM level but they are strongly dependent on the quality of the applied force field.[19,20] The accuracy of Potential Enery Surface (PES) description implies also the accuracy of the entropy analysis. Additionaly, the identification of the dominating conformations is a must because the number of considered binding modes is strongly limited by computing power. Recently, it has been shown that the accuracy of 2 kcal/mol can be achieved for estimates of changes of Gibbs energies upon binding (ΔG) in host•guest systems where one binding mode dominates over the others.[21]

Noncovalent interactions mentioned briefly above are well defined in sharp contrast to *e.g.* hydrophobicity which is interpreted as a property of molecule that is being repelled from water environment. Additionally, hydrophobicity requires much larger system size to be studied, when compared to easily isolated nature of noncovalent interactions.[22] Advantageously, all is being solved when appropriate solvation model is chosen because hydrophobicity is driven by the preferential interaction of the molecule with other partners than the solvent or more commonly preferential interaction of the solvent molecules with themself.[23] However, for the studies of both, noncovalent interactions and other effects such as hydrophobicity, it is crucial to have solid geometrical data at hand. Every studied image of molecule up to now was based on the data generated by crystallography.[24] For these reasons, a very fruitful cooperation between the areas of crystallography and computational chemistry, described also in this thesis, arised.

1.1 Use of Crystallographic Data in Computational Chemistry

Since 1914 and 1915 when first two Nobel prizes were awarded to Von Laue and Braggs, molecular crystallography dramatically expanded human understanding of the structure and function of many materials.[25-27] Contemporary crystallographic techniques providing very detailed structural information include X-ray, neutron and electron diffraction.[28] These data are afterwards used to refine proposed atomic arrangements inside of the sample.[29]

Over the past few decades, material sciences as well as day life were enriched by various applications of crystals reaching from crystal displays in mobile phones, computer and

TV screens to crystalline beads in catalytic car converters.[30,31] Nowadays, any studied image of a molecule is based on or use the data generated by crystallography. The main drawback represents the necessity of performing a crystallization of the studied sample that can be a difficult task. Flexible proteins, especially present inside of the membranes, were proven to be very difficult to study using X-ray crystallography.[32] Moreover, the orientation of the flexible molecules inside of the crystal can be different from its arrangement in natural environment. This phenomena is caused by *crystal packing*.[33] Therefore, information from X-ray crystallography is often combined with other techniques such as nuclear magnetic resonance (NMR) spectroscopy, electron micrographic or computational studies. These multidisciplinary studies are able to determine the structure of large assemblies (*e.g.* transfer RNA or ribosomes) and atomic details such as protonation states.[34] Crystallography played a prime role in filling the gaps in our knowledge and shed light on unexpected close contacts between atoms. Various studies of these close contacts revealed new types of interactions including halogen or chalcogen bonding in exotic species.[35]

Additionally, the crystallography uncover what happens when two complex regions of *e.g.* protein DNA-binding domain and the target DNA sequence, meet. Protein−DNA interactions are responsible for important processes in cells such as DNA replication, DNA repair and cell cycle regulation.[36] Further, the genetic information is stored inside of DNA which is tightly packed with histone proteins and in the same time the high fidelity recognition is required for the gene expression. Here, the understanding of rules governing the recognition process would constitute a major accomplishment in the fields of bioinformatics and computational biology. The possibility of specific base pair recognition in the major groove by amino acids accompanied by two hydrogen bond formation was first examined by Seeman.[37] It was described shortly after first experimental studies occurred and still constitutes an important tool guiding the prediction of protein−DNA binding sites. A second factor contributing to the specificity of sequence recognition is the DNA shape readout.[38] Local deviations from the native B-form have been observed in many protein−DNA complexes. It has been shown that the ability or tendency to adopt different conformations (*e.g.* kinks and bents) is sequence-dependent.[39] On the other hand an overall bend of the DNA double strand

represents a nonlocal effect that can enable the formation of interactions impossible in the B-form.[40-42]

Recently, research of cytochrome P450 enzymes (CYPs), family of biotransformation enzymes, was stimulated by emerging biochemical and X-ray structural data.[43] However, these data are based on the mutated or variously modified CYPs designed to enhance enzyme solubility. Despite the growth in the number of these complex structures the information about orientation of anchored CYPs remains very limited. In such situation computer modelling brought already much insight into several key aspects of the binding including position of ibuprofen in membrane or secondary structure predictions in cases when X-ray-resolved structure is not available.[44-46]

The continually increasing amount of structural data has opened space for theoretical studies in many areas and few of them were outlined above. On the other hand, there are several pilot studies using QM as the tools for resolution of X-ray structures in place of MM methodology dominating the field.[47,48] This represents a way forward because in many cases we may achieve perfect agreement with experiment, but still get some predictions wrong. It can be caused mainly by two reasons. Firstly, studied systems are outside of the parametrization set used for development of the force field methods. Secondly, the error cancellation behaves in various systems differently.

1.2 Accuracy of Production Calculations

The largely unsolved problems in computational chemistry are both the accurate predictions in large scale applications and the precision assessment of the applied methodology. Generally errors in any kind of calculation or measurement are defined as being systematic or random.[49] While the random errors cannot be predicted, the systematic errors are largely predictable in both magnitude and sign. The random errors propagate as the square root of the sum of squares. On the contrary, the systematic errors accumulate as a simple sum. Therefore it is highly advantageous for error analysis to routinely check not only the most studied root-mean-square-error but also average signed error (for definitions see section 1.2.2). It represents a shift of values in the measured or calculated values from the right value. It is important to note that it

vanishes when differences between individual measurements are calculated or only relative order of results is in the center of interest.

We have to consider following fundamental sources of the random error in computational chemistry: the accuracy of implemented numerical algorithm, truncation and round off errors. Computer device is not able to deal with certain numbers therefore they need to be rounded off. This is direct consequence of finite floating point number usage on computers.[50] The quality of the approximation is dependent on the word size and is typically negligible when compared to any other source of error. On the other hand, the truncation error is error connected to the given method because it takes place when series (both finite and infinite) need to be truncated to fewer number of terms.[51] These errors are typical for computational chemistry and are method and system dependent (see later). Because our intentions here are to study the noncovalent interactions in large systems containing several hundreds of atoms, the method accuracy cannot be routinely assessed by the exact non-relativistic energy calculated by *e.g.* Full Configurational Interaction (FCI).[52] However, one can study isolated errors of individual approximations at much smaller model system with a high-level method and compare it with the lower-level method where the respective approximation was used. Afterwards, the best overall method performing consistently in various scales ranging from few atom systems to few tens up to hundred is used afterwards.[53]

The ultimate test of methods represents comparison with experiment. However, there are two severe limitations in the field of noncovalent interactions. Firstly, the experimental data for the wide range of noncovalent interactions observed in nature are not available. Secondly, the measured quantity in experiments is not interaction energy (eq. 1) but the dissociation energy D_0. This quantity includes also on the deformation energy (E_{def}) and change of zero-point vibration energy (ΔZPVE) that requires description of not only electronic but additionally vibrational state of the system.[54]

$$\Delta E = E(\mathrm{AB}) - [E(\mathrm{A}) + E(\mathrm{B})] \qquad (1)$$

These limitations are usually overcomed by comparison with more accurate methods instead of the direct comparison with experiment. It is important to note here, that the conclusions from such studies are not limited only to interaction energies. However, it

provides a good clue on how well the methods would work in all kind of calculations, *e.g.* determination of more complex quantities that rely on the precise description of the potential energy surface.[11] We will now describe the data sets of geometries and interaction energies of dimers playing essential role in accuracy assessment, description and studies of noncovalent interactions. Afterwards, we will describe error analysis used throughout the text, the fundamental approximations and the most popular computational methods used nowadays.

1.2.1 Data Sets

Recent advances in methodology and computer hardware enabled construction of various data sets useful for both noncovalent interactions studies and thorough tests of methods. These data sets commonly include geometries and interaction energies in the local minima representing an interaction motif *e.g.* hydrogen bonding, π-π and σ-hole interaction. However, such data describe the system's behavior only in the one point in the PES. On the contrary, the experimental D_0 value brings additional information characterizing the PES around the equilibrium geometry.[11] Part of this information can be recovered when we calculate energy not only in the potential energy minimum but also in additional points along the dissociation curve. The prime example of such data set is *S66(x8)* data set where are included the most common interaction motifs between organic and biomolecular building blocks.[55]

The most important feature of any benchmark data set is its size. It has to be large enough to provide data for meaningful statistical analysis. Additionally, it should not be excessively large from the analysis point of view as well as the computational one.[56] This means that any important type of interaction has be present in the several versatile model systems and the system sizes included in the data set have to be manageable for calculations using the unified setup and methodology.[55] Further, the interaction energies in the set should be roughly in the same range of values because the systems interacting too weakly do not contribute to the error statistics equally as the ones interacting strongly.[57] The strongly interacting systems can override the contribution of the rest of the systems in the data set. This can be partly solved by calculations of relative errors.

However, this approach fails when interaction energies are close to zero. Finally, for method validation a great care has to be taken in distinguishing between data sets utilized for parametrization of the given method and an independent data set suitable for validation. This information has to be always included.

Recently, several specialized data sets were published among which following will be used throughout the text: *AA-sidechain*[4] (amino acid side chains), *L7*[53] (7 large complexes chosen for studying dispersion interaction), *1hsg*[58] (decomposition of the HIV-II protease crystal structure with a bound ligand indinavir into 21 interacting fragment pairs), *Charged HB*[57] (charged hydrogen bonds), *A24*[59] (set of 24 small noncovalently bound dimers small enough to be calculated with high level of theory), *S22*[60], *S22x5*[61] *Halogens* and *Halogensx10* data sets.[62] Most of these data sets include useful system division according to a dominant stabilization term based on DFT-SAPT analysis.[63]

Three additional data sets that consist of the structures and relative energies of different conformations of complexes and molecules will be utilized: the *Goddard water* set of water clusters[64], *Wat5*[65] (relative energies of different conformations of the water pentamer) and *PCONF*[66] (conformations and relative energies of the FGG tripeptide). The terms marked in italics represents the name of the data used in the text bellow.

1.2.2 Error Analysis

There are multiple statistical measures that can be used for the evaluation of an error in data set highliting different information about error measurements. Here, we will give an overview of the most commonly used statistical tools.

When a single error measure need to be considered, the root-mean-square error (*RMSE*) is the natural choice. It is being referred, especially for geometry differences, also as root-mean-square deviation (*RMSD*). When quantities X are being compared to reference quantities X^{ref}, it is defines as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - X_i^{ref})^2} \quad , \qquad (2)$$

where *N* stands for size of the data set. Further, this is the most robust and commonly used error function to be minimized in the parametrization of methods (see later for optimization of dispersion correction in section 3.2).[11] In addition to the *RMSE*, which is sensitive to the most problematic cases, it is advantageous to utilize mean signed error (*MSE*) that refers to the simple average of errors:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - X_i^{ref}) \quad . \quad (3)$$

It is not a measure of the overall accuracy but it isolates the systematic part of the error. Finally, it can be highly important to list and check also largest (unsigned) error.

For a relative comparison between the interaction energies of the different interaction types or different data sets with different ranges of energies, we often utilize the *RMSE* as the percentage of the average interaction energy in the group (*rRMSE*). Similarly to *rRMSE*, we do define also the relative mean signed error (*rMSE*). However, this analysis can fail when energies approach zero value because the relative error in such cases is very high.

1.3 Fundamental Approximations in Computational Chemistry

In this section, we would like to briefly summarize the conventional approximations representing the most important tools of quantum chemistry: The Born–Oppenheimer, frozen core and non-relativistic approximation.

1.3.1 The Born–Oppenheimer Approximation

Without a doubt the Born–Oppenheimer approximation (BOA) belongs nowadays among the most used approximations. It is based on the large mass difference between

electrons and nuclei. It follows the assumption that the electrons are able to be accommodated adiabatically in the field of the nuclei that are being sluggish in their motion relative to the electronic motions in the system.[67] In practice, at any given moment the electrons feel a Hamiltonian $\hat{H}_e$ that depends only on the nuclei positions at that instant. Therefore, the nuclei are treated as being stationary and their kinetic energy is neglected, while one is solving only electronic part of the Schrödinger equation written as:

$$\hat{H}_e|\Psi_e\rangle = E_e|\Psi_e\rangle \tag{4}$$

where

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne} + \hat{V}_{nn} \quad . \tag{5}$$

When such approximation break down? There are numerous special phenomena that originate from the failure of the BOA. It typically happens when two electronic energies of two separate electronic states become very close to each other at some point. A classic example is avoided crossing between two states (*e.g.* ionic and covalent states in the ground state of sodium chloride).[68] Another example of BOA breakdown was described in electron-transfer reactions involving long-distance electron tunneling. It happens when the time spent by electron in the barrier region between redox centers is comparable to periods of the nuclei vibrations.[69]

However, the BOA represents very solid approximation when two PES are well separated. Typically, it is the case in the realm of noncovalent interactions, and therefore it is safe to separate motions of electrons and nuclei. Additionally, only the systems containing few electrons can be studied without this approximation.[70]

1.3.2 The Frozen Core Approximation

In the last paragraph we proposed separation of electronic and nuclei degrees of freedom based upon the significant difference between the mass of an electron and the masses of the nuclei. We can push it further and lower even more the number of variables that

need to be handled in the calculations. One can assume that upon the interaction of two molecules the interesting chemistry happens only in the valence electron region. The resulting approximation is referred as the frozen core approximation stating that the lowest-lying molecular orbitals are constrained to be double-occupied.

Recently, Řezáč *et al.* prepared balanced data set, named *A24* (see section 1.2.1), of 24 small noncovalently bound dimers composed from period 1 and 2 elements.[59,71] These systems were studied by the highest possible level of theory applicable up to date. It is a must to study not only isolated types of interactions but more importantly a range of different systems in a systematic manner and of the same composition as are present in the target applications. Here, they included calculations with and without core electrons correlation at the CCSD(T)/CBS level (see description later). The maximum errors showed to be as large as: 0.04 and 0.01 kcal/mol in case of water ammonia and ethene ethyne pairs counting only 0.6% and 1.3% of interaction energy, respectively. Neither these reported marginal differences nor average relative error 0.57% of interaction energy for whole data set justify the computational expense accompanied with all core correlation calculations. Additionally, a second motivation exists for frozen core approximation because approximate solutions (effective potentials) open a way how to empirically include the non-relativistic effects that are described in next paragraph.[72]

1.3.3 The Non-Relativistic Treatment

Another approximation carried out routinely for small atom number elements is neglection of relativistic effects *i.e.* finite speed of light and the gravitational effects at the size scale of atoms.

Among the most celebrated manifestations of relativity are the yellow color of solid gold and the low melting point of the mercury. Lesser known example is the magnitude of electric potential difference in the lead-acid car battery being out of 90% attributed to the relativistic effects, therefore one can argue that "cars start due to relativity".[73]

Despite these cases with heavy atom effects, it is clearly less important for systems containing only first-two-row elements present at *A24* data set.[59] Authors studied the relativistic effects by means of parameter-free fourth-order Douglas−Kroll−Hess

(DKH) Hamiltonian in an all-electron CCSD(T) calculation using the extended aug-cc-pCVQZ-DK basis set. The average relative error caused by omitting the relativistic effects is 0.14% or in absolute range from −0.003 to 0.017 kcal/ mol. It represents the smallest effect studied in the original publication.[59]

2 METHODS

Both, the total energy (eq. 6) and interaction energy (eq. 1) are in *ab initio* computational chemistry commonly divided into two parts[74]: Hartree Fock (E_{HF}) and electron correlation (E_{corr}) contributions.

$$E = E_{HF} + E_{corr} \tag{6}$$

In general, the numerical methods capturing the most out of the E_{corr} are very computationally demanding and therefore cannot be used but for the small systems composed of few tens of atoms. Often we desire to understand a large system, *e.g.* whole protein, however these methods are not directly applicable for such system size. Therefore, the most expensive methodologies are used sololy for calculations of precise energies of small model systems. Afterwards, these data are assembled into data sets (such as the *A24* set described above, see section 1.2.1). They are utilized for method comparison and accuracy assessments of the methods applicable to large systems. The accurate and efficient calculations of correlation energies for diverse systems play the key role in this thesis.

2.1 Extrapolation to the Complete Basis Set Limit

The energy eigenfunctions (the Schrödinger electronic wave-functions $|\Psi_e\rangle$ in eq. 4) are constructed from finite number of one-electron wave functions (basis functions) while the true wavefunction would need a complete *i.e.* infinite orthogonal set. The most severe errors of *ab initio* methods typically originate from the slow energy convergence

with respect to the number of basis functions describing the electronic wave-function $|\Psi_e\rangle$.

Fortuitously, the basis set cardinal number (X) dependent extrapolation schemes have been shown to work very well. Large variety of these equations, which utilize this approach for obtaining complete basis set (CBS) limit results, has been published. Currently, the one proposed by Helgaker is the most commonly used.[75,76] We used here solely the two-point extrapolation scheme. It was pointed out that results calculated with the lowest basis set can be of much lower quality and can easily spoil results of the two other more expensive calculations.[59] The HF and correlation energy are being commonly extrapolated separately:

$$\Delta E_X^{HF} = \Delta E_{CBS}^{HF} + A\exp(-\beta X) \quad , \quad (7)$$

where ΔE_{CBS}^{HF} is basis set limit of the HF energy, A is pre-factor and β is a parameter that was fitted previously by Helgaker for two combinations: 1.43 when using double and triple zeta basis sets and 1.54 for triple and quadruple zeta basis set combination.[75,76] The correlation energy is determined from following formula:

$$\Delta E_X^{corr} = \Delta E_{CBS}^{corr} + BX^{-3} \quad , \quad (8)$$

where ΔE_{CBS}^{corr} is correlation energy extrapolated to the CBS limit and B is pre-power factor. Both terms are determined in extrapolation procedure. This scheme was developed together with the systematically improved family of Dunning‘s basis sets called Correlation Consistent Polarized Valence Double, Triple or Quadruple-Zeta basis sets (cc-pVXZ).[77] These basis sets are very often augmented with a set of diffuse functions (aug-cc-pVXZ). The accuracy of this and similar schemes is still an open question, because the performance of the given basis set is strongly dependent on the chosen method for the description of electron correlation energy. On the contrary, the HF energy has been shown to converge fairly quickly with the basis set size and even energies from QZ size basis sets are commonly used without any extrapolation.[59]

2.2 Post-HF Calculations of Correlation Energy

2.2.1 Møller-Plesset Perturbation Theory

In order to study the correlation energy one has to pass from HF method to *e.g.* perturbative methods, used in many areas of physics and chemistry. The perturbation of the Fock operator was introduced into computational chemistry by Møller and Plesset in 1934.[78] The first correction to HF energy arise from second order of perturbation resulting in MP2 correlation energy. Nowadays, Møller-Plesset perturbation (MPP) theory represents the most favorite approach for correlation energy calculations. This correction accounts typically for 80-90% of the correlation for moderate computational expense scaling roughly $\propto O(N^{4-5})$, where N stands for number of electrons. The truncations of the perturbation series up to third, fourth, *etc*. order are denoted as MP3, MP4, respectively.

Recently, it has been shown that remarkable results can be achieved when MP2 and MP3 results are combined and we use for such method acronyms MP2.5 or MP2.X. These methods, introduced by our laboratory, are based on the fact that MP2 method generaly overestimates the interactions and MP3 underestimates by about the same amount. The amount of over- and under-estimation is system and basis set dependent but it was observed that the fourth order effect of triples is roughly 50% of the MP3 third order correction, therefore this MP2.X scaled version is delivering results of MP4 quality comparable to the most precise methods.

It has been shown that with increasing size of basis set the golden ratio between MP2 and MP3 is slowly approaching approximately to 0.5.[79] MP2.5 method is known to provide excellent interaction energies for various types of noncovalent complexes and it is typically advantageous to use a composite scheme (eq. 9) similar to the one used for CCSD(T) and other methods (see later).

$$\Delta E^{MP2.5/CBS} = \Delta E^{HF/CBS} + \Delta E^{MP2corr/CBS} + [\frac{1}{2}(\Delta E^{MP3} - \Delta E^{MP2})]_{mediumsizebasisset} \qquad (9)$$

Here, equation 3rd term, named ΔMP2.5 correction, is calculated from the difference between the MP3 and MP2 interaction energies in a medium size basis set [*e.g.* aug-cc-pVDZ or 6-31G*(0.25)].[80-83]

The idea of scaling of energy components determined by MPP methodology gave birth one additional class of methods: spin component scaled (SCS) MP2 methods. In 2003 it was introduced for the first time by Grimme[84], who scaled the antiparallel (singlet) and parallel (triplet) spin components of the correlation energy by formula:

$$E^{SCS-MP2} = p_S E_S + p_T E_T \quad , \tag{10}$$

where p_S=6/5 and p_T=1/3 are scaling parameters while E_S and E_T are singlet and triplet components of the MP2 energy, respectively. The p_S scaling parameter was derived from theory, contrary to p_T that was parametrized against a data set of reaction and atomization energies. However, several studies showed that despite the success in improving the description of dispersion bound systems the different types of noncovalent complexes (*e.g.* hydrogen bonding) are described by SCS-MP2 and MP2 methods only comparably.[85] Since then, several similar and more empirical approaches were published such as: SCS-MI-MP2, SCSN-MP2 or SSS-MI-MP2.[86,87] All three methods were fitted against the *S22* or a specialized data set of nucleic acid base pairs. Both data sets contain reference CCSD(T) interaction energies. SCSN-MP2 and SSS-MI-MP2 utilize only one, the triplet component of MP2 energy. The most accurate SCS method for description of noncovalent interaction, SCS-MI-MP2, was tested in this thesis (see later).

Finally, the explicitly correlated methods represent another important direction that significantly improves basis set convergence toward CBS limit. The standard atomic orbital basis set is not properly describing the singularity of the Coulomb operator in the situations when distance between two electrons approaches zero. In this position the so-called correlation cusp occurs resulting in slow convergence of correlation energy with respect to the basis set size.[88] However, the explicit correlation effect can be added as a correction to the calculation in an AO basis resulting in R12 or F12 class of methods.[89-91]

2.2.2 Coupled Cluster Method

In this chapter coupled cluster (CC) methodology applied for noncovalent interactions is briefly outlined.[92] It is based on the exponential form of the wave operator that is expanded into the clusters of excitation operators. The size-extensivity and the rather fast convergence toward full configurational interaction value represents the most important features of CC ansatz. Due to the advances in both software and computer hardware the affordability of the methods is increasing. The computational cost of coupled cluster singles and doubles (CCSD) method scales as $\propto O(N_{virtual}^4 N_{occupied}^2)$ where $N_{virtual}$ and $N_{occupied}$ is the number of virtual and occupied orbitals, respectively. Inclusion of perturbative estimate of the energy contribution arising from triples leads to the so-called CCSD method augmented by non-iterative connected triples [CCSD(T)].[93] It increases the scaling to $\propto O(N_{virtual}^4 N_{occupied}^3)$. CCSD(T) level of theory at the CBS limit is generally considered to provide high-confidence benchmark interaction energies for many small complexes of up to about 50-70 atoms. Interaction energy $\Delta E^{\mathrm{CCSD(T)/CBS}}$ is calculated in following way:

$$\Delta E^{\mathrm{CCSD(T)/CBS}} = \Delta E^{\mathrm{HF/CBS}} + \Delta E^{\mathrm{MP2corr/CBS}} + (\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}})_{\mathrm{mediumsize\,basisset}} \quad . \quad (11)$$

The $\Delta E^{\mathrm{CCSD(T)/CBS}}$ determination involves two terms. The first term represents the extrapolated MP2 interaction energy ($\Delta E^{\mathrm{MP2/CBS}}$) determined using the Helgaker technique with the cc-pVXZ basis sets very often augmented with diffuse function. The second term is the ΔCCSD(T) correction and it is determined as the difference between $\Delta E^{\mathrm{CCSD(T)}}$ and ΔE^{MP2} energies calculated with a small or medium size basis set. The CCSD(T) calculations are impractical for larger complexes, and thus only the first two terms can be extrapolated to the CBS but advantageously the third term is much less dependent on the size basis set.[64] This has been thoroughly studied by Řezáč *et al.* where they analyzed three groups of the composite CCSD(T) schemes, differing in the size of basis sets used in ΔCCSD(T) correction, MP2 correlation and HF terms.[59] They pointed out that the CCSD(T) term calculated with aug-cc-pVDZ basis set combined with any of the MP2 terms tested yielded a reasonable error of about 2%. This is also the method of choice for assembling of benchmark data sets where aug-cc-pVT(D)Z

and aug-cc-pVQ(T)Z basis sets are used for MP2 calculation and aug-cc-pVDZ for ΔCCSD(T) correction term.[55] Additionally, very low errors less than 2% due to error cancellations were found also when 6-31G**(0.25,0.15) basis set was utilized (the exponent of the polarization function on hydrogen atoms is set to be 0.15 and in other elements reduced to 0.25). This shows great promise also for the calculations with smaller variant of above mentioned basis set trimmed off polarization functions on hydrogen atoms 6-31G*(0.25) for post-HF method (see later).

In previous section we mentioned the empirical approaches how to improve description of MP2 method by scaling of the same spin and opposite spin components of MP2 energy. The same approach has been successfully applied in case of CCSD method resulting in SCS-CCSD and SCS-MI-CCSD methods.[94,95] Both methods surpass their SCS-MP2 and SCS-MI-MP2 counterparts significantly.

2.3 High-Accuracy Alternative for Cross Check of Post-HF Methods: Quantum Monte Carlo Methods

Only few *ab initio* computational methods can achieve the desired accuracy of 1 kcal/mol (see later). Methods described above, such as FCI, CCSD(T) and MP2.5, suffer from exponentional or high polynomial scaling and they are therefore feasible for molecules with only up to ~500 electrons and medium size basis set. Therefore, classes of linear- and low-order polynomial scaling methods represent important direction of method development. One such attempt is fixed-node diffusion Monte Carlo (FN-DMC) method providing $\propto O(N_{electrons}^{3-4})$ scaling and intrinsic massive parallelism.[96] However, the large prefactor to the CPU cost and missing easy-to-use implementation (in a black box fashion) still limit its applications in computational chemistry community.

The FN-DMC calculations in this thesis were performed with software packages GAMESS[97] and QWalk[98]. Time step of 0.005 and Slater-Jastrow trial wave-functions were used.[96] For construction of the single determinant Slater part the valence B3LYP orbitals were utilized and expanded in valence triple-zeta basis sets (augmented with *s* functions while *f* and *g* shells were removed) and the effective-core potentials were

used for representation of the atomic cores.[99] The Jastrow part contained electron-electron and electron-nucleus terms.[100]

2.4 Density Functional Theory

During past decades the Density Functional Theory (DFT), a method dominating solid state physics, was brought into the spotlight in the fields of chemistry. With the increasing number of applications to large systems, it has become apparent that the accurate description of the London dispersion plays a fundamental role. However, standard density functionals (LDA, GGA, meta-GGA and hybrids) are based on the electron density that has local (or semi local) character resulting in inability to properly capture most of the long range van der Waals interaction.

In theory, the DFT is formally exact providing the same accurate prediction of observables that a solution of the electron Schrödinger equation provides. However, in reality we do not have the true functional at hand. It is since when 1964 the Hohenberg-Kohn proof was formulated and more rigorous foundations of DFT were developed.[101] It says that „The external potential is a unique functional of the electron density only." and „The functional that delivers the ground state energy of the system, gives the lowest energy if and only if the input density is true ground state energy". At first glance recent studies seems to contradict the second theorem. Medvedev *et al.* calculated the electron densities and derivatives for wide range of contemporary functionals.[102] They showed that since appearance of meta-GGA formalism there is no further improvement in RMSD of electron density but mostly worsening. The worst description showed those functionals either developed before 1985 or those simply obtained by the tuning of tens of parameters with the aim of getting the best energetic and geometric description of a broad data set. The chemical space is vast and the chosen test performed on the electron densities of few atom-size model systems shows inherent inconsistencies. Despite this fact, these methods are nowadays more and more popular. It sets up a disturbing trend of turning away from the path toward the celebrated exact functional.

Nowadays, most of the efforts continue to improve the description of attractive long-range van der Waals interactions, which is well documented in the number of new DFT

functionals addressing this issue published each year.[103,104] The most popular approaches are introducing truly non-local density functionals, reparametrization of the current functionals, double hybrid functionals or long range corrected functionals. However, in our laboratory we choose to follow different approach, namely a posteriori calculated empirical correction term as is described later. We tested thoroughly several functionals and in terms of efficiency and accuracy the GGA-type functional BLYP shows great promise when combined with properly parametrized empirical dispersion term.

2.4.1 Inclusion of Dispersion by Atom-Atom Empirical Dispersion Correction

An efficient approach how to compensate the lack of dispersion energy is included here via *a posteriori* calculated empirical correction term. The most successful versions, Grimme's D and D3, are utilized both, because the more recent one D3 version is limited only to few TZ size and larger basis sets, whereas, the former one does not have this limitation. The dispersion can be described by the damped pairwise interatomic potentials of the form [105]:

$$E_{\text{disp}} = -\sum_{\text{AB}} \sum_{\text{n=6,8,10}} \frac{C_{n,AB}}{r_{AB}^{n}} f_{d,n}(r_{AB}) \quad , \qquad (12)$$

where the expansion is usually terminated at n=8, letters A and B are indexes of the atoms in the system, $C_{n,AB}$ are dispersion coefficients derived from atom polarizabilities for atom pair AB and r_{AB} is the respective interatomic distance. For description of dispersion between atoms at shorter distances, where part of the dispersion is covered by the correlation part of the DFT functional, the damping term $f_{d,n}(r_{AB})$ has to be used. In the D3 zero damping approach, it is defined as a function of cutoff radii for AB pair R_0^{AB}, radii scaling parameter $s_{R,n}$ ($s_{R,8}$=1), global scaling factor s_6 =1, steepness parameters α_6=14 and α_8=16 and is given by:

$$f_{d,n}(r_{AB}) = \frac{s_n}{a + 6(r_{AB}/(s_{R,n}R_0^{AB}))^{-\alpha_n}} \quad (13)$$

and

$$R_0^{AB} = \sqrt{\frac{C_{8,AB}}{C_{6,AB}}} \quad . \quad (14)$$

Later, a revised rational damping to finite values for small interatomic distances was developed by Becke and Johnson (BJ).[106-108] It enters the calculations as[109]:

$$f_{d,n}(r_{AB}) = \frac{s_n r_{AB}^n}{r_{AB}^n + (a_1 R_0^{AB} + a_2)^n} \quad . \quad (15)$$

Here a_1 and a_2 are free fit parameters. Note that the number of adjustable parameters is two (s_8 and $s_{R,6}$) and three (s_8, a_1, a_2) for zero and BJ damping, respectively. These parameters are functional and basis set dependent.

2.5 Semiempirical Quantum Mechanical Methods

Semiempirical QM methods (SQM), representing a compromise between accuracy and economy, can be applied to systems with thousands of atoms. Recently they have been successfully used even for whole proteins.[110-112] Although the SQM methods are considered to be a very promising tool for large-scale calculations of any biological systems, their description of noncovalent interactions is not satisfactory and clearly represents the limiting factor.[113] This is caused by its development which was for a long time focused solely on description covalent bonding and thermochemical properties of individual small molecules. The most crucial approximation that are being made are the inclusion of only valence orbitals and electrons, neglecting of differences between *s*- and *p*- type orbitals, setting the overlap matrix set to zero and parametrization of Hamiltonian matrix. This often leads to two main drawbacks. Firstly, the missing repulsion between atoms because core–core term usually underestimates the H–H non-bonded distances in aliphatic systems. Secondly, these methods are parametrized

against none or only limited number of noncovalent interactions, therefore their applicability here is limited. First SQM methods capturing at least qualitatively the geometry of hydrogen bonded complexes were Austin Model 1 (AM1) and Parametrized Model 3 (PM3). This was achieved by additional core–core term specifically parametrized on hydrogen bonded complexes. Further development was done in order to cover also dispersion interaction by empirical R^{-6}, R^{-8} and R^{-10} terms. However, even with inclusion correction for missing repulsion energy none of the SQM methods were accurate enough for quantitative description of hydrogen bonds and dispersion interactions in noncovalent complexes. Novel approaches in parametrization of SQM methods were tested, however they were only partly successful resulting in *e.g.* RM1 method.

The breakthrough in both accuracy and efficiency, when MOZYME algorithm is applied, was done with introduction of PM6 method.[114] It was shown to be the most accurate SQM method available despite the insufficient core–core Gaussian functions used to mimic correlation or van der Waals effects. In order to bring PM6 method to the border of the chemical accuracy of 1 kcal/mol several versions of *a posteriori* calculated corrections were introduced in years 2009-2011 in our laboratory.[57,110,111,115] They are denoted by D*x*H*y* suffixes where *x* and *y* stand for version of dispersion and hydrogen bonding correction, respectively. All versions are included in cuby framework (http://cuby.molecular.cz).[116] Recently, the most favorite form of the dispersion correction is derived from the D3 correction proposed by Grimme for DFT methods, see previous paragraph. The main difference from correction applied for DFT methods represents the inclusion of additional correction connected to specific errors in short distances for hydrogen–hydrogen distances caused by underestimated Pauli repulsion for hydrocarbons.[57] At first glance, the most advanced version of hydrogen bond correction, H4, consists of multiplication of several contributions: proton transfer term, radial and angular parts that are scaled by free parameter determining strength. Finally, there are two specific corrections for charged groups and water hydrogen bonds. Polynomials are usually used for both angular and radial parts and appropriate cutoff radius is chosen (about 5.5 Å). The actual functional forms are available in the original paper.[57]

In 2012 the newest version of PM7 was introduced where both corrections for hydrogen bonding (with functional form H+ version) and dispersion correction as formulated by Jurečka *et al.* were introduced *prior to* parametrization.[117] The performance of this method will be discussed later in section 3.1.

2.6 Molecular Mechanics

The molecular mechanics (MM) methods, also referred as "classical" or "force fields", are based on the Newton's laws of physics where the atoms are modelled as charged spheres. The MM energy is defined as a sum of various terms arising from bonding (stretch, bend and torsion) and non-bonding (van der Waals and Coulomb) terms. The performance and reliability of MM methods strongly depend on the parameters found in the functional forms (typically harmonic potentials) in equations of the energetic terms. These values are assigned once, fixed and for most of the systems available in the literature. They are parametrized to reproduce data such as experimental frequencies and *ab initio* energies of some sets of molecules, while atomic charges are derived from *ab initio* calculations combined with Restrained Electrostatic Potential (RESP) fitting.[118]

In this thesis, results obtained with Amber99SB-ILDN protein force field[119] combined with Amber94 nucleic-acid parameters will be discussed for protein-DNA bioinformatic analysis.[120,121] Additionally, the entropy analysis and second derivatives were performed with Amber11 at 298 K in order to estimate change of the Gibbs energy in host•guest systems.[119] In all cases the rigid rotor harmonic oscillator (RRHO) approximation was used. This approximation enables to uncouple the motions of molecules into translations, rotations and vibrations.[122] It makes calculations of molecular motions mathematically tractable even for large molecules.[123]

2.7 Fragment-Based Methods

One of the ways of executing calculations for the large systems without parametrization or sacrificing accuracy is to use some kind of fragment-based method. It generally includes performing of independent fractional calculations on portion of the system at a time and afterwards combining the results from the fragment calculations to predict the same properties as for the whole. The golden grail for a numerous of these approaches represents linear scaling coupled with the outcome accuracy approaching that which would be obtained for the full calculation. These techniques are being called „embarrassingly parallel“ and commonly take advantage of massively parallel computers where each separate fragment calculation is performed on a separate compute node.[124] The way how fragment-based methods are trying to achieve it differ in the fragmentation of the system and description of the local environment in each independent electronic structure calculation. There is no such thing as a “free lunch” therefore one has to be cautious when any kind of system fragmentation is being applied. Not only computational time but mainly the desired accuracy has to be met. Getting an answer quickly can become attractive enough to sacrifice the accuracy along the way therefore any application of the fragmentation scheme should be accompanied by accuracy assessment. One has to consider several tasks to be fulfilled: (1) energies has to be described on small to large systems evenly (2) different conformations has to be considered and described with similar accuracy, *e.g.* opened and closed conformation arrangement (3) several electronic structure methods have to be tested. (4) several model systems should be tested thoroughly.

Fragment-based methods in literature are divided into several groups. First class of methods is dealing with all orbitals in the full system. These methods are based on localized type orbitals often defining molecular orbital groups. Or alternatively, the methods are applying some kind of divide and conquer idea where from the fragment densities an *ad hoc* molecular density is being constructed and afterwards single energy calculation follow.[125-129] The interaction between fragments is readily introduced through modification of the Fock operator by the Coulomb potential of a partner, often doing it self consistently until simple monopole representation of electrostatic potential of the rest of the system is converged. This takes into account both electrostatic and polarization effects. An important general development has been been made when

strong orthogonality between fragment wave-functions was introduced. This made construction of block Fock matrix possible without the need of diagonalization of full Fock matrix. Further discussion of this class of approaches is however beyond the scope of our work and can be found reviewed elsewhere.[130]

Second large class of fragment-based methods can be derived from the Multibody Expansion (MBE) or cluster expansion method. It has been developed in the field of solid-state chemistry in order to calculate the total energy as a linear combination of the energies calculated on the atom clusters using N-order interaction potentials. It proved to be a general concept in computational strategies of total energy evaluation from subsystem energies.[131] Individual methods then are distinguished from each other by the way how they generate overlapping or disjoint fragments, write down the order of truncation, neglect the non-overlapping fragment and non-chemically bonded interactions. Few such examples are molecular fractionation with conjugate caps (MFCC) method,[132,133] fragment energy method (FEM)[131] and kernel energy method (KEM) utilized by us here for host•guest project (see later in section 3.5).[134] The pairwise additive KEM total energy can be calculated from fragments (smaller kernels of atoms):

$$E_M = \sum_{i=1}^{M} E_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} (E_{ij} - E_i - E_j) \qquad (15)$$

or in case of summation from all double kernels reduced by those of any single kernels that have been overcounted in the sum over double kernels:[135]

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M-2) \sum_{i=1}^{M} E_i \quad , \qquad (16)$$

where M is total number of fragments, E_i is the energy of the system composed of the ith fragment, E_{ij} is the energy of the system composed of the i, j fragments *etc*. We were forced to apply it for MP3 calculation, because the system size exceeded 180 atoms.

2.8 Energy Decomposition Analysis

Noncovalent interactions are governed by various physical/chemical phenomena that are however hard to determine by experimental techniques and no QM operators exist that would enable to compute any energy or interaction energy components.[136] However, there are possibilities how to tackle this problem that will be described herein. Recently, they have been utilized even for whole proteins using fragmentation or QM/MM scheme. Hirao calculated the protein environment effect on an intermediate compound I of cytochrome P450cam. Author was successful in decomposing it into driving forces showing the electrostatic effect being the largest in the magnitude.[137]

There are two general approaches how to decompose the interaction energies. First approach is by means of perturbation theory, where the interaction between monomers is treated as a perturbation between two previously noninteracting monomer systems. This is referred as symmetry adapted perturbation theory (SAPT) scheme.[63] Second approach is used for variational calculations of the complex where it is defined by a stepwise evaluation and relaxation of the separate contributions for the complex starting from a state composed from noninteracting monomers – this is called energy decomposition analysis (EDA).[138] Both these approaches are nowadays being developed for quantifying various effects playing role in noncovalent interactions.

Schemes implemented at the higher levels than HF or DFT, *e.g.* at the coupled cluster (CC) level of theory, are offering the possibility to highly accurately describe the studied system however also significantly restricts the affordable size of the systems under study. Additionally, there is discussion in the current literature about ill-defined polarization and charge transfer at close intermolecular distance that can be considered as polarized electron density of one molecule extended into the space occupied by the other molecule where it is hard to separate it from the remaining polarization.[138] This is especially true when large basis set is utilized.

Here, the interaction energy EDA method was readily utilized for the partitioning of the DFT interaction energy into the electrostatic interaction energy ($\Delta E_{electro}$), the exchange-repulsion ($\Delta E_{exch\text{-}rep}$), the orbital relaxation energy ($\Delta E_{orb\text{-}rel}$), the correlation interaction (ΔE_{corr}) and finally the Grimme's D3 empirical dispersion interaction energy (ΔE_{disp}):

$$\Delta E = \Delta E_{electro} + \Delta E_{exch-rep} + \Delta E_{orb-rel} + \Delta E_{corr} + \Delta E_{disp} \quad . \qquad (17)$$

2.9 Solvation

The solvation effects or interactions of the environment with solute can be described as the either 'specific' or 'nonspecific'. By the specific we mean the directional noncovalent interactions between participating molecules in the system. Another important component of solvation could be called 'nonspecific' since environment simply attenuates the electrostatic interaction in our system when we compare it to situation in *vacuo*. On the other hand solvent environment affects also other interactions (*e.g.* dispersion interaction) however to a considerably less extent than electrostatic energy. Therefore, one has to be vigilant when comparing magnitudes of different contributions described in previous section. The attraction from dispersion energy, although usually being small, could be in *e.g.* water environment comparable, or even higher than from the electrostatic energy. This can easily result in large change of the physical nature of the binding when different environment is considered. In the field of computational chemistry two approaches how to describe the effect of solvent on the solute are routinely being used.

First, the popular in applications using more approximate methods, is inclusion of the explicit water molecules. This is the most robust and general approach, how to account for solvation requiring long equilibration or geometry optimization of the system especially when the positions of the water molecules are unknown. This drastically increases the complexity of the calculation because of both the harder sampling of the potential energy surface and increase of the system size.

On the contrary, the second approach, implicit solvation model, is more approximate and efficient because it depends only on the coordinates of the solute itself. Therefore, it leaves out the problems with equilibration of the system caused by addition of large number of water molecules and its connected number of degrees of freedom in the system. Additionally, in order to achieve converged solute-solvent interactions one does not have to calculate relatively long times for achieving a reasonable statistical certainty but the solvation free energy is updated instantaneously. This can be proved

to be useful in studies of long processes by molecular dynamics. All effects introduced by solvent (entropy cost for cavity formation, polar screening of solute charges, van der Waals interactions *etc*.) are typically introduced by suitable functions resulting in:

$$\Delta G_{solv}(r^{\vec{M}}) = \Delta G_{cav}(r^{\vec{M}}) + \Delta G_{hydroph}(r^{\vec{M}}) + \Delta H_{elstat}(r^{\vec{M}}) \quad , \quad (18)$$

where first two penalty terms, cavity formation and hydrophobicity, have both the enthalpic and entropic character. While the third term is purely enthalpic and represents the shielding of electrostatic interactions by the polarized solvent.

Poisson equation (eq. 19) provides the foundation of the description of electrostatic interactions in the field of continuous electric dielectricum and also all implicit solvent models such as Conductor-like Screening Model (COSMO) or density-based solvation model (SMD). The main difference between these is that the later method calculates free energy using not partial atomic charges but rather full solute electron density with noticeably increased computational expense. The electrostatic potential $\Phi(\vec{r})$ can be described as:

$$\Delta\Phi(\vec{r}) = -4\pi \frac{\rho(\vec{r})}{\varepsilon} \quad , \quad (19)$$

where $\rho(\vec{r})$ is charge distribution and ε is uniform dielectric constant. The choice of dielectric constant is an open question as it is impossible to derive from experiment under most conditions. It can easily differ in the cavity of the molecule and on its surface and its true value is still an ongoing debate in the literature.[139,140] Second severe downfall of implicit solvent models is the lack of description of the direct hydrogen bonding mediated via water hydrogen network system.

The hybrid models utilizing both approaches in the same time were still not proven to be superior to any of the two approaches but are commonly accepted because of the both efficiency and introduction of the direct interacting water molecules.[141-143] The direct hydrogen bonds were readily proved to be necessary in order to adequately describe the energetic or the geometrical aspects of the systems. But how do we determine the positions of the water molecules? One favorite approach is the inclusion

of the first or also second solvation shell - it is layer of the water molecules near or in the „direct contact“ with the solute. This is still not affordable for some applications. Second approach is run a molecular dynamics of the water molecules in the presence of the frozen solute and determine the important water molecules from the clustering procedure, *i.e.* determination of water sites with high occupation number directly related to the frequency of water molecule being found in the given position. In this way one can determine the most important water molecules, incorporate them into the calculation prior to the system equilibration. Such procedure has been automated *e.g.* in the Schrödinger package in the WaterMap module.[144-146]

3 PROJECTS

Projects are organized as follows. First, we discuss the PM7 performance and parametrization of D3 correction for various basis set and functional combinations on a number of benchmark data sets of noncovalent interactions (attachment A+B). Next, we verify the bioinformatic findings and accuracy of *ab initio* and MM methods for protein•DNA interactions using available crystal geometries (attachment C+D). Finally, we describe two projects; in one of them quantum chemistry tools helped to determine the leading interaction motifs governing interplay between larger organic molecules (attachment E) and in the second one we studied and designed the host•guest binding (attachment F+G+H).

3.1 Performance of the Semiempirical Quantum Mechanical PM6 and PM7 Methods

The SQM methods are considered a promising tool for large-scale calculations in both material science and applications on biological systems. Here, we compared the performance of recently released semiempirical method PM7 with its predecessor, PM6 with post-SCF (DH+, DH2 and D3H4X) corrections for various types of noncovalent interactions. These corrections were included during development of the PM7 method therefore the respective parameters were adjusted along with remaining parameters of the PM7 method. For these reasons a more balanced description of different interactions types can be expected and any double counting should be automatically avoided. It has been shown that PM7 method considerably improved the description of such properties as the heat of formation or the height of the reaction barriers for reactions.[147] However, its description of noncovalent interactions was not thoroughly tested up to date and

usually represents the limiting factor for the most of the contemporary SQM methods. For these reasons, we utilized various benchmark data sets of interaction energies and geometries available in the literature.

3.1.1 Computational Details

Results of PM7 and PM6 methods augmented with empirical corrections were validated on a total of 13 data sets listed below while their brief description can be found in the method section 1.2.1. Eleven data sets contain geometries and interaction energies of molecular complexes. The two remaining data sets consist of the structures and relative energies of different conformations of complexes and complex molecules: *Wat5*[65] and *PCONF*,[66] whereas the reference served the global minima the puckered ring and FGG.99.

The two last data sets, *Wat5* and *PCONF*, and 4 others are the most important for the validation, because they were utilized in the parametrization of neither the PM7, PM6 nor any post-SCF corrections: the *Goddard water* set[64] (only electroneutral clusters of water molecules were considered), *AA-sidechains*,[4] *1hsg*[58] and *L7*[53]. The seven remaining data sets; the *S22*,[60] *S22x5*,[61] *S66* and *S66x8*,[55] *Charged HB*,[57] *Halogens* and *Halogensx10*[62] were used for the parametrization of the new PM7 method as well as the DH+ and D3H4 corrections. Or more specifically, DH+ and DH2 used a *S22* data set for parametrization; the same is true for *S66* in the case of D3H4 and PM7; *Charged HB* for D3H4; *Halogens*(x10) for halogen corrections (DH2X and D3H4X) and for the PM7 method. These data sets were included here for their useful system division according to a dominant stabilization term based on DFT-SAPT analysis.[148]

Most of these data sets were published with the well-constructed benchmark CCSD(T)/CBS energies. The first exception is the *L7* set, for which Sedlák *et al.* published QCISD(T)/CBS energies calculated in a similar way to CCSD(T)/CBS estimates with the inclusion of the ΔQCISD(T) correction term instead of the CCSD(T) one. For more details, go to the original paper.[53] Second exception represents the *Wat5* data set. We calculated CCSD(T)/CBS estimates on the previously published structures according to the scheme mentioned above (eq. 11) with aug-cc-pVDZ basis set for the correction term and aug-cc-pVTZ and aug-cc-pVQZ basis sets for $\Delta E^{\mathrm{MP2corr/CBS}}$ and

$\Delta E^{HF/CBS}$ extrapolations according to Helgaker.[75] Benchmark interaction energies were corrected for basis set superposition error (BSSE) using the counterpoise scheme of Boys and Bernardi.[149] The PM6 and PM7 calculations were performed using the MOPAC2012 program[150] and all benchmark calculations were performed using the Molpro2012 program.[151]

3.1.2 Results and Discussion

Table 1 summarizes a statistical analysis of the PM7, PM6, PM6-DH+, PM6-DH2(X) and PM6-D3H4(X) methods calculated on various data sets. First, we will analyze results of the following data sets (used for parametrization of SQM methods): *S22*(x5), *S66*(x8), *Halogens*(x10) and *Charged HB*. For these data sets the PM7 method showed substantial improvement over its older version PM6. Additionally, it delivers results of the similar accuracy as the three modified PM6 methods. These results provide a good clue on how well were the methods parametrization handled. Because there are achieved similar error magnitudes with PM7 as for other methods it indicates that the parametrization of the correction terms analogous to the ones added to PM6 was made properly. Further, it is evident that all but the methods with the specific corrections for halogen bond failed to describe systems containing this specific noncovalent interaction (*Halogens* and *Halogensx10*).

Next, all the SQM methods have problems with the relative energies of the molecules and complexes included in the *PCONF* and *Wat5* sets. This is clearly reflected in a generally much higher RMSE for these data sets when compared to any other in the list. Figure 1 shows how all SQM methods fail to assign the order of the stabilities of both, water pentamer and FGG tripeptide conformations. Among the methods tested, the DH2 and D3H4 approaches show the best performance. The very poor performance is slightly surprising especially in the case of DH+ correction due to the fact that this particular version has been parametrized to reproduce the cooperativity effects in water clusters. Additionally, overall performance for the rest of the data sets is slightly worse when compared to the DH2 and D3H4 corrections (see Table 1). The 3rd generation correction, DH+, evidently lacks the accuracy of the DH2 and D3H4 versions.

	PM6	PM6-DH+	PM6-DH2(X)	PM6-D3H4(X)	PM7
S22	4.18	0.78	**0.53**	0.8	0.85
S22x5	3.25	0.84	**0.61**	0.88	1.16
S66	3.07	0.84	0.94	**0.68**	1
S66x8	2.49	0.76	0.79	**0.66**	0.98
Charged HB	4.56	2.18	2.51	**1.65**	1.85
Halogens	2.8	2.56	**2.60 (2.32)**	**2.60 (2.32)**	3
Halogensx10	4.15	4.18	4.19 (3.65)	4.20 (3.29)	**3.4**
Wat5	9.38	7.15	5.5	**5.43**	6.71
PCONF	**2.69**	2.74	3.03	3.15	2.98
Goddard water	27.47	26.47	**4.17**	4.54	12.4
AA-sidechains	4.08	1.89	1.32	**1.17**	1.49
1hsg	2.09	1.37	1.13	**0.72**	1.4
L7	15.79	**3.22**	3.35	3.93	4.93

Table 1. Comparison of SQM methods with benchmark data. The numbers listed here are RMSE for the respective data sets and the results were colored according to the relative difference between the method's RMSE and 'the best' SQM method tested here.

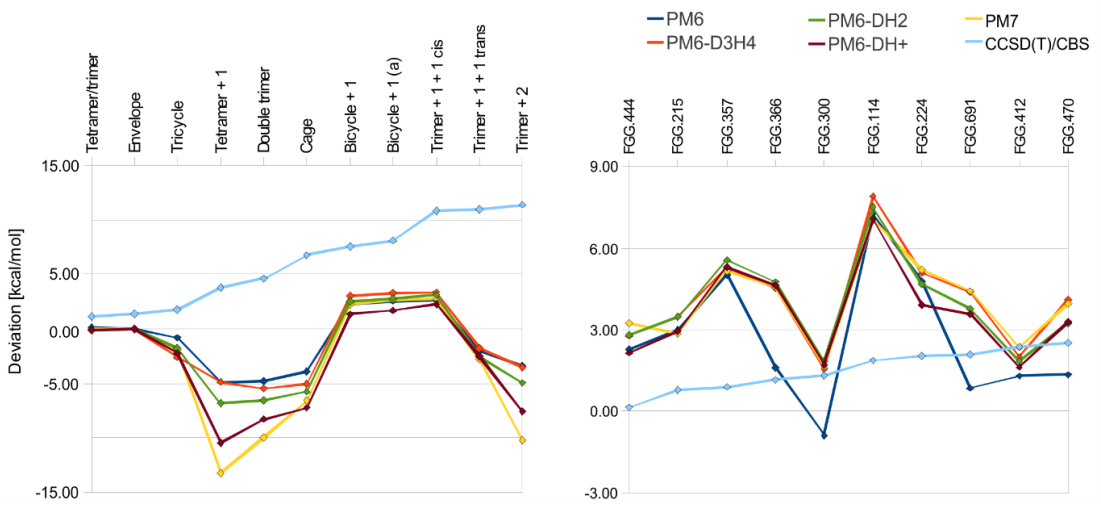


Figure 1. The relative stabilities of the conformers of water pentamer (*Wat5*) and the FGG tripeptide (*PCONF*). As the reference the global minima have been chosen: the puckered ring and FGG.99.

The poor stability order estimations of water pentamers has lead us to investigate systems of water clusters in more detail. The relative deviations of interaction energies for the *Goddard water* data set as a function of the number of interacting molecules are summarized in the Figure 2. PM6 method strongly underestimates the interaction energy showing the urgent need of hydrogen bonding correction. On the contrary, PM7 and PM6-DH+ methods strongly overestimate it indicating that they are not well suited for this class of systems. Finally, DH2 and D3H4 approaches again systematically provide the best results. The very good performance of the D3H4 correction was achieved not only for the neutral hydrogen bonds from the *Goddard water* set but it was followed by comparably accurate results in systems with charged hydrogen bonds (*Charged HB* set in Table 1). It originates from the inclusion of specific parameters for charged hydrogen bonds (parametrized on this set).

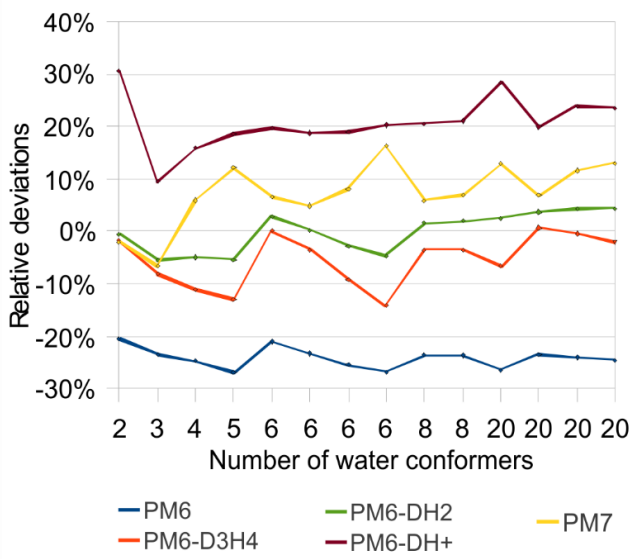


Figure 2. The relative deviations of interaction energies for the *Goddard water* data set.

Next, the results for the two biological data sets, *AA-sidechains* and *1hsg*, will be described. The aforementioned data set consists of representative structures of amino acid side chains, while the second one includes fragmented parts of the HIV-II protease

– ligand (indinavir) complex. In both cases the D3H4 approach provided the best results and reached almost chemical accuracy (~1 kcal/mol).

The *L7* data set consisting mostly of large dispersion-bound systems (of a size ranging from 48 up to 121 atoms) provides a good clue on how the methods studied would work in larger complexes. We found that PM6 failed, while all the other approaches provide reasonable agreement with the benchmark results. Important method comparison represents the RMSE, plotted as a percentage of the average stabilization energy (rRMSE), for the *L7* data set and three datasets of smaller complexes (*1hsg*, *S22*, *AA-sidechains*) in Figure 3. The rRMSEs increased roughly two-times for DH+, while for all other methods the increase was fourfold.

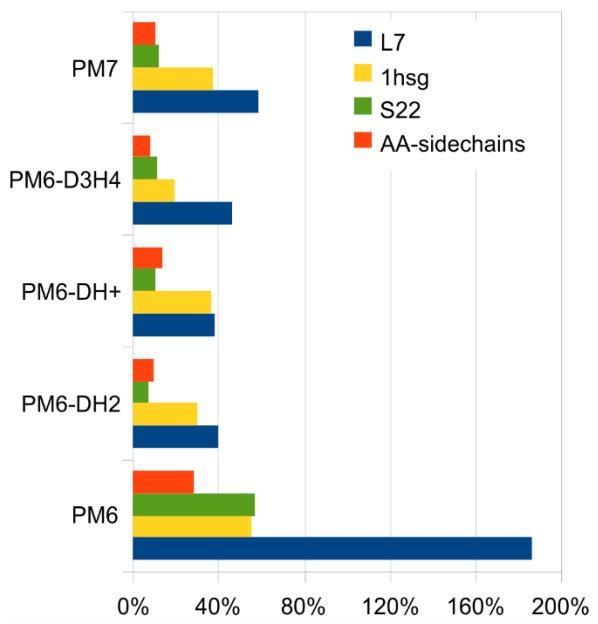


Figure 3. The rRMSE for L7 and a variety of medium-sized data sets.

Several data sets [*S66(x8)*, *S22(x5)* and *Halogens(x10)*] are split into different classes with a common interaction motifs: hydrogen bond, dispersion interaction (π-π, π-aliphatic, aliphatic-aliphatic), halogen bond (X-O,N,S,π) and mixed. The RMSE were plotted as a percentage of the average interaction energy for all the SQM methods under study and summarized in Figure 4. The poor description of PM7 method for dispersion bound complexes is surprising considering the fact this method includes similar

correction for dispersion interaction as the other methods tested with the exception of PM6. The most striking results were obtained in the case of π-π dispersion in *Halogensx10* and *S66x8* sets, where the RMSE is twice as large for PM7 than for PM6-D3H4. The halogen bonded complexes are comparably well described in both PM7 and PM6-D3H4X methods. Great improvement was observed for the X-π interaction, where the RMSE of PM7 is half of the one of corrected PM6.

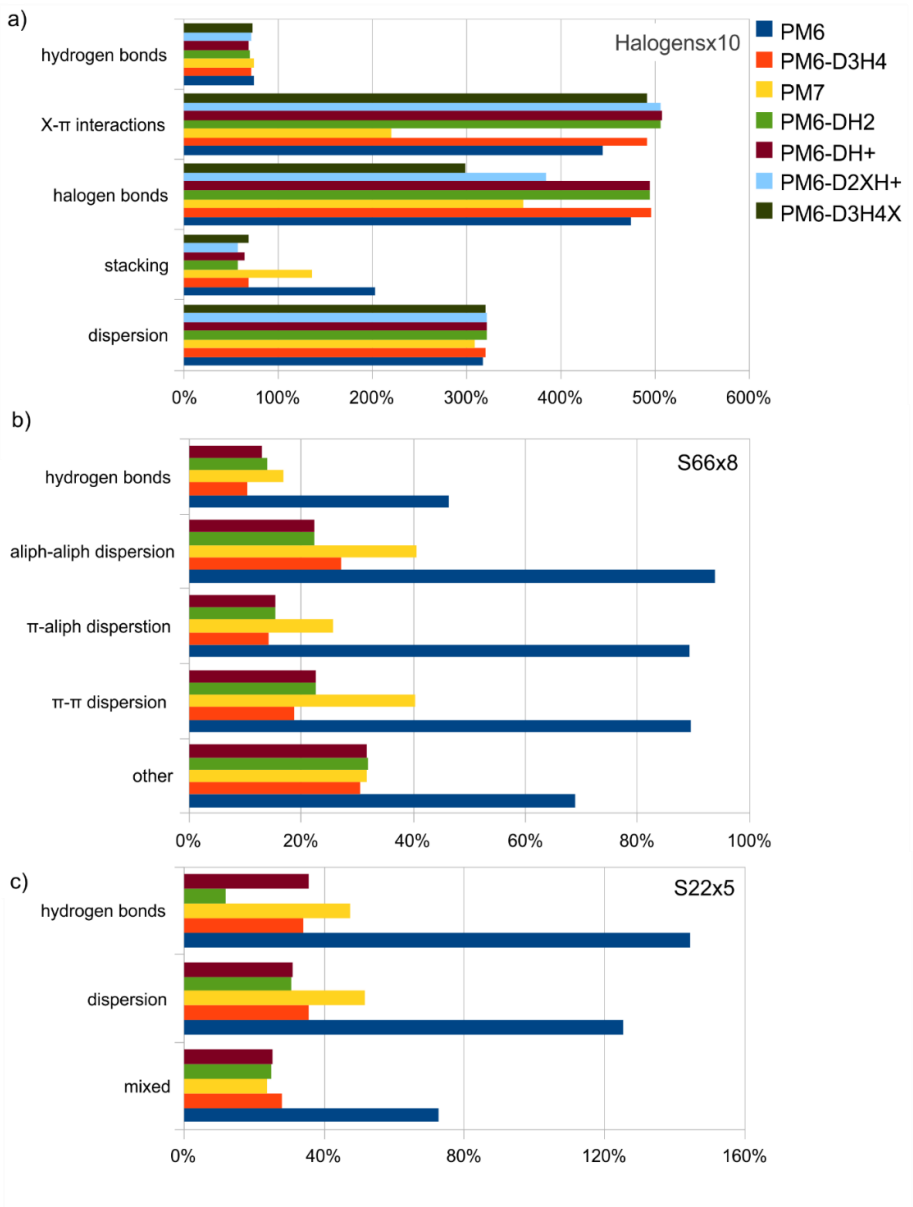


Figure 4. The rRMSE plotted for groups of different interaction types for *Halogensx10*, *S66x8* and *S22x5* data sets.

3.1.2.1 Geometry Optimization

In previous section we focused on the energetics of noncovalent interactions. However, additional information characterizing the PES around the equilibrium geometry is necessary to demonstrate the accuracy of SQM methods. Several data sets *S66x8*, *S22x5* and *Halogensx10* data sets contain besides the potential energy minimum also additional points along the dissociation curves. However, these several points provide only limited knowledge about the accuracy of the whole PES. From the full geometry optimization more information on the relation between geometry and energy changes can be gathered. Additionally, it is important in those cases when the structure is not known and geometry optimization is a must.

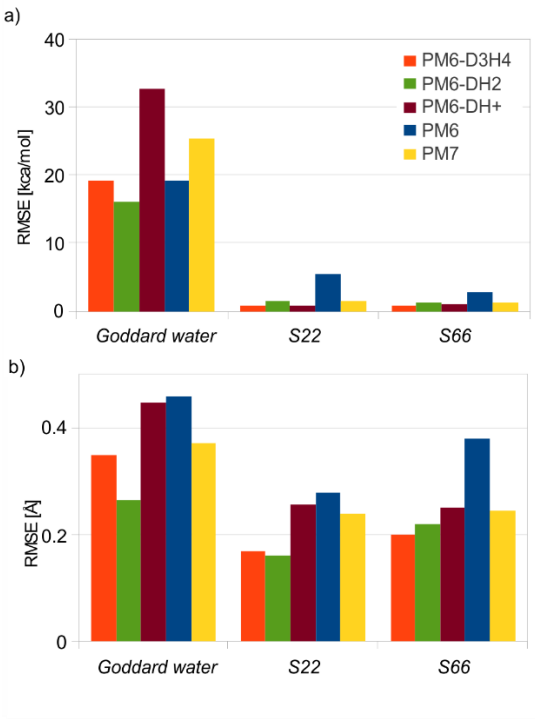


Figure 5. The RMSE for the energetics (a) and geometries (b) for *Goddard water* complexes, *S22* and *S66*, optimized and calculated by the SQM methods.

Bottleneck of the most of the SQM and MM methods represent the accurate predictions of hydrogen bond networks. In order to demonstrate it we investigated neutral water clusters from the *Goddard water* as well as all structures in the *S22* and *S66* sets. Results are presented in Figure 5 and 6. The RMSE (in Å) of the optimized structures is rather

high for all SQM methods and no significant differences have been found. However, interaction energy of these optimized structures represents a more sensitive measure. Inspecting Figure 6a, we find that the RMSE of the interaction energy is surprisingly large, exceeding 15 kcal/mol in all cases. The relative deviations of water cluster interaction energies are depicted in Figure 7. We found that PM7, PM6 and PM6-DH+ methods provide inaccurate results. A slightly better results are given by DH2 and D3H4 approaches. From visual comparison we must conclude that all of the SQM methods investigated failed seriously in reproducing the structure of water clusters. Evidently, the SQM methods investigated are not well suited for water simulations.

The situation is very different in the case of binary complexes from *S22* and *S66* data sets. Both, the structures and interaction energies for the optimized dimers differ only slightly from the benchmark results. For the *S66* data set the PM6-DH+ and PM6-D3H4 methods exhibit the lowest value (0.7 and 1.0 kcal/mol) and PM7 and PM6-DH2 yield only slightly worse numbers.

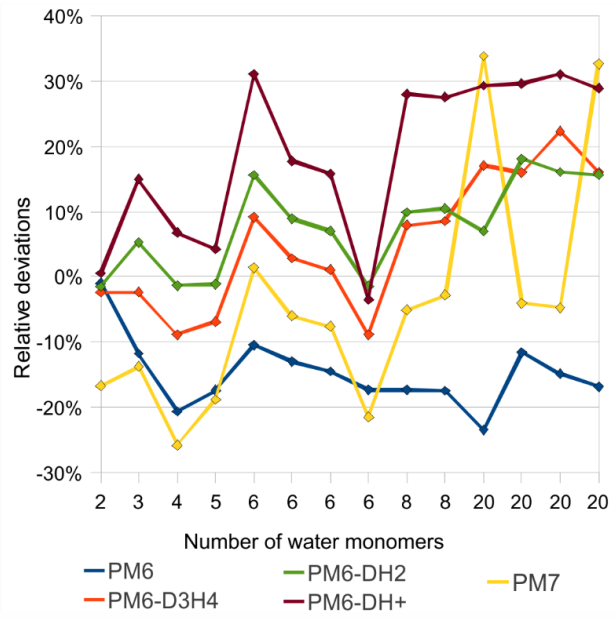


Figure 6. The relative deviations of interaction energies for the *Goddard water* data set after optimization.

Additionally, inspecting Figure 5a we can see that the best performing methods for *S22* data sets are PM6-D3H4 and PM6-DH2 with the RMSE under 0.2 Å. However, the second method, PM6-DH2, has to be performed with numerical gradient optimization due to missing the analytical gradients and cannot be therefore recommended for routine geometry optimizations.

3.1.3 Conclusion

We can summarize this study by stating that none of the methods tested showed a clear dominance and can thus be unambiguously recommended to the use in the field of noncovalent interactions. Although PM6-D3H4(X) and PM6-DH2(X) are slightly superior to the others in some cases, it is impossible to generalize this statement for the whole range of noncovalent interactions investigated. It should be taken into consideration that the modified PM6 methods are designed exclusively for noncovalent interactions while PM7 showed a remarkable improvement over many properties of isolated systems (not studied here). However, relatively large errors of PM7 for the interaction energy calculations of dispersion bound complexes and water clusters are a limiting factor in the possible applications. For other types of binary complexes in the 13 data sets, PM7 yields only slightly worse results as compared to PM6 with the inclusion of the newest post-SCF correction for dispersion, hydrogen and halogen bonds (PM6-D3H4X). Nevertheless, due to the improvements in the description of the properties of isolated molecules reported in Stewart's original paper, PM7 constitute the most robust tool among the semiempirical methods.[147]

Finally, when one wants to achieve accuracy of 1 kcal/mol and better across the wide range of exotic noncovalent interactions, *e.g.* other than hydrogen bonds and dispersion dominated, the less approximate methods need to be chosen. One of such examples is DFT methodology accompanied with range of post-HF methods used in the rest of the projects here.

3.2 Accurate DFT-D3 Calculations in Small Basis Set

The key idea to express the electronic energy as a functional of the electron density led to methods called Density Functional Theory (DFT). The substantial increase in the accuracy of both energetic and geometric descriptions in the last decades caused development of DFT to focus toward more sophisticated applications in *e.g.* energy decomposition, QM/MM and vibrational analysis.[11,152,153] Affordable scaling enabled its use for assemblies up to few hundreds of atoms opening space for increasing number of these applications in the areas of catalysis, supramolecular chemistry, molecular biology or biotechnology.[154] Shortly, it has been recognized that the accurate description of the London dispersion is of crucial importance.[155]

The improvement of the description of attractive long-range van der Waals interactions is in the center of interest of many researchers that is well reflected in the number of new DFT functionals addressing this issue published each year.[103,104] We chose to follow an alternative approach to compensate the lack of dispersion energy, namely *a posteriori* calculated empirical correction term decribed in previous section 2.4.1.[105] The most successful version, Grimme's D3, was up to now limited to triple-zeta (TZ) and larger basis sets because of the severe BSSE native to smaller basis sets.[109] For the treatment of large systems, it is highly desirable to use a small basis set if the error can be kept reasonably small. To find the best setup, we have searched a library of basis sets ranging from a minimal basis set to split valence (SV) and double-zeta (DZ) basis sets, selected those with the smallest BSSE and parameterised the Grimme's D3 correction (see section 2.3.1) for them.[105]

3.2.1 Computational Details

We started by building a library of DFT interaction energies calculated for the S66 data set[11,55,57] using multiple common DFT functionals: BLYP[156], B97-D[157], PBE[158] (GGA functionals), B3LYP[159], PBE0[160] (hybrid functionals) and TPSS[161] (meta-GGA) in the following basis sets: STO-3G, MINI, MINI3, MINIS, MINI3S (minimal basis), 6-31G, 6-31G*, 6-31G**, SV, SV(P), SVP, def2-SV(P), def2-SVP, def2-SVPD, MIDI, MIDIX, DZ, DZP, DZVP, DZVP-DFT, cc-pVDZ and aug-cc-pVDZ (split-valence and

double-zeta).[81,82,162-168] We calculated average BSSE (over *S66x8*[55] data set) of these combinations and interestingly DZVP-DFT showed superior results to other basis sets of comparable size.[164] Next, we chose Generalized Gradient Approximation (GGA) functionals BLYP[156] and PBE,[158] meta-GGA TPSS[161] and hybrids PBE0[160] and B3LYP[159] combined with def2-TZVP,[167] DZVP-DFT,[164] def2-SVP[167] and 6-31G*[81,82] basis sets for further studies. Both the standard zero and BJ damping were tested for dispersion correction.[106-108] Parametrization of dispersion was exclusively performed on the subset of the *S66x8* data set, *i.e.* geometries and interaction energies of dispersion bound complexes at points along dissociation curves.[11,57] Note that we excluded systems containing a hydrogen bond from the parametrization. In these, the BSSE is largest what would lead to underestimated dispersion in all the other systems. We chose RMSE as the only optimized criterion for parameter search. The Cuby framework (http://cuby4.molecular.cz) was used to automate the calculations.[116] In some combinations of functional and basis set, the parameters of the BJ damping function reach unphysical values resulting in positive dispersion energy balancing the overstabilization caused by the large BSSE. These combinations were omitted from the discussion. The final parameters will be included in the new version of Cuby package and in original publication.[116]

3.2.2 Results and Discussion

3.2.2.1 Validation for Interaction Energies

The transferability and the quality of parametrization were tested on data sets *X40* and *L7*.[53,62] The former data set plays a prominent role in the validation, because it contains halogen elements not present in the training set. The *L7* data set represents mostly dispersion-bound complexes of larger size and is then a step between small model systems and biologically relevant systems. We utilized RMSE of interaction energies and RMSD of the geometry as the most important descriptive statistic criterions for ranking the methods. Additionally, for purpose of deeper analysis we calculated signed and unsigned mean error indicators.

Here, results are summarized in a plot of RMSE (Figure 7). The combinations of basis sets and functionals are ordered by their complexity. Generally, the BJ damping proved to be superior over zero damping although this behavior is strongly system, basis set and functional dependent.

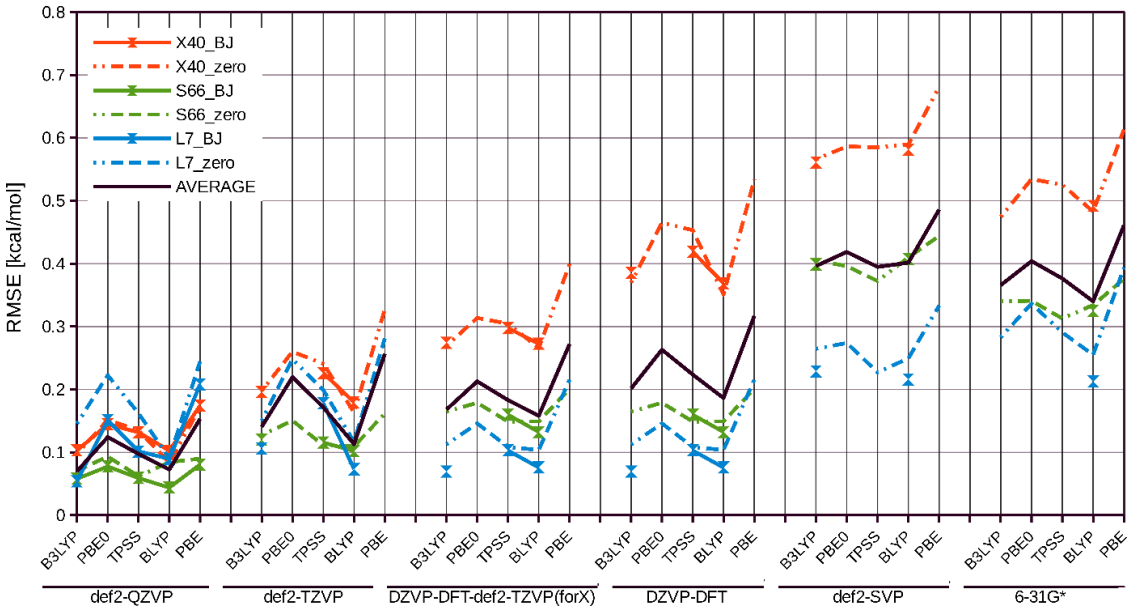


Figure 7. RMSE for the three data sets (*X40*, *S66* and *L7*) and two damping functions (zero and BJ). DZVP-DFT-def2-TZVP(forX) stands for mixed basis set where def2-TZVP is used for halogen atoms and DZVP-DFT for the rest.

Clearly, there is no simple increase of accuracy with the complexity of the functional from right to the left. The BLYP and B3LYP functionals yielded in average more accurate results than PBE, TPSS and PBE0 while the BLYP has lesser computational demands when compared to hybrid B3LYP functional. Secondly, we found large diversity originating from the different diffuse character of basis sets (*i.e.* various exponential values of Gaussian *d*-polarization function) and magnitude of BSSE. This is most profound in the case of more diffuse DZVP-DFT basis set having average RMSE 0.186 kcal/mol delivering superior results of almost TZ size basis set quality. It is thus the first choice we can recommend for practical calculations. Closer inspection of our results show that the cases limiting the accuracy of the DZVP-DFT calculations

are the halogenated molecules in the X40. It can be, however, improved by using larger basis set (def2-TZVP) only for these atoms (Cl, Br, I) at negligible additional cost.

3.2.2.2 Validation for Geometries

In the second step, we analyze influence of the basis set size on the quality of geometries. Using smaller basis set size for geometry optimization prior to interaction energy calculations is a common practice in many studies. This is based on the assumption that geometries are less sensitive to the basis set size. Because we are interested only in the effect of the basis set size in DFT calculations, we take geometries optimized with BLYP-D3/def2-QZVP as a benchmark. It is used for validation of medium and small size basis sets both combined with BLYP functional and D3 empirical dispersion. Table 2 summarizes the statistical analysis of geometry optimizations performed with def2-TZVP, DZVP-DFT and def2-SVP basis set. Clearly, results strongly depend on quality of basis set. For example when passing within the same basis set family from triple-zeta (def2-TZVP) to double-zeta (def2-SVP) size basis set the RMSD increased seven times. On the contrary, DZVP-DFT basis set yielded good results delivering three times smaller RMSD when compared to basis set of similar size (def2-SVP). This demonstrates the potential of DZVP-DFT basis set to be useful not only for calculations of energetics but even more advantageous for geometry optimizations.

	def2-TZVP	DZVP-DFT	def2-SVP
RMSD	0.01	0.023	0.07

Shortest Distance (average in *S66* data set: 2.45Å):

MSE	-0.011	-0.018	-0.044
MUE	0.011	0.029	0.054
RMSD	0.017	0.038	0.07

Table 2. Average RMSD and additional statistical analysis of the shortest distances between monomer geometries in the complexes.

Finally, we discuss the computational demands of the calculations. Table 3 shows the computational times for systems increasing in size for given basis set size and rationalizing our focus on DZ size basis sets. All the DFT calculations were performed using Turbomole 7.0 package.[169]

Basis set / system size	CB[5]=90 atoms	CB[8]=144 atoms
6-31G*	5.8	6.8
def2-SVP	6.4	8
DZVP-DFT	7.7	9.3
def2-TZVP	25	31.1
def2-QZVP	108.5	161.2

Table 3. Computational times for two cucurbit[n]uril systems where n=5,8 with 90 and 144 atoms, respectively. We used BLYP functional and one computational node: Intel Xeon E5630 2.53 GHz, 8 cores, 5.8 GB RAM per core.

3.2.3 Conclusions

We have shown that DFT method combined with DZ size basis set can yield accurate results for noncovalent interactions. Among the tested combinations of functionals and basis sets we recommend DZVP-DFT basis set and BLYP functional for both geometry optimizations and energy calculations in large systems. The DZVP-DFT basis set was already used successfully in studies of brominated carborane cages.[170] It is important to note that even def2-SVP or 6-31G* basis sets combined with any tested functional clearly surpassed SQM methods tested in previous studies.

3.3 Large-Scale Quantitative Assessment of Binding Preferences in Protein−Nucleic Acid Complexes

New technologies, methods and advances in the field of crystallography, molecular biology, biochemistry, computational analysis and crystallization enabled enormous growth of wealth of protein•DNA structure data. It allows a question whether there is sufficient enough information for studies of universal rules governing the DNA sequence recognition process. Understanding of these rules would be a major leap forward to understanding of process such as DNA replication, gene expression, DNA repair and cycle regulation.

Recently, the relative abundance of various modes of amino acid−base contacts (dispersion bound, hydrogen bonded) was studied, however such information regarding the accurate energetics and types of noncovalent interactions in protein•DNA complexes is still missing.[171] These interactions typically account for from few up to several tens of kcal/mol in *vacuo*. Therefore the quantitative assessment of binding preferences in protein•DNA complexes require results that would be validated by a highly accurate *ab initio* method. There were up to now available no specialized data sets for such purpose.

For these reasons we have quantitatively examined the protein•DNA interactions by comparing the interaction energies for all 20 × 4 amino acid−DNA base pair combinations. We analyzed all available protein−DNA complexes and therefore we could have drawn conclusions that are not limited to any single DNA binding motif or protein family. Additionally, this opened a chance to further study the performance of computational techniques used nowadays.

3.3.1 Computational Details

In total 50,205 nucleotide−amino acid pairs were gathered from the Protein−DNA interaction atlas generated according to the method described by Luscombe *et al.* and updated as of March 2014.[172] The atlas was assembled from 1,569 unique structures of protein–DNA structures. Amino acid–nucleotide pairs were extracted by the procedure similar to the SIRIUS set of scripts introduced by Singh and Thornton.[173] According to

certain distance criteria (here 4.5 Å) these programs pick out the pairs of residues and mark them as interacting. The procedure resulted in 20 x 4 sets of contacts. In each of these sets comprising of a unique pair of a single amino acid with a single DNA nucleotide, all dimers were transformed to occupy the same frame of reference of the DNA base. We therefore generated 20 distributions of amino acid residues around each of the DNA bases (see Figure 8). In all the distributions, the RMSD between atom positions was calculated for all pairs of amino acid side chains. The dimer with the highest number of structures within the RMSD of 1.5 Å was removed together with all these "neighboring" structures and denoted as a cluster representative. The contacts removed with cluster representative were regarded as corresponding cluster. This procedure was repeated up to six times depending on the size of the cluster.

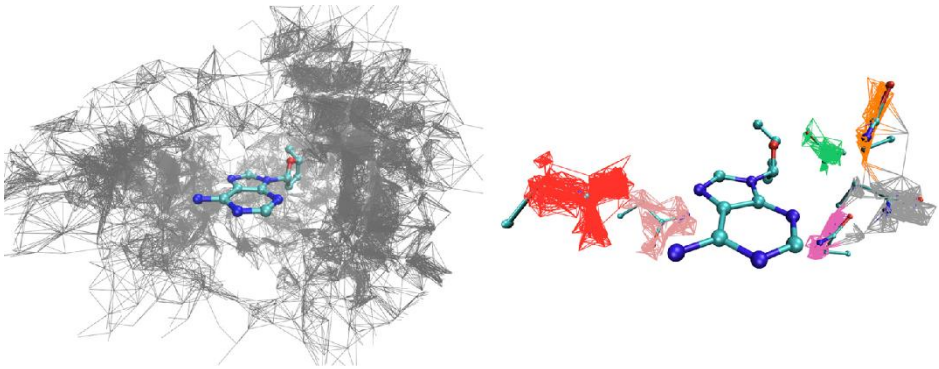


Figure 8. Ade-Asn distribution and all identified clusters. The ball-and-stick representations of cluster representatives are depicted at the right hand side.

Only those with the distance of less than 4.5 Å between any DNA-base atom and any amino acid side chain atom were further considered. A total of 272 clusters and the same number of cluster representatives, were constructed in this way. The geometries of all pairs are available at http://pdna-iea.uochb.cas.cz/.

The Cα representations of amino acids were prepared by replacing the carbonyl and amide groups with hydrogen atoms as described in Berka *et al.*[4] Any potentially non-specific interactions between the backbone and the DNA base were eliminated by this termination of each amino acid by a methyl group. Histidine was protonated on the ε-N atom and proline was considered uncharged. Regarding the nucleotides only the base

was preserved and the deoxyribose C1’ carbon from N-glycosidic bond was replaced with a hydrogen atom.

As the atlas contains only the positions of heavy atoms, hydrogens were added to each cluster representative and optimized at the B3LYP-D3/def2-TZVPP level.

The cluster representatives were classified according to the physico-chemical character of each amino acid: polar (T, S, N, Q, C, M; 69 contacts), non-polar (G, A, V, I, L, P; 76 contacts), aromatic (F, Y, W, H; 63 contacts), positively (K, R; 33 contacts) and negatively (D, E; 31 contacts) charged cluster representatives as shown in Figure 9. A more detailed description of the methodology can be found in our publication, where mostly bioinformatic aspects and features derived from the distributions involving tens of thousands of contacts were highlighted.[174]

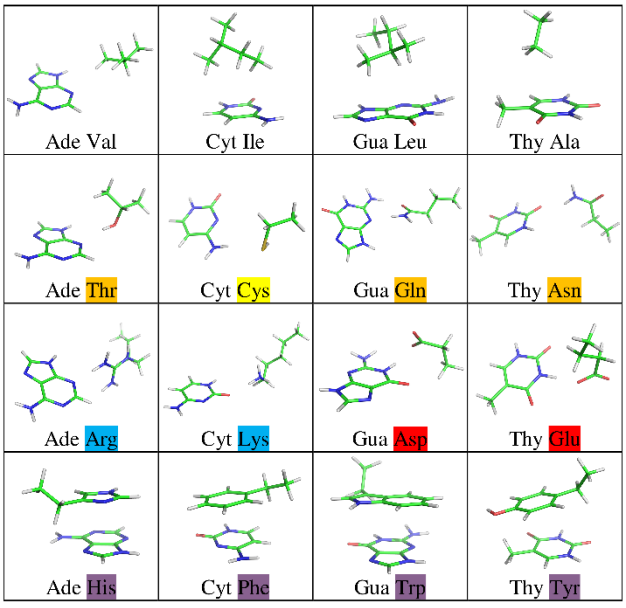


Figure 9. Examples of geometries of DNA bases and Cα representations of amino acid ordered from the top to the bottom column according to the amino acid type: non-polar, polar, charged and aromatic.

3.3.1.1 Benchmark CCSD(T)/CBS Interaction Energies

The reference CCSD(T)/CBS interaction energies were approximated by the eq. 11, specifically large aug-cc-pVQZ basis set was utilized for HF calculations while the second term, $\Delta E^{\mathrm{MP2corr/CBS}}$, was determined using aug-cc-pVTZ and aug-cc-pVQZ basis sets. The ΔCCSD(T) correction term was calculated in a smaller aug-cc-pVDZ basis set. All interaction energies were corrected for BSSE using the counterpoise scheme of Boys and Bernardi.[149] The resolution of identity was used to accelerate the MP2 calculations and the frozen-core approximation was applied systematically to all calculations of correlation energy.[175] The same setup was used to generate extensive data sets of benchmark interaction energies such as *S66x8* and *X40x10*.[55,62]

3.3.1.2 Methods Under Study

The CCSD(T) method has been proven accurate, robust, size-consistent and suitable for single reference calculations of noncovalent interactions. For larger systems than several dozen atoms, the highest accuracy can be achieved with empirically scaled methods based on the scaling of the same- and opposite-spin contributions (such as SCS-MP2 and SCS-MI-MP2)[86] and methods dependent on MP3 energy (MP2.5, MP2.X).[79,176] Next, explicitly correlated MP2 methods represents a systematic way how significantly speed up the basis set convergence toward the CBS limit. More information about these methods can be found above in the sections 2.2.1 and 2.2.2.

Additionally, it is even more important to test methods that are nowadays routinely being applied for large complexes with hundreds of atoms. The balance between the accuracy and computational cost is the reason why DFT is the method of choice for many applications. Here, we included results of B3LYP, BLYP and TPSS functionals combined with def2-TZVPP and def2-QZVP basis sets.[163] Both were augmented with the D3 empirical dispersion term utilizing the BJ damping function (see section 2.3.1).[109,177] Average absolute value of three-body nonadditive terms for complexes presented here was under 0.05 kcal/mol therefore it was not considered.

Further, we have investigated the performance of the SQM PM6-D3H4 method that showed best performance in previous more general tests in section 3.1.[57,150] The empirical force-field calculations were performed with the Gromacs-4.5.5 package[178]

with the combined Amber99SB-ILDN protein force field[119] and Amber94 nucleic-acid parameters.[120,121] This force field was selected as the most commonly used one and the parameters of the Cα representations of amino acids and DNA-base parameters were based on the existing topologies of amino acids and free nucleotides, respectively. More information about parameters used can be found in our original publication.[174] All calculations were performed in the gas phase using a double-precision setup.

3.3.2 Results and Discussion

The wall time that was spent by the methods in the standard computational cluster is listed in Table 4. In Figure 10 correlations are depicted between the benchmark CCSD(T)/CBS estimates and the results obtained by the respective method while Figure 11 shows relative RMSE for the four groups of complexes between the DNA basis and the amino acid residues.

Method and basis set	Time [hours]
CCSD(T) / aug-cc-pVDZ	273
RI-MP2 / aug-cc-pVQZ	117
RI-DFT / def2-QZVP	31.3
RI-MP2 /aug- cc-pVTZ	12.8
RI-MP3 / aug-cc-pVDZ	7.3
RI-MP2-F12 / cc-pVDZ	6.5
RI-DFT / def2-TZVPP	3.3
RI-MP3 / 6-31G*(0.25)	0.6
Amber force field	0.001

Table 4. Computational times for the ade–trp system with 37 atoms. The indentical computational node was used for each calculation: Intel Xeon E5630 2.53 GHz, 8 cores, 5.8 GB RAM per core.

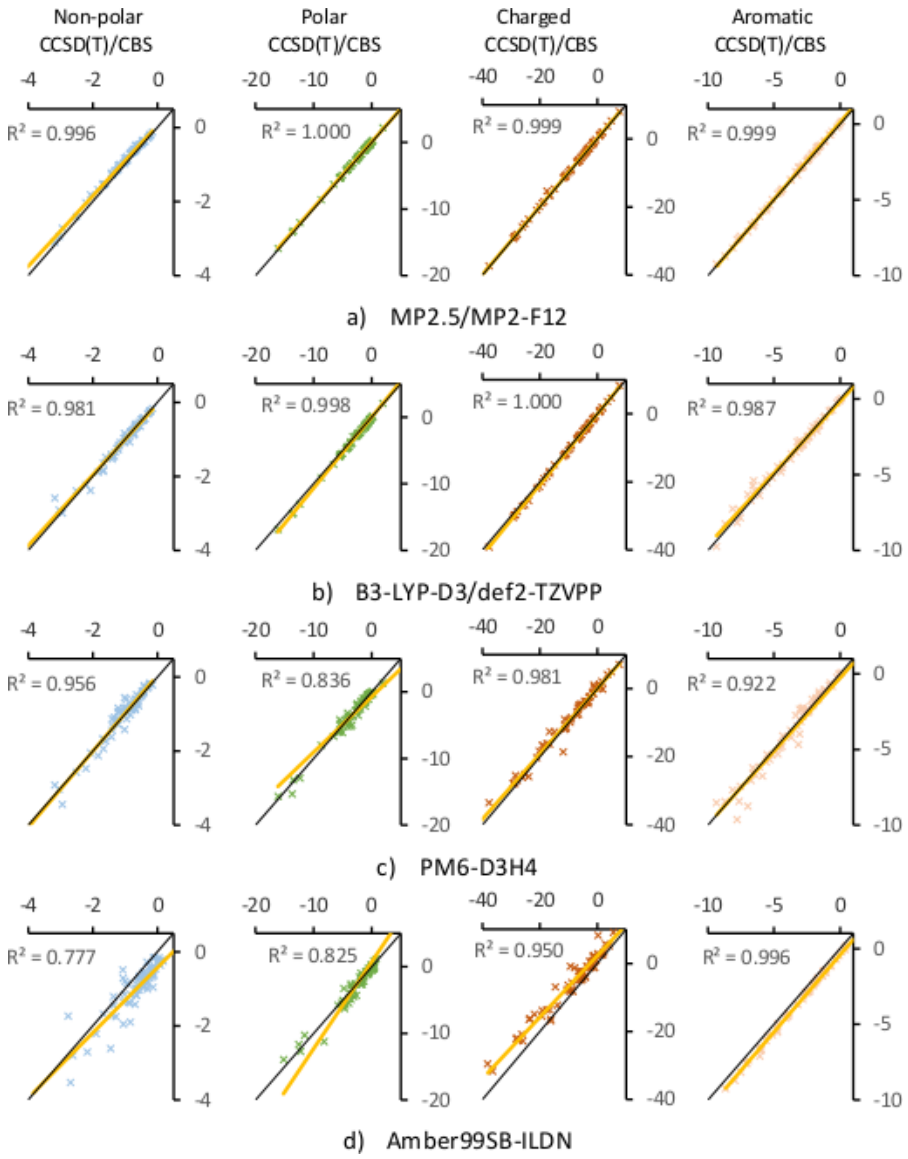


Figure 10. Correlation plots between CCSD(T)/CBS reference and MP2.5/MP2-F12, B3LYP-D3/def2-TZVPP, PM6-D3H4 and Amber99SB-ILDN methods. The yellow line represents the linear regression and the black one has a slope of 1. All energies in kcal/mol.

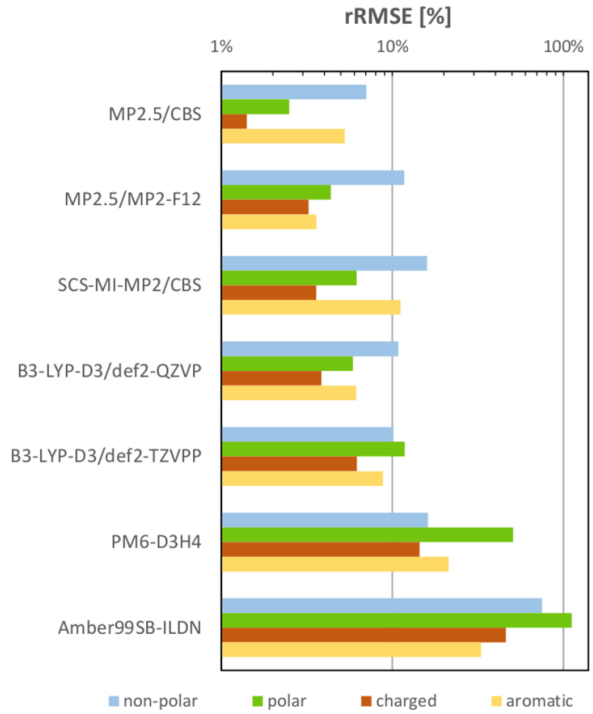


Figure 11. rRMSE for the S66 data set.

3.3.2.1 The Performance of MP2.5/CBS

In previous studies it has been shown that MP2 method strongly overestimates interaction energies in complexes with dominant dispersion interaction.[53] One of the ways how to correct this behavior for systems around 100 atoms it is possible to calculate ΔMP2.5 correction term from the difference between MP3 and MP2 energies (see section 2.2.1). In our study, we used two basis sets, 6-31G*(0.25) and aug-cc-pVDZ. When the smaller basis set 6-31G*(0.25), was used, the results were comparable (RMSE 0.11 and 0.13 kcal/mol larger smaller basis set, respectively) to the case of the considerably larger, and thus much more time-consuming, aug-cc-pVDZ basis set (see Table 4). The usage of smaller basis set deteriorated results only for the systems with aromatic amino acids where interaction energies were slightly overestimated (with MSE -0.07 kcal/mol).

Next, we studied the correlation and rRMSE between the ΔCCSD(T) and ΔMP2.5 correction terms. Surprisingly, contradictory results have been obtained. The correlation of more than 95% has been achieved for all types of interactions with the exception of

charged systems, where it was only 72% and 67% for the larger and smaller basis sets, respectively. On the other hand, rRMSE was for charged systems below 2% while for neutral complexes was around 4%. This is caused by relative the highest absolute interaction energies of charged systems among all interaction types and probably the relatively small basis set size. These findings support previous studies concerning the very good performance of the 6-31G*(0.25) basis set, which makes it a promising tool for various applications in the field of noncovalent interactions.

3.3.2.2 Performance of MP2-F12 method and Composite Schemes

Significant improvement of MP2 method for basis set convergence toward CBS limit can be achieved when explicitly correlated methods are used. We tested here MP2-F12 that can be utilized instead of two separate MP2 calculations with a systematically increasing size of correlation-consistent basis sets. We compared MP2-F12/cc-pVDZ(-F12) results with MP2/CBS benchmark interaction energies. MP2-F12 method performed well for the complexes with non-polar and polar amino acids with a RMSE below 0.10 kcal/mol. Inspecting the complexes containing aromatic amino acids and charged systems, we found larger discrepancies resulting in RMSEs of 0.17 and 0.31 kcal/mol, respectively.

Partial error cancellation takes place when MP2-F12 and the ΔMP2.5/6-31G*(0.25) correction term are combined. As mentioned above, ΔMP2.5 correction term for systems with aromatic amino acids has a mean signed error with the opposite sign when compared to MP2-F12 with MSE 0.19 kcal/mol. The resulting RMSE for aromatic complexes is about 0.12 kcal/mol. This combination of MP2.5 and MP2-F12 method afforded only 0.20 kcal/mol overall RMSE for our data set of DNA-base dimers. This approach results in computational savings of almost 2 orders of magnitude when compared to the CCSD(T)/CBS calculation.

3.3.2.3 A Comparison of DFT-D3, PM6 and Amber Force Field with CCSD(T)/CBS Methods

Figures 10 and 11 show the performance of Amber99SB-ILDN force field, PM6 and B3LYP-D3 method (with def2-TZVPP and def2-QZVP basis sets). The key findings will be highlighted and discussed next.

The Amber99SB-ILDN force field performs well for neutral polar and aromatic systems, however, significantly underestimates charged complexes (see the second column from the right in the Figure 10). The inclusion of deformation energy slightly improved the overall performance and decreased RMSE from 2.6 to 2.3 kcal/mol. Clearly, the systematically worse agreement with the benchmark data was achieved for positively charged systems.

The SQM PM6 method without corrections exhibit slightly worse results than force field calculations (RMSE 2.5 kcal/mol). Significant improvement can be achieved when post-SCF correction for dispersion interactions and hydrogen bonding are applied. The largest improvement was shown for aromatic systems where RMSE decreased to one third (to 0.8 kcal/mol). The same trends were observed in non-polar systems (a decrease of the RMSE from 1.0 kcal/mol to 0.2 kcal/mol), charged complexes (the RMSE decreased from 3.5 kcal/mol to 1.5 kcal/mol) and polar systems (decreasing the RMSE from 2.2 kcal/mol to 1.5 kcal/mol).

Finally, the DFT-D method was tested. The larger def2-QZVP basis set lowers both the absolute and the relative errors roughly by one third (RMSE 0.25 kcal/mol) when compared to smaller TZ-size basis set, that overestimates mainly polar and negatively charged systems. Although we found that both basis sets systematically overestimates the strength of the interactions, they perform well for non-polar and aromatic systems. Both basis sets can be recommended for applications aiming at larger complexes containing several hundreds of atoms. The other two functionals, meta-GGA TPSS and GGA-type BLYP, performed slightly worse (with the RMSE being 0.34 kcal/mol and 0.35 kcal/mol, respectively) with significant computational savings when compared to B3LYP hybrid functional.

3.3.2.4 Interaction Energy Distributions Calculated by MM and DFT Methods

In previous sections we analyzed the performance of *ab initio*, DFT and MM methodologies. Next, a brief analysis of the interaction energy profiles of clusters associated with a certain amino acid−base pair follows. These energy profiles clearly revealed that the cluster representatives derived according to straightforward distance criteria and clustering procedure (see section 3.3.1) indeed represent the most typical interaction energy value found within the members of its associated cluster. The similar findings were reported by Berka *et al.* for amino acid side chain−side chain interactions.[4]

Attention should be paid to comparison of interaction energy profiles for individual clusters to the energy profile of the entire distribution. Following general observations could be drawn in several distributions: (I.) the clusters accommodate most of the constants within the interaction energy range (II.) those contacts with the highest stabilization energy in the energy profile form the clusters (III) the low energetic contacts do not correspond to any cluster (IV.) the peaks of the cluster interaction energy distributions corresponds well with the position of the cluster representatives.

These properties were found for several clusters occurring in the guanine−glutamine, adenine−asparagine, cytosine−asparagine, cytosine−tyrosine, adenine−glutamine and adenine−lysine energy distribution profiles. An example of such an energy profile calculated using Amber03 is shown in Figure 12. Next, the lowest lying energy structures of the adenine−asparagine, adenine−glutamine and the guanine−glutamine contacts were studied. We have found the presence of two hydrogen bonds and a close resemblance between these structures in biomolecules and the local minima determined by full geometry optimizations.

We verified our observations concerning the distributions by recalculating all cluster-associated contacts for adenine−glutamine contacts using the DFT-D/B3LYP-D3/def2-TZVPP method. Despite the small shift toward more negative values, the respective interaction energy DFT profiles verified the empirical results very well (Figure 12).

Detail bioinformatic consideration about the limits of our study, presence of low-lying hydrogen bonds, rest of the pairs and the examination of changes in geometries after full optimization can be found in our original publication.[174]

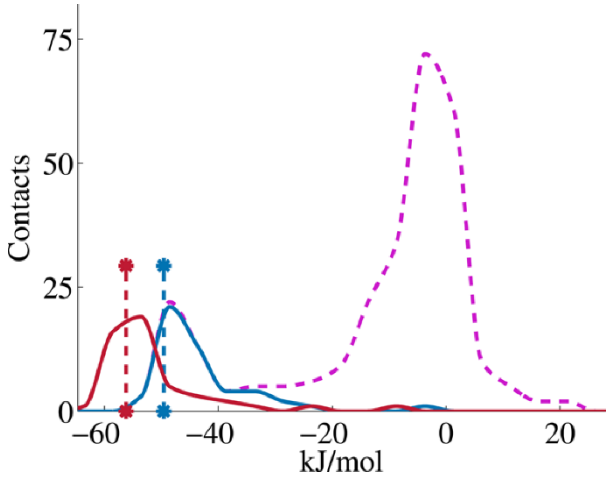


Figure 12. Ade−glu interaction energy profiles calculated with force field are depicted as dashed purple curve; pronounced low-lying cluster is in solid blue curve; additionally red solid profile of all cluster-associated contacts was calculated with B3LYP-D3/def2-TZVPP method; dashed vertical lines corresponds to energy of cluster representatives. (all in kJ/mol)

3.3.3 Conclusions

We have quantitatively investigated amino acid-base preferences based on the crystal structures. In several cases of amino acid−base pairs we found unique low lying interaction energies distinct from the rest of the distributions. These findings were verified with DFT-D/ B3LYP-D3/def2-TZVPP methodology.

Additionally, we have analyzed 272 representative pairs of amino acid side chains with nucleic-acid bases. For these the benchmark CCSD(T)/CBS interaction energies were calculated and used for testing various methods. We found that MP2.5 method achieved small RMSE of 0.11 kcal/mol (relative error of 2%) when compared to CCSD(T) method. This technique shows great promise for the future larger scale applications. Among DFT functionals the B3LYP systematically overestimated the strength of the binding by up to 0.31 kcal/mol in positively charged systems. The RMSE value of 0.25 kcal/mol is slightly increased to 0.34 and 0.35 kcal/mol, when two tested functionals, TPSS and BLYP, are used instead, respectively. PM6-D3H4 and Amber99SB-ILDN force field were in reasonable agreement with benchmark method.

3.4 Computational Analysis of X-ray Crystal Structures of Organic Compounds

It is known that many molecules undergo large conformational movements upon binding of specific ligand molecules and there are not always X-ray structural data available to determine these changes easily by their fairly straightforward investigation. Furthermore, the positions of water molecules, the role of crystal contacts and crystal packing on the complex structure are hard to determine. In these situations the position of computational chemistry seems bright. Firstly, not always there is possibility to obtain clean X-ray crystal structure of high quality. Secondly, the effect of environment can be easily seen if appropriate efficient and accurate computational method is chosen. Thirdly, virtual experiments can be performed in order to isolate different effects under experimentally unreachable conditions of *e.g.* different solvent, counterions, pH or temperatures. Here, we present one case, where computational chemistry 1) brought deeper insight to the binding preferences and properties 2) reproduced experimental data with reasonable accuracy.

Choline (Ch+) molecule is the precursor for synthesis of neurotransmitter acetylcholine (ACh+) as well as phosphatidylcholine and sphingomyelin, two classes of phospholipids having essential role in cell membranes. The first step in the respective biosynthesis is always the selective binding of this water soluble vitamin. One of such examples has been described in case of ChoX protein from the Sinorhizobium meliloti family, a plant root-associated bacterium. This protein discriminate all competitors such as acetylcholine and glycine betaine in favor of choline. The X-ray crystal structure of Ch+•ChoX complex shows aromatic cage binding trimethyl ammonium cation through strongly directional cation-π interactions in site I.[179] This represents the main binding of these moieties. The site II with two carboxyl groups fix the hydroxyl tail through hydrogen bonds and determine the binding specificity between *e.g.* choline, acetylcholine.

The studies of naturally assembled protein cavities constitute a driving force for synthesis of self-assembled artificial mimics featuring similar binding modes and properties.[180] Despite of great success in case of suitable artificial receptors, those showing high selectivity for choline remain very rare. The strongest competitor for choline is referred to be acetylcholine, therefore the selectivity is typically defined as

the ratio between equilibrium constants of choline and acetylcholine: K(Ch+)/K(ACh+). In nature, the reported selectivity by ChoX protein is 54, while among artificial mimics the choline, to the best of our knowledge, was reported to bind no more than 3-fold stronger than acetylcholine.[179,181]

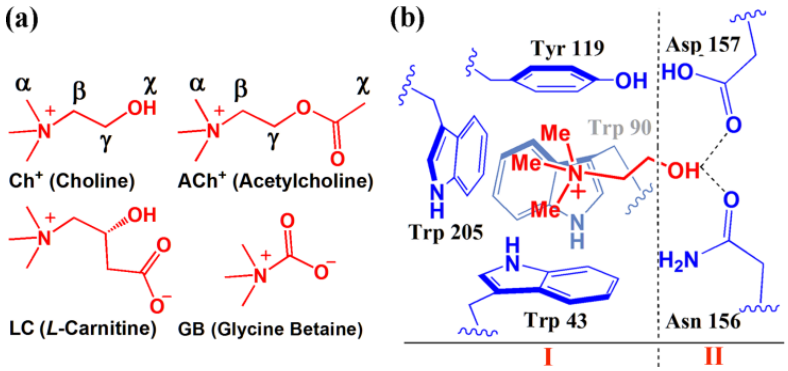


Figure 13. (a) Structures of choline and its three derivatives/competitors; (b) binding motif of choline inside of the protein ChoX (with two binding sites: I and II).[179]

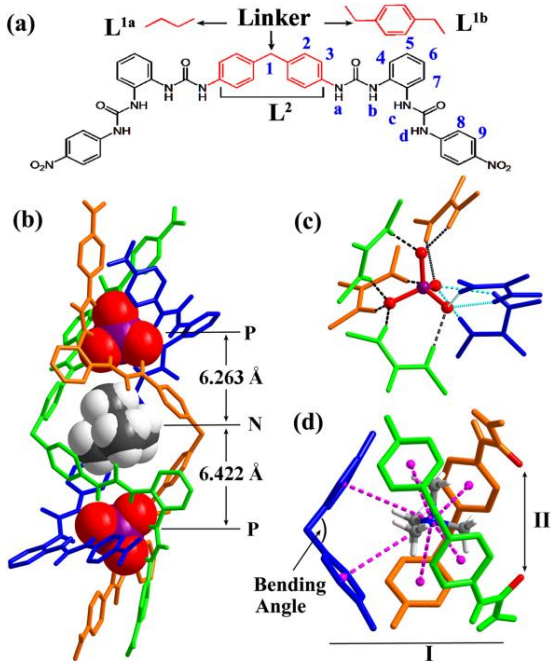


Figure 14. (a) Structures of $\mathbf{L}^{1a}$, $\mathbf{L}^{1b}$ and $\mathbf{L}^{2}$ organic molecules (b) Crystal structure of $(TMA)_5[(TMA)\bullet(PO_4)_2(\mathbf{L}_2)_3]$ complex; (c) Hydrogen bonds formed between a PO_4^{3-} ion and six urea units of $\mathbf{L}^{2}$ molecules; (d) The aromatic cage trapping a TMA+ through cation-π interactions (purple dashed lines).

Here, we studied for the first time a a triple helicate assembled from a bis-biurea ligands. It was functionalized with an "aromatic cage" that resembles ChoX protein binding pocket and achieved selective binding of choline with the binding affinity of Ch+ as high as 20 times of that of ACh+. The equilibrium constants were measured in 1H NMR competition experiments. Additionally, we obtained crystal structure of $(TMA)_5[(TMA)\bullet(PO_4)_2(\mathbf{L}^2)_3]$ (Figure 14b), which possess desired large cage formed by six phenyl rings of the three 4,4'-methylenebis(phenyl)- linkers. In this binding pocket a tetramethylammonium (TMA+) cation was trapped.

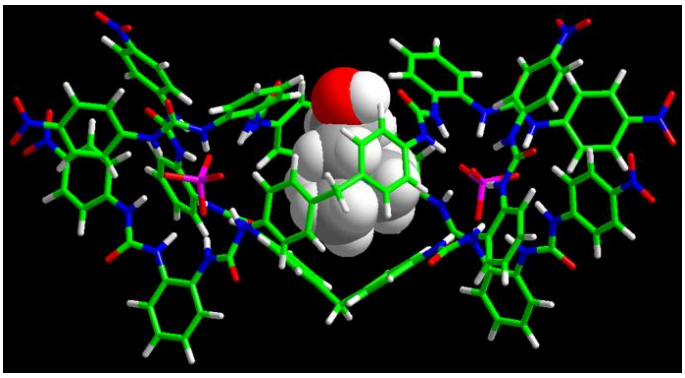
3.4.1 Computational Details

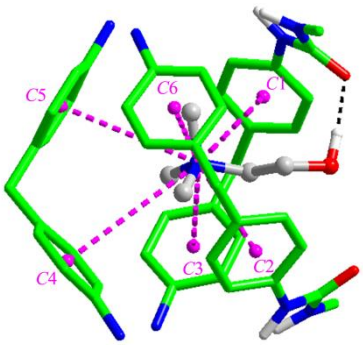
As the X-ray crystal structures of acetylcholine and choline molecules with triple anion helicates were not available, the starting geometries for DFT-D/BLYP-D/def2-SVP geometry optimizations were prepared by modification of the crystal structure of the complex with tetramethylammonium$^+$ moiety. All DFT calculations have been performed utilizing COSMO solvation model in order to include effect of the environment (acetone). To understand the nature of the stability of the complexes, the DFT-D/BLYP-D3/def2-TZVPP interaction energy and other parameters of binding were calculated: dispersion energy (ΔE_{disp}), deformation energies E_{def}(guest) and $E_{def}([(PO_4)_2(\mathbf{L^2})_3]$ and electrostatic energies ($\Delta E_{\text{electro}}$). The $\Delta E_{\text{electro}}$ was determined by the Coulomb law calculated by using NBO atomic charges. We will show that these crude estimates of $\Delta E_{\text{electro}}$ correlate well with more advanced Energy Decomposition Analysis method (see later in case of host•guest systems, section 3.5.2.4). The second derivatives and entropy analysis were calculated with Amber at 298 K and change of the Gibbs energy (ΔG) were derived.[119]

3.4.2 Results and Discussion

The DFT optimized structure of Ch+•2 (Figure 15) demonstrates a surprisingly similar dual-site binding mode with that displayed by the crystal structure of Ch+•ChoX (Figure

13b). The DFT optimized structure of the analogous ACh+•2 (Figure 16) displays a similar aromatic cage encapsulating the trimethylammonium head and the methyl protons of the acetyl group are extending out of the cavity which is consistent with the results demonstrated by 1H NMR. In contrast to Ch+•2, no hydrogen bond was formed in ACh+•2 with the tail group. Overall, the DFT results show that the binding of Ach+ is 2.8 kcal/mol weaker than that of Ch+ (Table 5). The N···centroid distances in ACh+•2 were also measured as ranging from 4.28 to 4.86 Å with an average of 4.62 Å.

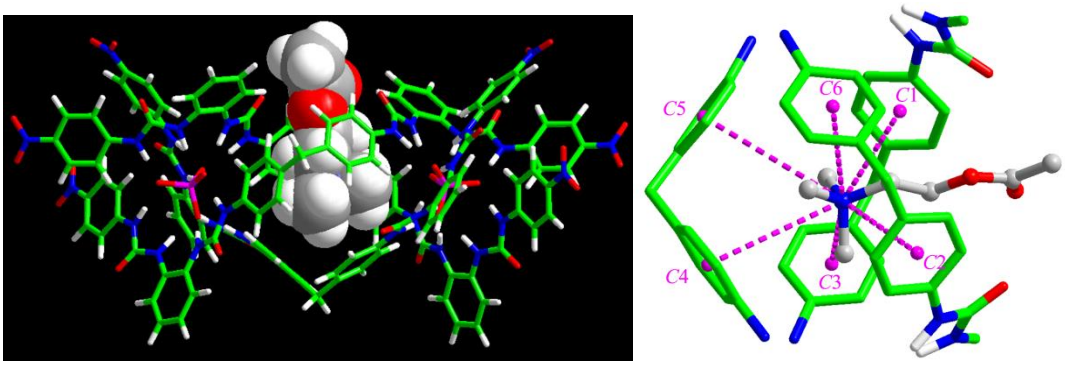




N-C1	N-C2	N-C3	N-C4	N-C5	N-C6	Average
4.35	4.93	4.81	4.82	4.52	4.67	4.68

Figure 15. DFT optimized structure of Ch^{+}•**2** (left) and choline binding sites (right) with data of N···centroid distances for evaluating the cation-π interactions (purple dashed lines, distances are in Å).

Inspecting the Table 5, we find that the almost 3 kcal/mol difference is originating mainly from the entropic contribution. This has lead us to investigate the role of size of the system for entropy calculations. It is known for protein•ligand complexes, that the usage of whole protein structure for entropy calculations deteriorated results. However, we calculated entropy term for two system sizes here for ACh+•2 complex (Figure 14b and 14d) with essentially the same results (4.5 and 4.6 kcal/mol, see Table 5).



N-C1	N-C2	N-C3	N-C4	N-C5	N-C6	Average
4.28	4.83	4.7	4.86	4.51	4.54	4.62

Figure 16. DFT optimized structure of ACh^{+}•**2** (left) and choline binding sites (right) with data of N···centroid distances for evaluating the cation-π interactions (purple dashed lines, distances are in Å).

Investigating binding parameters in Table 5 we can conclude that ACh+ molecule is more dispersion bound when compared to more electrostatically bound Ch+ guest molecule. This is probably caused by the presence of strong hydrogen bond between Choline and tail Triple Helicate group (Figure 15). Both these contributions are compensating the deformation energies (E_{def}(guest) and E_{def}([$(PO_4)_2$(**L²**)$_3$]) that are larger for Ch+ molecule.

	ΔE	E_{def}(guest)	E_{def}([$(PO_4)_2$(**L²**)$_3$]	ΔG	-TΔS	ΔH
Ch^{+}	-46	0.9	4.6	-39	1.5	-40.9
ACh^{+}	-45	0.3	3.5	-37	4.6(4.5)	-41.2

ΔE decomposition			
	ΔE_{disp} (2-body)	$\Delta E_{electro}$	ΔE_{disp} (3-body)
Ch^{+}	-43.3	-270.5	1.5
ACh^{+}	-49.6	-260.8	1.6

Table 5. DFT results for ACh+•**2** and Ch+•**2** complexes.

3.4.3 Conclusions

Our computational approach determined that the Ch+ is bound 2.8 kcal/mol stronger (ΔG) than Ach+, corresponding a K/K = 109. This is in nice correlation with the experimental value of 20. When comparing the calculated enthalpic and entropic contributions to the change of the Gibbs energy, we can conclude that surprisingly relative increase of the entropic contribution from Ch+ to ACh+ is causing the discrimination of Ch+ over ACh+ guest molecule.

However, it is clear that these rather small differences in ΔG are within the accuracy of MM method. In order to bring deeper insight, show trends or draw any important conclusions into host•guest binding several requirements on studied guest molecules would have to be fulfilled. Firstly, not only monocationic guest molecules should be studied. Secondly, it would be beneficial to include guest molecules of various sizes. And finally, larger range of equilibrium constants is mandatory. All these needs were met in the next section thoroughly discussing interactions of the cucurbit[n]uril host molecules with list of guest molecules.

3.5 Cucurbit[n]uril•Guest Binding Interactions

The computational chemistry already found its place in the field of new drug discovery. However, in this realm of protein•drug complexes its utility is rather limited without a firm foundation of experimental structural data. The supramolecular chemistry of host•guest complexes represents an logical next step from noncovalently bound small molecules in *vacuo* (*e.g.* amino acid–DNA-base dimers as described above), self-assembled triple helicate rigid cages in nonpolar solvents toward protein•drug complexes in water.[182] Especially the chemistry of cucurbit[n]uril (CB[n], n=5,6,7,8,10) macrocycles has undergone wide development and growth in the last decade and great interest has been stimulated leading to promising applications in materials chemistry, molecular recognition, drug discovery and chemosensing. CB[n] molecules found its applications *e.g.* in reaction inhibition (as protective groups), catalysis (in hydrolysis, photoreactions and dipolar cyclo-additions), recognition of peptides and native proteins (Trp-rich peptide or N-terminal Phe residues) and as drug carriers (increasing solubility, reducing cytoxicity or drug degradation during manufacture).[183-187]

The 'cucurbit' prefix in the name of cucurbit[n]uril's (Figure 17) is derived from its gourd or pumpkin-like shape and the 'uril' suffix originates from its methylene-ligated glycoluril building blocks.

Adamantane and diamantane salts, *e.g.* adamantane-1-ammonium or diamantane-4,9-diamonium, are known to be relatively strong binding guests to CB[7]. Very high values of equilibrium constants were measured in 50 mM NaO_2CCD_3 buffer at pD 4.74 and it can be as high as 10^{17} in case of CB[7]•diamantane-4,9-di(NMe_3I) complex reported as a 1000-fold stronger nature's best effort (avidin with its biotin cofactor). These assemblies, representing the strongest noncovalently bound host•guest complexes known, are a challenging type of systems for any contemporary computational method.

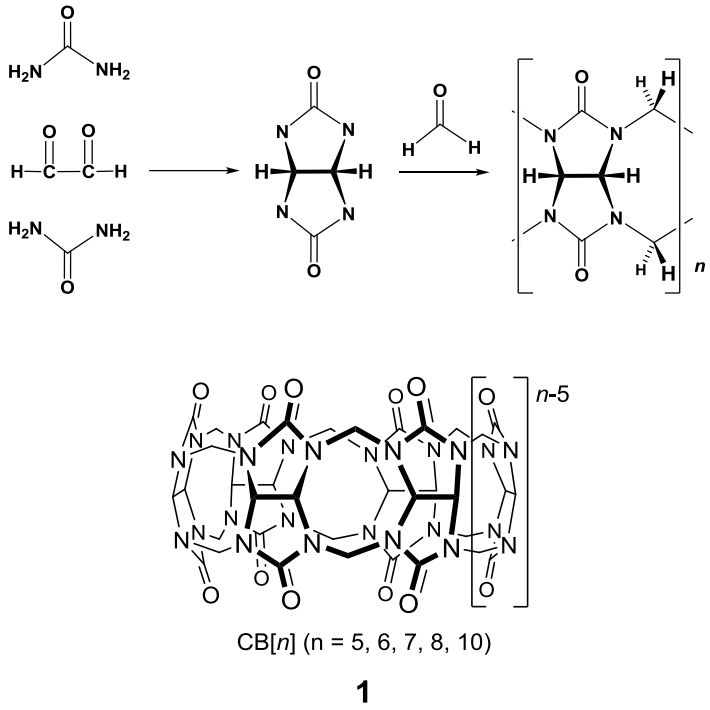


Figure 17. Synthesis and scheme of cucurbit[n]uril molecule.

Generally, the errors and limitations in Δ*G* predictions in any protein•ligand applications are caused by two factors: (1) unpredictably large errors arising from estimation of desolvation free energies (*i.e.* the absence of benchmark experimental data for charged species) and (2) missing structural data about the studied systems and (3) system size that restricts the range of affordable computational methods to mostly SQM or force field methods. Fortuitously, the <200 total number of atoms in CB[n]•guest complexes is small enough for accurate QM calculations of interaction energies.[188] Another very important feature of CB[n]•guest complexes is that the X-ray crystal structures of several complexes are available and they constitute the actual binding geometries as was verified by several NMR studies. In all applications and types of the sytems, the accurate description of noncovalent interactions plays a crucial role wherein a balanced treatment of hydrogen bonding, halogen bonding, London dispersion forces, polar, electrostatic interactions, *etc*. is of prime importance.[71] Here, the DFT has been chosen for the description of driving forces in CB[n]•guest complexes. Current development in this field has been described in great detail in recent publications of Grimme *et al.*[189] and Jensen[190].

Lately, various experimental parameters (*e.g.* equilibrium constants and changes in free energy) has been more or less successfully reproduced.[189,191,192] The bottleneck in these studies is the accurate prediction of guest's spatial arrangement inside the host cavities as was reported by Gilson *et al*.[193] Among the techniques tested in the recent blind test challenges (SAMPL4 and 5) none fulfilled all the goals regarding prediction of complex geometries and binding affinities, although DFT-D3 performed rather well in some of them.[189,194] The ranking of the guests was not completely reproduced because host•guest complex were not represented well by a single configuration or small guest molecule arrangement within the host cavity. The situation is greatly simplified when high quality X-ray structures are available.

In this project, we studied eleven X-ray structures of systematically varied geometry guest molecules within a (CB[7 or 8]) hosts. Structures of these biomimetic complexes showed that no high-energy water molecules present inside of the host cavities are required for modeling their binding interactions. This finding greatly simplifies modeling of the studied complexes. The final goal of this rational design approach was to computationally ascertain the various binding motifs (see Figure 18) of host•guest interactions and then maximize the strength of the binding through design of new guest molecules reaching unused CB[n] host sites.

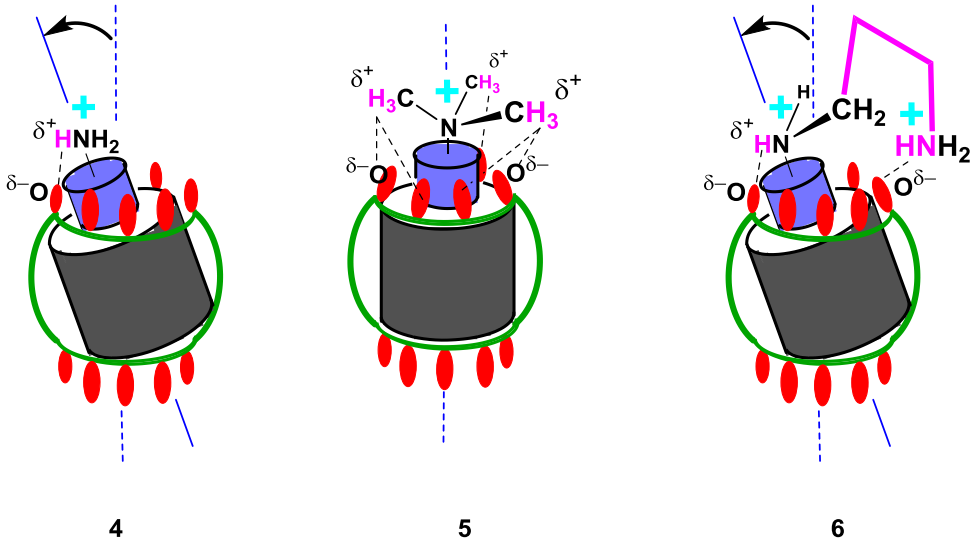


Figure 18. Three distinctive binding modes: primary (4), tertiary (5) and loop (6).

3.5.1 Methods

3.5.1.1 Studied Complexes

Training set of 11 complexes based on X-ray crystal structures was used for our computational protocol. The guests studied herein are depicted in Figure 19. In several cases only preliminary or no X-ray structure were available. Therefore the starting geometry of CB[7]•2 was derived from CB[7]•**9** by removing both methyl substituents and CB[7]•**3** by deleting half of the diamantane-4,9-di(NMe_3) guest (CB[7]•**8**). Similarly, the starting geometries of the complexes containing loops (CB[7]•**15**, CB[7]•**16** and CB[7]•17) were prepared by adding loops to similar crystal structures of host•guest complexes (different conformations were systematically considered and after optimization procedures only the one with lowest energy used). Finally, we studied the accuracy of our computational protocol on smaller complex of glycoluril (**18**) and 2,3-dimethyl-2,9-diaza-decane (**19**) prepared from the CB[7]•7 structure, wherein hydrogen atoms were added and geometry optimized at the DFT-D level.

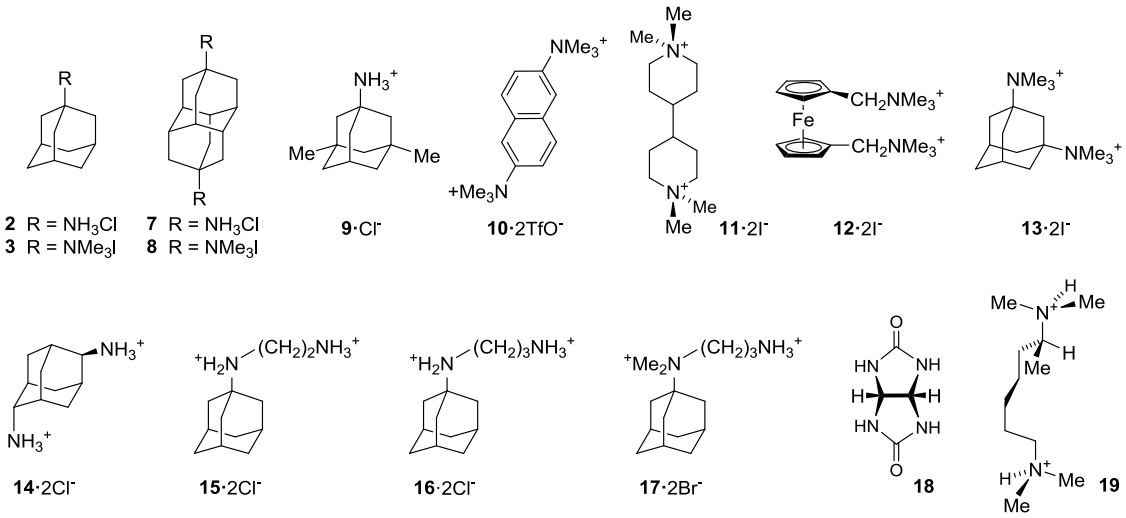


Figure 19. Illustration of guest molecules considered in our study: adamantane/diamantane, naphthalene, bipiperidine and ferrocene derivates together with small model systems glycoluril, and 2,3-dimethyl-2,9-diaza-decane.

Table 6 shows binding constants (K_a) and changes of Gibbs energies upon binding (ΔG_{exptl}) of guests **2-3**,**7-17** to CB[7] and CB[8] hosts. Both these quantities were determined under the same conditions with the same protocol. Comparison of experimental data reported by several laboratories and measured by different techniques can lead to inconsistencies. Because pKa values of weakly acid ammonium salts differ in various media, it can be expected that the data not measured under the same conditions as here (50 mM NaO_2CCD_3 buffered D_2O at pD 4.74) would deteriorate results. Secondly, the role and location of counterions are not yet fully understood and therefore counterions are omitted from our calculations.

Complex	Guest	K_a	ΔG_{exptl}[a]
CB[7]•2	Ada-1-NH_3Cl	$(4.23\pm1.00)\times10^{12}$ [b]	–17.20±0.14
CB[7]•3	Ada-1-NMe_3Cl	$(1.71\pm0.40)\times10^{12}$ [b]	–16.66±0.14
CB[7]•7	Diam-4,9-di(NMe_3I)	$(2.0\pm0.5)\times10^{15}$ [c]	–20.84±0.15
CB[8]•8	Diam-4,9-di(NH_3Cl)	$(8.3\pm2.3)\times10^{11}$ [c]	–16.22±0.17
CB[8]•9	3,5-diMeAda-1-NH_3Cl	$(4.33\pm1.11)\times10^{11}$	–15.84±0.16
CB[7]•10	Naph-2,6-di(NMe_3Tfl)	$(1.7\pm0.4)\times10^{11}$ [d]	–15.29±0.14
CB[7]•11	4,4'-Bipip-N,N'-di(NMe_2I)	$(1.9\pm0.4)\times10^{11}$ [d]	–15.36±0.13
CB[7]•12	Ferro-1,1'-di(CH_2NMe_3I)	$(1.9\pm0.4)\times10^{13}$ [c]	–18.09±0.13
CB[8]•13	Ada-1,3-di(NMe_3I)	$(1.11\pm0.28)\times10^{11}$ [b]	–15.04±0.15
CB[7]•14	Ada-2,6-di(NH_3Cl)	$(1.2\pm0.4)\times10^{12}$	–16.43±0.21
CB[8]•14	Ada-2,6-di(NH_3Cl)	$(4.7\pm1.2)\times10^{8}$	–11.80±0.15
CB[7]•15	Ada-1-$NH_2(CH_2)_2NH_3Cl_2$	$(2.4\pm0.6)\times10^{13}$ [c]	–18.22±0.15
CB[7]•16	Ada-1-$NH_2(CH_2)_3NH_3Cl_2$	$(1.5\pm0.4)\times10^{13}$	–17.94±0.16
CB[7]•17	Ada-1-$NMe_2(CH_2)_3NH_3Br_2$	$(6.8\pm1.6)\times10^{12}$	–17.48±0.14

Table 6. Experimental values: K_a (M^{-1}) and ΔG_{exptl} for CB[n] complexes with various mono-cationic and di-cationic guest molecules. Footnotes: [a] Calculated at 298 K, energies are in kcal/mol. [b] Data taken from ref. 2. [c] Data taken from ref. 192. [d] Data taken from ref. 188.

3.5.1.2 Computational Details

As well as in previous studies the DFT-D/BLYP/def2-SVP/COSMO//DFT-D/BLYP/def2-TZVPP/COSMO has been chosen here for description of geometry and several binding parameters of host•guest systems. In several cases (CB[7]•**3**,**7**,**9**,**10**, or **11** and CB[8]•**8**) the larger def2-TZVPP basis set was used for geometry optimizations in order to validate the above approach. The RMSD between both sets of geometries was negligible ranging from 0.039 and 0.115 Å. It justifies the above approach.

Several methods [FN-DMC, CCSD(T)/CBS and MP2.5/CBS] were used for accuracy assessment of BLYP-D3/def2-TZVPP method for interaction energy calculations, as described later. MP2.5 and CCSD(T) interaction energies were corrected for BSSE and constructed as described previously in the section 2.2.

Deformation energies ($E_{def\ (host)}$ and $E_{def\ (guest)}$) were calculated with the same method as the geometry optimizations. The solvents effects were thoroughly studied using COSMO and SMD continuum solvent models with applied permittivity ε equal 78.5. The entropy analysis and second derivatives were obtained with Amber force field at 298 K.[119] The interaction energy decomposition has been performed by EDA method (see section 2.8) using BLYP functional and def2-TZVPP basis set. Additionally, NBO atomic charges were used for estimates of the electrostatic interaction energies ($\Delta E_{electro}$) determined by the Coulomb law.

Following three implicit solvent models were considered for the calculations of desolvation free energies: SQM PM6 method and DFT/BLYP-D/def2-SVP method both linked with the COSMO continuous solvation model (resulting in $COSMO^{PM6}$, and $COSMO^{DFT}$) and SMD model based upon the HF/6-31G* electron density (SMD). The desolvation free energy was calculated as the difference between the gas phase energy of the host and energy of the host calculated with the particular implicit solvent model applied.

The explicit solvation model has been utilized as implemented in WaterMap (WM) module in Schrödinger software package (see section 2.8). We performed MD simulation (10ps long) of a CB[n] molecules dipped in 12 Å/side periodic cubic box filled with explicit water molecules. The atom positions of CB[n] molecules were frozen during our simulation. The MD simulation is accompanied by trajectory analysis resulting in following water molecule properties: location, occupancy, enthalpy,

entropy and free energy. Enthalpy and entropy is associated with the transfer of the solute molecule from a bulk water to environment.

Finally, the dynamic elasticity of CB[n] hosts in water environment was studied with MD simulations using PM6-D3 and BLYP-D/def2-SVP methods. The trajectory analysis has been performed with the Visual Molecular Dynamics software.[195] We report here the $RMSD_{average}$ calculated as:

$$RMSD_{\text{average}} = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} [ri(t_1) - ri(t_2)]^2}{N_{atoms}}} \tag{20}$$

where *ri(t)* is the position of the *i*th atom at time t and N_{atoms} is the total number of atoms of a CB[n] host molecule. The systems were simulated under the following conditions: 298 K and implicit solvent was applied (no explicit waters were included). The equilibrium properties were calculated after discarding the initial 1 ps long nonstationary segment of the simulated trajectory.

3.5.2 Results and Discussion

In this section properties of host and guest molecules will be discussed as first. Accuracy assessment of DFT method for cucurbit[n]uril host•guest complexes will follow. Next, correlation between experimental ΔG and the calculated estimates will be described. Finally, the binding parameters and design of new guest molecules will be discussed.

3.5.2.1 Desolvation Free Energy and Stiffness of Cucurbit[n]uril (n=5,6,7,8) Host Molecules

Section 2.9 described the solvation effects in detail. Here, we would like to remind that since the CB[n] host is polar and the studied guest molecules are charged, the significant part of the interaction energy (calculated *e.g.* by DFT methodology) is originating from the electrostatic energy between them. However, it is strongly damped when passing

from *vacuo* into water environment. Clearly, the environment can severely change the physical nature of binding and thus the accurate description of solvent effects represents a crucial point in our approach. Secondly, Nau and coworkers pointed out that the release of high energy water from the cavity of CB[n] macrocycles is a major determinant for guest binding in aqueous solutions.[196,197] Now, we will report our investigation of solvation of isolated host molecules (both the explicit and implicit solvation models were compared).

Characteristics for CB[n] hosts composed of different numbers of glycoluril units (n = 5,6,7,8) with the increasing diameter size will be now compared with existing literature data.[196,197] The second, third and fourth columns in Table 7 show the linear increase of ΔG_{desolv} with the host's increasing diameter. It is important to note that these results show only the nonspecific solvation because the high energy waters in cavity are not modeled. On the contrary, these are described by WaterMap calculations. The specific solvation free energies results are in fifth column ($\Delta G_{pot}{}^{WM}$) while the last column contains their enthalpy and entropy components. The calculations of CB[7] and CB[8] host molecules showed the most favorable solvation (*i.e.* the smallest ΔG). This finding is not in accord with the results of Nau and coworkers as listed in the left-hand data-column showing theoretical difference in potential energy ($-\Delta E_{pot}$) of water molecules in a spherical cavity within the aqueous bulk and inside the host cavity. However, if one takes into account large error bars then it is apparent that reported potential energies for CB[7] and CB[8] are statistically the similar. Therefore, their findings basically agree with our WaterMap generated results.

Next, WaterMap simulations determine the number of water molecules with high energy residing within the host cavity in the absence of an encapsulated guest. A nice agreement depicted in Table 8 has been found with both MD simulation literature results in the first column and Packing Coefficients (PC) analysis results in the second column.

	$-\Delta E_{pot}$ [a]	COSMO^{DFT} [b]	COSMO^{PM6}	SMD	ΔG_{pot}^{WM}	$\Delta H/-T\Delta S_{pot}^{WM}$
CB[5]	41.6±28.8	71.5	84.1	134	12	7.6 / 4.4
CB[6]	51.1±29.	93.6	117.1	158.7	21.4	12.3 / 9.1
CB[7]	102.4±31.3	108.2	138.8	180.8	5.3	–4.9 / 10.2
CB[8]	66.2±10.7	122.7	162	202.1	4.9	–10.5 / 14.9

Table 7. Calculated energies related to solvation of CB[n] hosts, all units are kcal/mol. Footnote: [a]Data taken from ref. 193.

			WaterMap		
	MD [a]	PC analysis [a]	Water Sites	Avg. Occupancy	$N_{water\ molecules}$ [b]
CB[5]	2 [2.0]	**2**[c]	2	94%	~2
CB[6]	4 [3.3]	**4**[c]	6	56%	~4
CB[7]	7 [7.9]	**8**[c]	12	41%	~8
CB[8]	10 [13.1]	**16**[c]	19	41%	~16

Table 8. The number of water molecules trapped within a CB[n]-host's cavity as studied by MD, PC analysis, and WaterMap methodologies. Footnotes: [a] Data taken from ref. 196. Data in square brackets taken from ref. 197. [b] Number of water molecules determined by analysis of water sites.

Finally, MD simulations of CB[n] host molecules in implicit water solvent environment have been used for the studies of the dynamic elasticity. Table 9 shows the increase of CB[n] host's deformability (dynamic stiffness) increases with the cross-sectional diameter host's. Both computational methods (PM6-D3 and DFT methods) clearly show that this increase is non-linear since deformability sharply increased in case of CB[8]. The increased deformability is important, since it suggests that CB[8] (with its increased flexibility) can more readily encapsulate larger guest with higher spatial demands.

	PM6-D3	BLYP-D/def2-SVP
CB[5]	0.27	0.20
CB[6]	0.30	0.23
CB[7]	0.36	0.24
CB[8]	0.50	0.31

Table 9. The dynamic elasticity of CB[n] hosts calculated in COSMO water environment. Values in the table are $RMSD_{average}$ calculated according to eq. 20.

3.5.2.2 The Role of Solvation and Comparison of Different Solvation Models for Guest Desolvation

It is known, that relative solvation energies of neutral small molecules are well reproduced with implicit solvent models such as the COSMO[198] model that was utilized in previous section for CB[*n*] molecules. In similar cases clear distinction has been reported between MM- and QM- based implicit models, showing those based on the QM electron density as superior to others.[23] However, when comparing neutral, monocationic and dicationic molecules the situation is getting more complicated.[199] It is evident from the magnitudes of the main contributions to the ΔG_{calcd} (described later), solvation and interaction energies, that the accurate description of both terms is mandatory. It has been shown that in the rigid molecule case, where the optimized geometry represent both solvated and gas phase conformational ensembles adequately, the implicit solvent models provide reasonable accuracy.[142]

Six of the training set guests had adamantane skeletons, three had diamantane frameworks, one of naphthalene, one of 4,4'-bipiperidine and one of ferrocene with their charges ranging from +1 to +2. Due to guest's charge and geometry differences, it is crucial to select a sufficiently accurate model for the calculation of ΔG_{desolv} associated with complex formation and concomitant guest and host desolvation (*i.e.* removal of the water molecules). The lack of experimental data prevents a direct comparison between theory and experiment. Figure 20 compares desolvation free energies for each guest as calculated by the $COSMO^{PM6}$, $COSMO^{DFT}$ (using BLYP/def2-SVP method), or or SMD implicit solvation techniques. It is noted that the three desolvation free energy values for each particular guest are very similar (with SMD energies being

consistently the highest for each guest in the series). The fact that similar desolvation free energies were obtained by each of the three different methods indicate that any of the method tested here may be used for subsequent calculation of ΔG_{calcd}. Further, inspection of Figure 20 also shows some chemically sensible general trends. Quaternary aminium guests have lower desolvation free energies than primary ammonium guests: (**3** < **2**,**9** and **7**,**10-13** < **8**,**14-16**). Mono-ammonium/aminium guests have markedly lower desolvation free energies than di-ammonium/-aminium guests: **2**,**3**,**9** << (**7**,**8**,**10-16**).

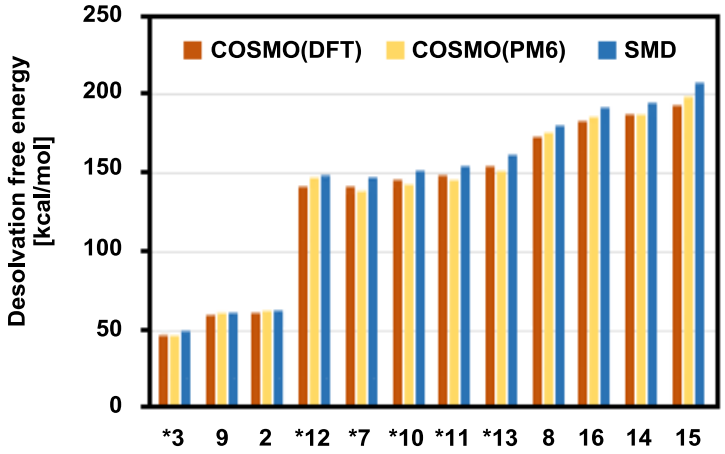


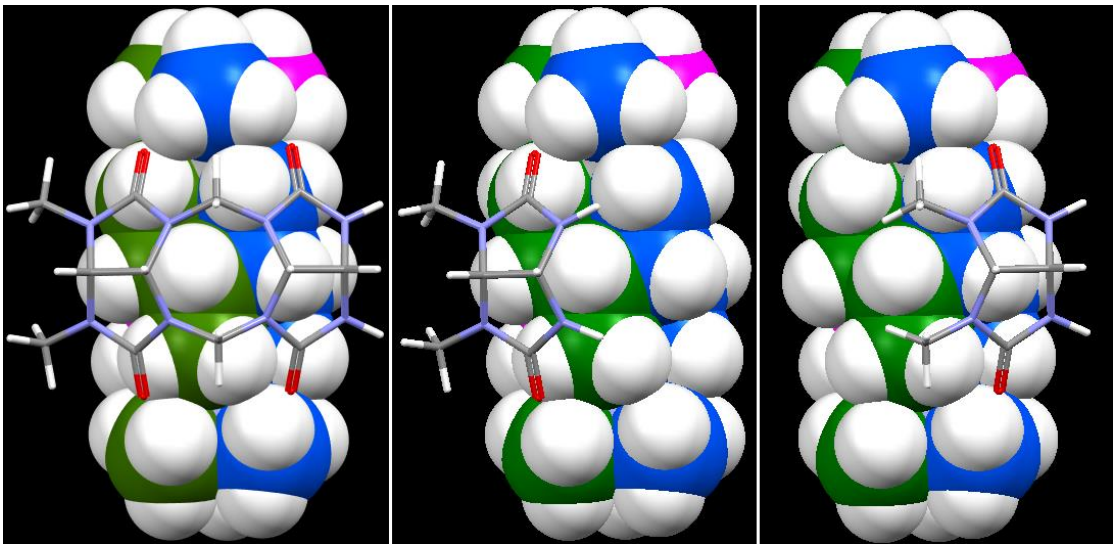
Figure 20. Desolvation free energies calculated by three implicit solvation models: COSMO^{DFT}, COSMO^{PM6} and SMD (all energies are in kcal/mol), the asterisks mark quaternary aminium guests.

3.5.2.3 Accuracy Assessment of ΔE Interaction Energies Using KEM Method

In the previous sections we concluded that the COSMO solvation model is well suited for predictions of solvation energies in host•guest complexes. Next, the accuracy of the BLYP-D3/def2-TZVPP method will be assessed against the more accurate methods such as CCSD(T)/CBS, MP2.5/CBS and FN-DMC.

However, these benchmark methods are limited to the systems containing less than few tens of atoms. Here, we decided to calculate $\Delta E^{MP2.5/CBS}$ interaction energy estimate for CB[7]•**7** assembly therefore necessary additional approximations had to take place. The

$\Delta E^{MP2.5/CBS}$ interaction energy in eq. 9 is determined by two terms. First one, $\Delta E^{MP2/CBS}$ interaction energy, was calculated here using smaller basis sets (non-augmented, cc-pVDZ and cc-pVTZ) while the second, ΔMP2.5 correction term, could not be obtained for the whole CB[7]•**7** complex due to CPU time limitations. However, the fragmentation of the CB[n] host was deemed reasonable since it possessed a periodic structure of glycoluril building blocks ligated by methylene linkers. Therefore it was decided that the ΔMP2.5 correction term would be calculated as a sum the correction terms between the guest **7** and each of seven individual CB[7] host fragments to reduce the input structure's size. Host CB[7] was subdivided into seven methylene-ligated glycoluril units $(CH_2)_{top}(CH_2)_{bottom}$glycoluril, and affixed descriptors **[i]** where i = 0→6. **[0]**,**[1]**,**[2]**,**[3]**,**[4]**,**[5]**, and **[6]**. The fragmented-complex (Fragment-**[0-6]•7**, see Figure 21 for illustration**)** could now be considered to be a suitable set of systems for the so called KEM (see section 2.7) calculations of the ΔMP2.5 correction. Resulting –145.6 kcal/mol $\Delta E^{MP2.5/CBS}$ extrapolated value for the CB[7]•**7** complex was found to be in excellent agreement (within 2%) with that of the –147.6 kcal/mol $\Delta E^{BLYP\text{-}D3/def2\text{-}TZVPP}$ value (used in estimates of ΔG_{calcd}). This close agreement gives full credibility to the DFT-D approach used later for the whole set of host•guest complexes.



Construct-**[12]** Fragment-**[1]** Fragment-**[2]**

Figure 21. Construct-**[12]** (two methylene-linked glycoluril building blocks and a whole **7** guest) subdivided into two fragments **[1]** and **[2]**, each with a **7** guest.

While both, KEM and limited basis set, are mandatory in obtaining the $\Delta E^{MP2.5/CBS}$ interaction energy, it follows that one should study how *e.g.* the number and size of guest fragments or increase of the basis set size would influence the final results. We have tried to address these open questions in next paragraphs.

Let us first examine the accuracy of $\Delta E^{MP2.5/CBS}$ extrapolation. Since glycoluril•(2,3-dimethyl-2,9-diazadecane) complex **18•19** was small, +2 charged, and represented the essential binding motif between CB[n] and N-methylated diamantanes, it was chosen to test the suitability of using time-efficient non-augmented (cc-pVDZ and cc-pVTZ) basis sets for $\Delta E^{MP2/CBS}$ extrapolation (see eq. 8). The comparison with the results obtained when the recommended (but much more time-expensive) augmented basis sets shown only negligible difference of 0.3% relative and 0.08 kcal/mol absolute errors. Further, the resulting $\Delta E^{MP2.5/CBS}$ and $\Delta E^{CCSD(T)/CBS}$ energies (larger, aug-cc-pVDZ, basis set was utilized for calculation of ΔCCSD(T) correction term) were found to be in close agreement as evidenced by their low 0.9% relative and 0.3 kcal/mol absolute error values. This comparison clearly supports the use of the more time-economical extrapolation.

A series of additional calculations on the **18•19** model system were performed to test the fragmentation technique's accuracy in the use of KEM for the calculation of ΔMP2.5's correction term, only the main points will be summarized here. Guest **19** was divided into three-fragment set 1 (**[a]**,**[b]**,**[c]**) and set 2 (**[d]**,**[e]**,**[f]**) differing in the broken C—C bonds. Larger fragments were then fused from the smaller units to provide **[ab]**,**[bc]** or **[de]**,**[ef]** units. ΔMP2.5 correction terms were then calculated for complexes between host **18** and each of the six smaller fragments (and also each of the four larger units). Four ΔMP2.5 energies arise from (**[a]**+**[b]**+**[c]**); (**[d]**+**[e]**+**[f]**); (**[ab]**+**[bc]**–**[b]**} and (**[de]**+**[ef]**–**[e]**) additions. Subtraction of **[b]** or **[e]** from their respective (**[ab]**+**[bc]**) or (**[de]**+**[ef]**) sums insures that individual fragments in the host composite are counted only once. In the first two cases, similar underestimation (0.2 kcal/mol) of ΔMP2.5 correction term was noted for the fragmented guest assemblies versus that for the **18•**unidivided-**19** complex. The error for ΔMP2.5 correction term [compared to ΔCCSD(T) correction term] exhibited the opposite sign (*i.e.* overestimation by 0.18 kcal/mol). As a result, favorable partial error cancellation appeared to occur when the ΔMP2.5 correction term was calculated via the

fragmentation technique. These results strongly suggest that both fragmentation technique used for ΔMP2.5 correction term is reliable.

These promising results with the fragmentation method prompted us to investigate Δ*E* of a methylene-ligated glycoluril double unit ensemble, versus the combined energies of its two single fragments, (see Figure 21). Resulting the $\Delta E^{\text{FN-DMC}}$ (kcal/mol) –53.0±1.2 (Construct-**[12]**]) value was statistically indistinguishable from the –54.9±1.6 (Fragment-**[1]**+**[2]**) sum of –27.4±0.9 (Fragment-[**1**]), and –27.5±0.8 (Fragment-[**2**]]). This finding strongly suggest that the fragmentation method is reliable in terms of both the total and subcomponent Δ*E* energies for the host•guest complexes in this study. However, this conclusion is not general, since $\Delta E^{\text{BLYP-D3/def2TZVPP}}$ comparison for the same system exhibited noticeable errors (10% relative error, see Table 10).

The discrepancies reported in last paragraph are probably caused by method- and fragment-size dependent error addition/cancellation effects. Nonetheless it had been already shown that it is possible to provide a valid $\Delta E^{\text{MP2/CBS}}$ total energy calculated for the whole non-fragmented system and ΔMP2.5 correction term of much smaller magnitude calculated for the fragmented system. The success in using the fragmentation method to obtain the ΔMP2.5 term is probably the result of the more additive character of dispersion energy compared to less additive induction energy dominating the total energy. The authors are of the belief that the above conclusion is general, and not just limited to our host•guest complexes.

	$\Delta E^{\text{FN-DMC}}$ [kca/mol]	$\Delta E^{\text{BLYP-D3/def2-TZVPP}}$ [kcal/mol]
Δ*E*(Fragment-[1])	-28.3	–28.5
Δ*E*(Fragment-[2])	-28.3	–28.4
Δ*E*(from fragments)	-56.5	–56.9
Δ*E*(Construct-[12])	-54.2	–51.4

Table 10. FN-DMC and DFT interaction energies determined from two separate interaction energies calculated on Fragment-[**1**] and Fragment-[**2**] compared to interaction energy of Construct-**[12]**.

3.5.2.4 Complex Formation – Correlation of ΔG_{calcd} and ΔG_{exptl}

It now remains to examine the correlation between the calculated estimate of change of Gibbs free energy (ΔG_{calcd}) in COSMO implicit solvent and ΔG_{exptl} values for 14 complexes measured in buffer solution. The ΔG_{calcd} binding free energy was calculated as the sum of six binding energy parameters: (1) ΔE, (2) the two body dispersion energy $\Delta D3_{2body}$ calculated as proposed by Grimme *et al.*[177], (3) ΔG_{solv}^{COSMO}, (4) the Amber 2nd derivative gas phase entropy $-T\Delta S$, and finally (5,6) the host's and guest's deformation energies E_{def}(host) and E_{def}(guest):

$$\Delta G_{calcd} = \Delta E^{BLYP/def2-TZVPP} + \Delta D3_{2body} + \Delta G_{solv}^{COSMO} - T\Delta S + E_{def}(host) + E_{def}(guest) \quad (21)$$

The reliability of COSMO implicit solvation has been tested on $\Delta G_{desolv(guest)}$ calculated with different implicit solvent models and the accuracy of the DFT-D3 computational method was assesed by *ab initio* MP2.5/CBS method. However, each term in eq. 21 has clear physical meaning and we did not perform any parametrization or adjustment by any means to experimental data. Additionally, three remaining terms of eq. 21, entropy and deformation energies, are calculated in approximative way due to computational economy. Therefore, only relative values will be relied on and we will seek after correlation between theoretical and experimental values. For review about ΔG_{calcd} predictions and detail discussion see ref. [199]. Similar trend (*i.e.* shift of values) was reported by Gilson *et al.*[200]

A list of binding energy parameters (ΔG_{cosmo}, ΔE, ΔG_{solv}, $\Delta E_{electro}$, ΔE_{disp}, $-T\Delta S$, $E_{def}(guest)$, $E_{def}(host)$, ΔG_{calcd}, Ka_{exptl}, and ΔG_{exptl}) are summarized in Table 11. Entropy change for CB[7]•**12** complex formation could not be calculated due to missing parameters for atom Fe. The electrostatic interaction energies ($\Delta E_{electro}$) for the complexes were determined by the Coulomb law and EDA methods. The Figure 22 shows that the $\Delta E_{electro}$(EDA) correlated well (correlation coefficient $\rho^2 = 0.77$) with $\Delta E_{electro}$(NBO). All findings discussed in the following text which utilize the $\Delta E_{electro}$ energies were confirmed when any of those two, $\Delta E_{electro}$(EDA) and $\Delta E_{electro}$, energies were used.

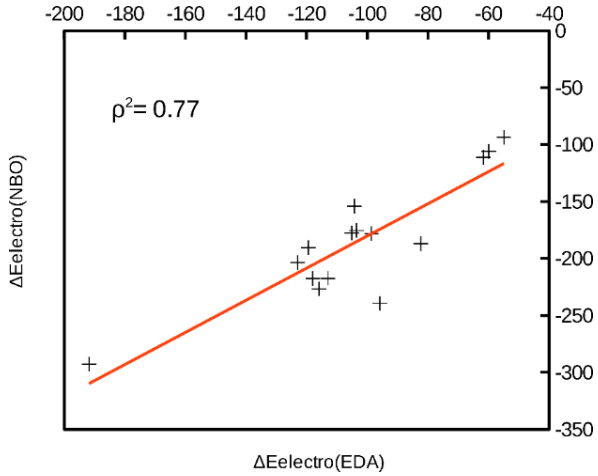


Figure 22. Correlation between $\Delta E_{electro}$ terms calculated by EDA method and coulomb law using NBO charges. (all in kcal/mol)

CB[n]•Guest	ΔG_{cosmo}	ΔE	ΔG_{solv}	$\Delta E_{electro}$	ΔE_{disp}	$-T\Delta S^{a}$	E_{def}(g)	E_{def}(h)	ΔG_{calcd}	$Ka_{exptl}{}^{a}$	$\Delta G_{exptl}{}^{a}$
CB[7]•**2**	–35.2	–92.5	57.3	–111.4	–47.6	4.4	0.8	1.6	–28.3	4.23E+12	–17.2
CB[7]•**3**	–33.9	–84.9	51	–93.4	–53.5	3.3	0.2	0.4	–30.0	1.71E+12	–16.7
CB[7]•**7**	–47.3	–147.6	100.3	–175.5	–72.4	4.5	0.5	1	–41.2	2.00E+15	–20.8
CB[8]•**8**	–40.6	–162.1	121.5	–217.5	–42.1	5.2	1.9	5.4	–28.2	8.30E+11	–16.2
CB[8]•**9**	–28.8	–85.5	56.7	–105.8	–37.2	4.9	1.2	2.2	–20.4	4.33E+11	–15.8
CB[7]•**10**	–35.7	–144.6	108.9	–177.5	–47.9	3.9	0.3	1.7	–29.7	1.70E+11	–15.3
CB[7]•**11**	–31.8	–139.4	107.6	–178.4	–48.9	5.1	0.4	1.5	–24.8	1.90E+11	–15.4
CB[7]•**12**	–38.2	–141.2	102.9	–239.6	–62.5	.[b]	1.1	1.1	.	1.90E+13	–18.1
CB[8]•**13**	–30.7	–136.2	105.5	–186.9	–55.2	5.1	0.9	3.6	–21.2	1.11E+11	–15.0
CB[7]•**14**	–44.2	–173.4	129.2	–226.7	–53.8	4.4	2	5.1	–32.7	1.20E+12	–16.4
CB[8]•**14**	–35.2	–161.0	125.8	–217.4	–38.2	5.2	4.1	10.1	–15.7	4.70E+08	–11.8
CB[7]•**15**	–50.1	–177.1	127	–203.7	–54.5	4.5	3	3.7	–38.8	2.40E+13	–18.2
CB[7]•**16**	–50.7	–168.8	118.1	–190.4	–55.9	4.9	5.9	2.9	–37.0	1.50E+13	–17.9
CB[7]•**17**	–42.7	–144.5	101.8	–154.1	–59.1	4.1	0.7	1.1	–36.8	6.80E+12	–17.5
Prediction											
CB[7]•**23**	–69.8	–262.8	193	–292.9	–81.9	2.1	9.3	2.7	–55.7		

Table 11. Binding energies and other parameters for CB[n]•guest complexes, all values listed are in kcal/mol with the exception Ka_{exptl} in M^{-1}. Footnotes: [a] Temperature 298 K. [b] not calculated.

Figure 23 shows a close correlation of ΔG_{calcd} with ΔG_{exptl} for thirteen complexes with $\rho^2 = 0.84$. Our approach was successful in capturing both the strongest and weakest complexes in our learning set. Since the 4.6(6) $-T\Delta S_{mean}$ value, with its low 0.6 estimated standard deviation (esd) for the 13 complexes, is small enough the entropy can be treated as a constant for a first approximation (see Figure 24a). This proved to be valid since after the addition of ferrocene complex CB[7]•**12** (blue error bar) the plot kept correlation coefficient $\rho^2 = 0.84$. Clearly, the accuracy of entropy calculations is sufficient enough and do not represent a significant term for the overall correlationship's fidelity.

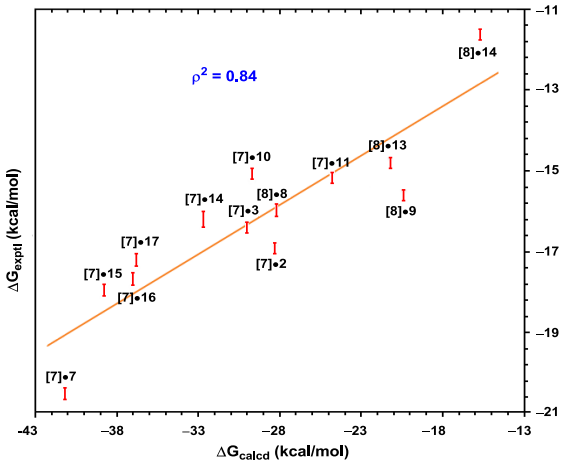


Figure 23. Correlation between theoretical ΔG_{calcd} and experimental ΔG_{exptl} (the equation of the fit: $\Delta G_{exptl} = 0.247\Delta G_{calcd} - 9.179$, $\rho^2 = 0.84$, n = 13).

Next, we investigate the role of the individual terms in eq. 21, besides the entropy discussed above, for quality correlationship with ΔG_{exptl}. It is known that for large scale systems the 3-body dispersion energy improve both, the total dispersion and interaction energy.[177,201] The Figure 24b shows slightly reduced correlationship with correlation coefficient lowered to $\rho^2 = 0.80$. It can be explained by the fact that the presented host guest systems are not large enough. On the other hand, both, the solvation and dispersion terms proved to be mandatory, because the correlation was completely lost (not shown) when either of these two terms was omitted from ΔG_{calcd}.

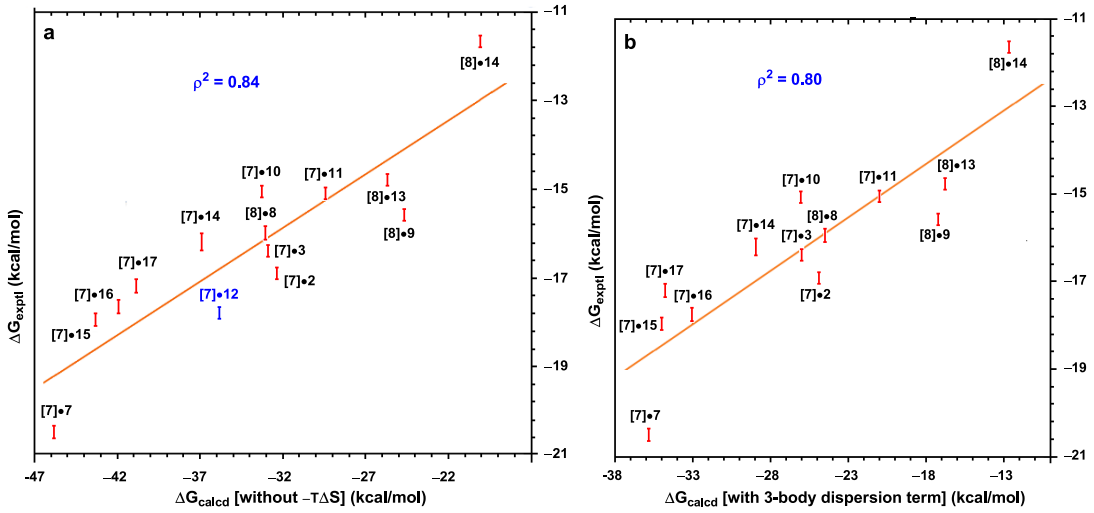


Figure 24. Correlation between ΔG_{calcd} and ΔG_{exptl} after (a) entropy term exclusion and addition of CB[7]•**12** complex (in blue). (b) addition of 3-body dispersion term.

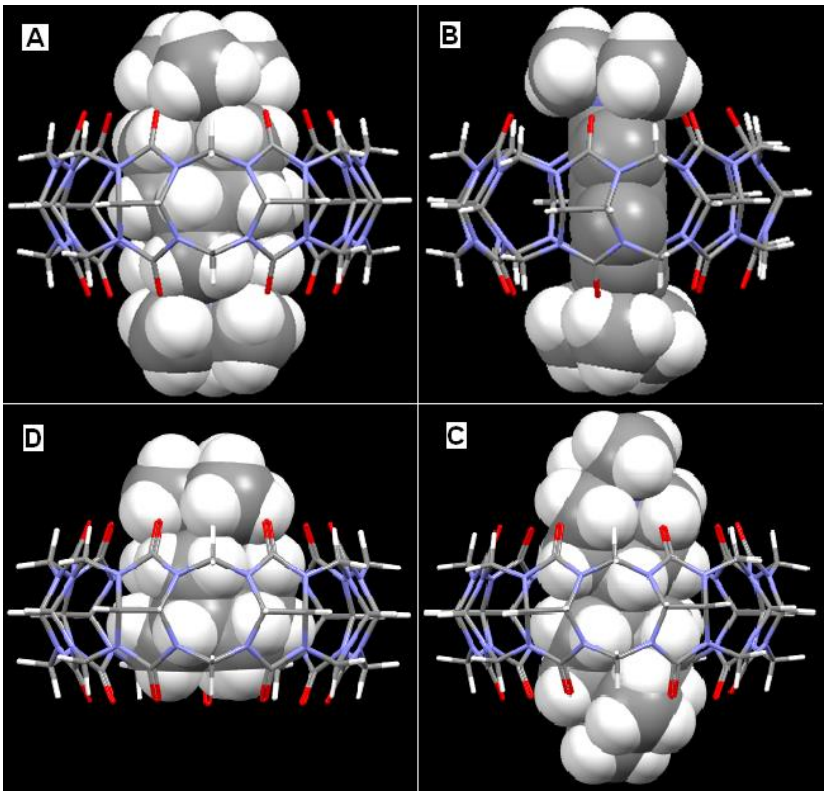


Figure 25. Illustration of four CB[7] molecules complexed with di-substituted diamantane (A, code name: CB[7]•**7**), naphthalene (B, code name: CB[7]•**10**), virtual complex CB[7]•Diam-4,9-diCMe$_3$ (C, CB[7]•3t-But) and mono-substituted adamantane (D, code name: CB[7]•**3**).

In order to understand the role of charged —$^{+}NR_3$ groups, space filling (dispersion) and electrostatic interactions we compared the binding modes for 4 CB[7]•guests containing tertiary ammonium functional groups (shown in Figure 25).

The "good fit" and its binding consequences in terms of significant dispersion forces are immediately apparent when one compares diamantane (A, CB[7]•**7**, ΔE_{disp} = –72.4 kcal/mol) and naphthalene (B, CB[7]•**10**, ΔE_{disp} = –47.9 kcal/mol) scaffolds. Additionally, significant difference in space filling dispersion forces were observed also for the former (diamantane) versus adamantane (B, CB[7]•**3** complex) derivates. Both these finding provides insight into the role of dispersion in the exceptionally high binding affinities of substituted diamantine derivates. Next, similar $N^{+}\cdots N^{+}$ distances (7.94 Å, 8.07 Å and 7.51 Å) accompanied with marginal difference also in $\Delta E_{electro}$ (–175.5 kcal/mol, –177.5 kcal/mol and –178.4 kcal/mol) were determined for CB[7]•**7**, CB[7]•**10** and CB[7]•**11**, respectively. Showing marginal difference when guests of different sizes and the same charge are compared.

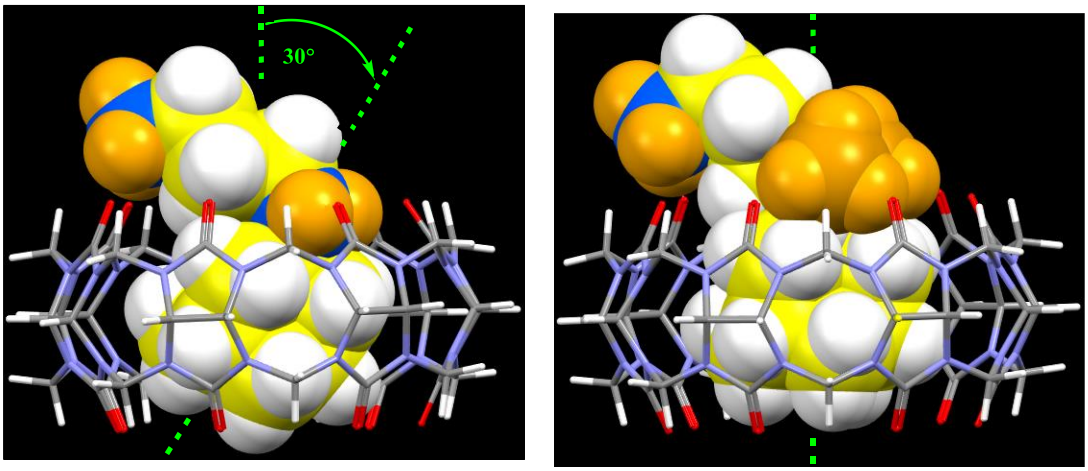
Finally, in case of Ada-2,6-di(NH_3) **14** guest two complexes (with CB[7] and CB[8] hosts) were previously reported and calculated here for the first time. The larger host complex suffers from only 4% and 3 % decrease in $\Delta E_{electro}$ and ΔG_{solv}, respectively. While ΔE_{disp} is reduced by 29% resulting in substantial decrease in experimental equilibrium binding constant from $(1.2\pm0.4)\times10^{12}$ M^{-1} K_a to $(4.7\pm1.2)\times10^{8}$ M^{-1} K_a.

3.5.2.5 Host•guest Complexes Containing Amino Loops

We applied the same methodology as utilized previously for maximizing the strength of the binding through design of new guest molecules through reaching unused CB[n] host sites. In a similar way as the binding motif of the primary amines was previously enhanced via methylation to tertiary binding motif reaching all 7 carboxyl oxygens as depicted in Figure 18.

Here, two CB[7]•adamantane-1-NH_2R^{+1} complexes (CB[7]•**15**,CB[7]•**16** containing an R = $[–(CH_2)_nNH_3]^{+1}$ primary amino loop [where n = 2 (ethano) or 3 (propano)]) and Ada-1-$NMe_2(CH_2)_3NH_3$ (CB[7]•**17**) were studied for the first time. Additional loops formed $CH_2)_n\underline{\mathbf{N}}–H^{\delta+}\cdots{}^{-\delta}\underline{\mathbf{O}}$=C 2.7 Å hydrogen bonds that are in accord with an increase of ΔG binding free energy increment versus the parent complex (see Figure 26). The

increase of both, $\Delta E_{electro}$ and ΔE_{disp} (see Table 11), is substantial and is partially compensated by increased solvation plus the guest's deformation energy.

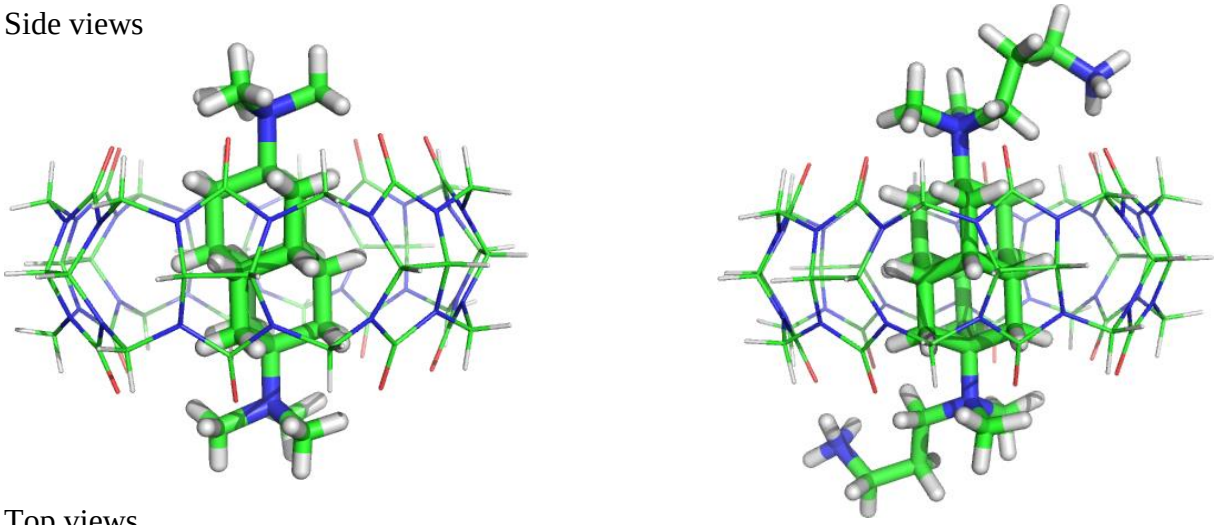


21 **22**

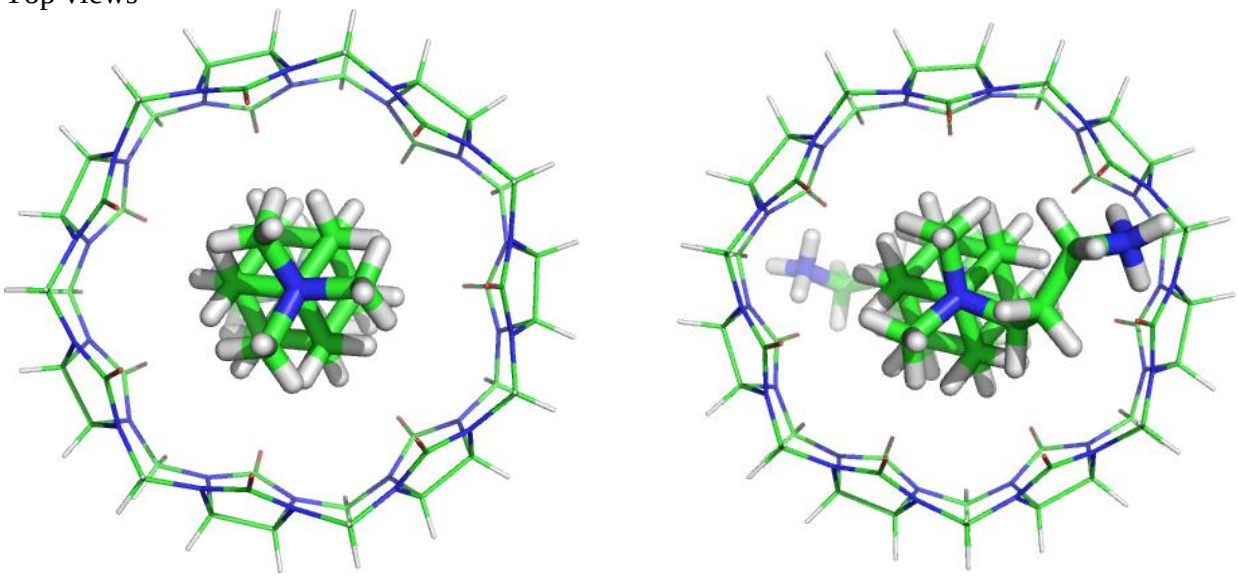
Figure 26. Comparison of primary and tertiary amino binding motif combined with addition of amino loops. (CB[7]•**16** and CB[7]•**17**, respectively)

The Ada-1-$NMe_2(CH_2)_3NH_3$ (**17**) guest represents success in K_a enhancement of CB[7]•quaternary-guest by a $[–(CH_2)_3NH_3]^{+1}$ loop (see model **22**). The experimental and theoretical observation that K_a values of quaternary-aminium guests are augmented by $[–(CH_2)_nNH_3]+1$ loops to a lower extent than those of primary-ammonium guests (compare CB[7]•**2** with CB[7]•**16** and CB[7]•**3** with CB[7]•**17**) must still be proven for larger number of guests. Finally, the Diam-4,9-di(NMe_2propanoNH_3) (**23)** is now being prepared to test this hypothesis. It remains to be seen if a predicted new record ultra-high value will experimentally be found for complex CB[7]•**23** (see Table 11 and Figure 27 for comparison with the current world record CB[7]•**8**).

Side views



Top views

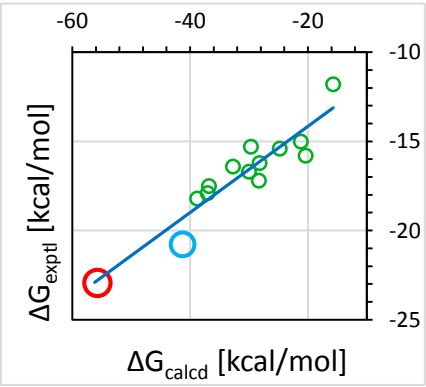


CB[7]•8

World record binder

K_a = (7.2 ± 0.9) x 10^{17} M^{-1}

ΔG_{calcd} = – 41.2 kcal/mol



CB[7]•23

New Proposed Guest

K_a to be measured

ΔG_{calcd} = – 55.7 kcal/mol

Figure 27. Visual comparison of current world record binder on the left hand side with the new proposed guest enriched by two additional amino loop.

3.5.3 Conclusions

In our last project we described properties of isolated host and guest molecules, discussed the accuracy of solvation and interaction terms in ΔG_{calcd} predictions and assessed the correlation between ΔG_{calcd} and ΔG_{exptl}. Finally, we described differences between primary and tertiary binding motifs leading to proposal of loop addition as a potential way for enhancement of binding strength. Following key findings have been highlighted:

- Sharp increase in deformability of CB[*n*] molecules has been found when passing from CB[7] to CB[8] hosts.

- Previously reported solvation properties of CB[*n*] molecules has been well reproduced by explicit (WaterMap) and implicit (COSMO and SMD) solvent models.

- Presence of high-energy water molecules was not required for modeling the CB[*n*]-host•guest complexes investigated in our study.

- BLYP-D3/def2-TZVPP interaction energy has been found in close agreement with MP2.5/CBS value (utilizing the kernel energy method).

- Nice correlation between theoretical (ΔG_{calcd}) and experimental (ΔG_{explt}) changes of Gibbs free energies has been reported here. ($\rho^2 = 0.84$)

- Prediction has been made that Diam-4,9-di(NMe_2propanoNH_3) could become next world record binder. Currently, the synthesis work is in progress in group of Pavel Majer.

4 Concluding Remarks

In this thesis, research concerning noncovalently driven recognition processes in both natural and artificial complexes has been presented. Every careful scientists must make an informed choice of the applied methods based on the understanding of the chemical system and drawbacks and merits of various methods. In those cases when binding is highly probably primarily enthalpy-driven, the QM description brings many advantages in description of many body effects, electron and proton transfers, formation and dissociation of a covalent bond *etc*. Major part of the effort here was made to obtain a detail understanding of the X-ray structures of host•guest systems, method performance and development of empirical dispersion for DFT methodology. While the studies of binding preferences between DNA and proteins must be considered as the initial investigation to set the stage for more elaborate analysis of larger molecular fragments (*e.g.* containing parts of the sugar−phosphate backbone) especially in solvents. The same is true in case of self-assembled artificial cage binding choline and acetylcholine guest molecules.

We have described the performance of various SQM method (used nowadays for ΔG predictions in protein•ligand systems). We showed that, when one wants to achieve accuracy of 1 kcal/mol and higher across the wide range of exotic noncovalent interactions, the less approximate methods need to be chosen. One of such examples is DFT methodology accompanied with rage of post-HF methods used in the rest of the presented projects. It has been shown that it is mandatory to compensate the lack of dispersion energy in DFT methods, therefore we developed here *a posteriori* calculated empirical correction term for small basis sets.

Next, we quantitatively assessed the binding preferences in protein•DNA complexes. It has been proven that amino acid−base geometries capable of one-to-one amino

acid−base recognition correspond to unique energy minima with interaction energies distinct from the rest of the distribution.

This thesis has demonstrated that the development in the field of computational chemistry has reached the point when it can truly help us to understand and interpret the experimental data. Our approach determined that choline guest is bound to self assembled triple helicate aromatic cage 2.8 kcal/mol stronger (ΔG) than acetylcholine corresponding to K(Ch+)/K(ACh+) = 109 selectivity. The respective experimental value is 20 agreeing well with our prediction reported together with decomposition into various physically meaningful terms.

Finally, results concerning CB[n] host•guest systems have been presented. The training set of 13 complexes based on X-ray crystal structures was studied including several new suggested derivatives containing amino loops. A new guest, namely Diam-4,9-di(NMe_2propanoNH_3) has been proposed as the potential new world record binder to CB[7]. Additionally, the predicted substantial increase of flexibility of CB[8] indicates that it can more readily encapsulate larger and promising guests with higher spatial demands. It still remains to be proven experimentally however our recent pilot study of several isodiamantane complexes suggests this notion (attachment H). It has been reported that the stability of complexes with CB[8] equals the most stable complexes containing CB[7] which are known as the strongest artificial complexes formed in water.

5 Bibliography

(1) Laskowski, R. A.; Thornton, J. M. Understanding the molecular machinery of genetics through 3D structures. *Nature Reviews Genetics* **2008**, *9*, 141-151.

(2) Liu, S. M.; Ruspic, C.; Mukhopadhyay, P.; Chakrabarti, S.; Zavalij, P. Y.; Isaacs, L. The cucurbit n uril family: Prime components for self-sorting systems. *Journal of the American Chemical Society* **2005**, *127*, 15959-15967.

(3) Bergmann, N. M.; Peppas, N. A. Molecularly imprinted polymers with specific recognition for macromolecules and proteins. *Prog. Polym. Sci.* **2008**, *33*, 271-288.

(4) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures. *J Chem Theory Comput* **2009**, *5*, 982-992.

(5) Gaj, T.; Gersbach, C. A.; Barbas, C. F., III. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology* **2013**, *31*, 397-405.

(6) Navrátil, V.; Schimer, J.; Tykvart, J.; Knedlik, T.; Vik, V.; Majer, P.; Konvalinka, J.; Sacha, P. DNA-linked Inhibitor Antibody Assay (DIANA) for sensitive and selective enzyme detection and inhibitor screening. *Nucl Acids Res* **2016**, *45*, e10.

(7) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95-115.

(8) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10360-10365.

(9) Wilcken, R.; Liu, X. R.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant p53. *Journal of the American Chemical Society* **2012**, *134*, 6810-6818.

(10) Haldar, S.; Kolar, M.; Sedlak, R.; Hobza, P. Adsorption of Organic Electron Acceptors on Graphene-like Molecules: Quantum Chemical and Molecular Mechanical Study. *J. Phys. Chem. C* **2012**, *116*, 25328-25336.

(11) Rezac, J.; Hobza, P. Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications. *Chemical Reviews* **2016**, *116*, 5038-5071.

(12) van der Veken, B. J.; Herrebout, W. A.; Szostak, R.; Shchepkin, D. N.; Havlas, Z.; Hobza, P. The nature of improper, blue-shifting hydrogen bonding verified experimentally. *Journal of the American Chemical Society* **2001**, *123*, 12290-12293.

(13) Pimentel, G. C.; McClella.Al. HYDROGEN BONDING. *Annu. Rev. Phys. Chem.* **1971**, *22*, 347-&.

(14) Arunan, E.; Desiraju, G. R.; Klein, R. A.; Sadlej, J.; Scheiner, S.; Alkorta, I.; Clary, D. C.; Crabtree, R. H.; Dannenberg, J. J.; Hobza, P.; Kjaergaard, H. G.; Legon, A. C.; Mennucci, B.; Nesbitt, D.

J. Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure Appl. Chem.* **2011**, *83*, 1637-1641.

(15) El Kerdawy, A.; Murray, J. S.; Politzer, P.; Bleiziffer, P.; Hesselmann, A.; Goerling, A.; Clark, T. Directional Noncovalent Interactions: Repulsion and Dispersion. *J Chem Theory Comput* **2013**, *9*, 2264-2275.

(16) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the sigma-hole. *Journal of Molecular Modeling* **2007**, *13*, 291-296.

(17) Kolar, M.; Kubar, T.; Hobza, P. On the Role of London Dispersion Forces in Biomolecular Structure Determination. *J. Phys. Chem. B* **2011**, *115*, 8038-8046.

(18) Zhou, H. X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chemical Reviews* **2009**, *109*, 4092-4107.

(19) Yilmazer, N. D.; Korth, M. Recent Progress in Treating Protein-Ligand Interactions with Quantum-Mechanical Methods. *Int. J. Mol. Sci.* **2016**, *17*, 12.

(20) Paton, R. S.; Goodman, J. M. Hydrogen Bonding and pi-Stacking: How Reliable are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *J. Chem Inf. Model.* **2009**, *49*, 944-955.

(21) Antony, J.; Sure, R.; Grimme, S. Using dispersion-corrected density functional theory to understand supramolecular binding thermodynamics. *Chemical Communications* **2015**, *51*, 1764-1774.

(22) Silverstein, K. A. T.; Haymet, A. D. J.; Dill, K. A. A simple model of water and the hydrophobic effect. *Journal of the American Chemical Society* **1998**, *120*, 3166-3175.

(23) Kolar, M.; Fanfrlik, J.; Lepsik, M.; Forti, F.; Luque, F. J.; Hobza, P. Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches. *J. Phys. Chem. B* **2013**, *117*, 5950-5962.

(24) Spek, A. L. Structure validation in chemical crystallography. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **2009**, *65*, 148-155.

(25) Bragg, W. L. The Diffraction of Short Electromagnetic Waves by a Crystal. *Proceedings of the Cambridge Philosophical Society* **1913**, *17*, 43–57.

(26) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect and Prospect. *Angew. Chem.-Int. Edit.* **2014**, *53*, 662-671.

(27) Brammer, L. Developments in inorganic crystal engineering. *Chem. Soc. Rev.* **2004**, *33*, 476-489.

(28) Betts, K. Crystallography: Understanding the Nature of Chemical Bonds and Molecular Structure. *American Chemical Society* **2014**.

(29) Sayre, D. X-ray crystallography: The past and present of the phase problem. *Struct. Chem.* **2002**, *13*, 81-96.

(30) Bonet, F.; Grugeon, S.; Urbina, R. H.; Tekaia-Elhsissen, K.; Tarascon, J. M. In situ deposition of silver and palladium nanoparticles prepared by the polyol process, and their performance as catalytic converters of automobile exhaust gases. *Solid State Sci.* **2002**, *4*, 665-670.

(31) Lee, S. H.; Bhattacharyya, S. S.; Jin, H. S.; Jeong, K. U. Devices and materials for high-performance mobile liquid crystal displays. *J. Mater. Chem.* **2012**, *22*, 11893-11903.

(32) Carpenter, E. P.; Beis, K.; Cameron, A. D.; Iwata, S. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology* **2008**, *18*, 581-586.

(33) Dauber, P.; Hagler, A. T. CRYSTAL PACKING, HYDROGEN-BONDING, AND THE EFFECT OF CRYSTAL FORCES ON MOLECULAR-CONFORMATION. *Accounts of Chemical Research* **1980**, *13*, 105-112.

(34) Yu, N.; Hayik, S. A.; Wang, B.; Liao, N.; Reynolds, C. H.; Merz, K. M. Assigning the protonation states of the key aspartates in beta-secretase using QM/MM X-ray structure refinement. *J Chem Theory Comput* **2006**, *2*, 1057-1069.

(35) Fanfrlik, J.; Prada, A.; Padelkova, Z.; Pecina, A.; Machacek, J.; Lepsik, M.; Holub, J.; Ruzicka, A.; Hnyk, D.; Hobza, P. The Dominant Role of Chalcogen Bonding in the Crystal Packing of 2D/3D Aromatics. *Angew. Chem.-Int. Edit.* **2014**, *53*, 10139-10142.

(36) Zhu, C.; Byers, K. J. R. P.; McCord, R. P.; Shi, Z.; Berger, M. F.; Newburger, D. E.; Saulrieta, K.; Smith, Z.; Shah, M. V.; Radhakrishnan, M.; Philippakis, A. A.; Hu, Y.; De Masi, F.; Pacek, M.; Rolfs, A.; Murthy, T.; LaBaer, J.; Bulyk, M. L. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research* **2009**, *19*, 556-566.

(37) Seeman, N. C.; Rosenberg, J. M.; Rich, A. SEQUENCE-SPECIFIC RECOGNITION OF DOUBLE HELICAL NUCLEIC-ACIDS BY PROTEINS. *Proc. Natl. Acad. Sci. U. S. A.* **1976**, *73*, 804-808.

(38) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S.: Origins of Specificity in Protein-DNA Recognition. In *Annual Review of Biochemistry, Vol 79*; Kornberg, R. D., Raetz, C. R. H., Rothman, J. E., Thorner, J. W., Eds.; Annual Review of Biochemistry, 2010; Vol. 79; pp 233-269.

(39) Rohs, R.; West, S. M.; Liu, P.; Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Current Opinion in Structural Biology* **2009**, *19*, 171-177.

(40) Kim, Y. C.; Geiger, J. H.; Hahn, S.; Sigler, P. B. CRYSTAL-STRUCTURE OF A YEAST TBP TATA-BOX COMPLEX. *Nature* **1993**, *365*, 512-520.

(41) Otwinowski, Z.; Schevitz, R. W.; Zhang, R. G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. CRYSTAL-STRUCTURE OF TRP REPRESSOR OPERATOR COMPLEX AT ATOMIC RESOLUTION. *Nature* **1988**, *335*, 321-329.

(42) Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. Protein-DNA interactions: A structural analysis. *Journal of Molecular Biology* **1999**, *287*, 877-896.

(43) Anzenbacher, P.; Anzenbacherova, E. Cytochromes P450 and metabolism of xenobiotics. *Cell. Mol. Life Sci.* **2001**, *58*, 737-747.

(44) Friedman, F. K.; Robinson, R. C.; Dai, R. Molecular modeling of mammalian cytochrome P450s. *Front. Biosci.* **2004**, *9*, 2796-2806.

(45) Hudecek, J.; Anzenbacher, P. SECONDARY STRUCTURE PREDICTION OF LIVER MICROSOMAL CYTOCHROME-P-450 - PROPOSED MODEL OF SPATIAL ARRANGEMENT IN A MEMBRANE. *Biochimica Et Biophysica Acta* **1988**, *955*, 361-370.

(46) Berka, K.; Hendrychova, T.; Anzenbacher, P.; Otyepka, M. Membrane Position of Ibuprofen Agrees with Suggested Access Path Entrance to Cytochrome P450 2C9 Active Site. *Journal of Physical Chemistry A* **2011**, *115*, 11248-11255.

(47) Koritsanszky, T. S.; Coppens, P. Chemical applications of X-ray charge-density analysis. *Chemical Reviews* **2001**, *101*, 1583-1627.

(48) Borbulevych, O.; Martin, R. I.; Tickle, I. J.; Westerhoff, L. M. XModeScore: a novel method for accurate protonation/tautomer-state determination using quantum-mechanically driven macromolecular X-ray crystallographic refinement. *Acta Crystallogr. Sect. D-Struct. Biol.* **2016**, *72*, 586-598.

(49) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M., Jr. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J Chem Theory Comput* **2011**, *7*, 790-797.

(50) Kehlet, B.; Logg, A. A posteriori error analysis of round-off errors in the numerical solution of ordinary differential equations. *ArXiv e-prints* **2015**.

(51) Simova, L.; Rezac, J.; Hobza, P. Convergence of the Interaction Energies in Noncovalent Complexes in the Coupled-Cluster Methods Up to Full Configuration Interaction. *J Chem Theory Comput* **2013**, *9*, 3420-3428.

(52) Hobza, P.; Zahradnik, R. INTERMOLECULAR INTERACTIONS BETWEEN MEDIUM-SIZED SYSTEMS - NONEMPIRICAL AND EMPIRICAL CALCULATIONS OF INTERACTION ENERGIES - SUCCESSES AND FAILURES. *Chemical Reviews* **1988**, *88*, 871-897.

(53) Sedlak, R.; Janowski, T.; Pitonak, M.; Rezac, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *J Chem Theory Comput* **2013**, *9*, 3364-3374.

(54) Haldar, S.; Gnanasekaran, R.; Hobza, P. A comparison of ab initio quantum-mechanical and experimental D-0 binding energies of eleven H-bonded and eleven dispersion-bound complexes. *Phys Chem Chem Phys* **2015**, *17*, 26645-26652.

(55) Rezac, J.; Riley, K. E.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J Chem Theory Comput* **2011**, *7*, 2427-2438.

(56) Hobza, P. Calculations on Noncovalent Interactions and Databases of Benchmark Interaction Energies. *Accounts of Chemical Research* **2012**, *45*, 663-672.

(57) Rezac, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J Chem Theory Comput* **2012**, *8*, 141-151.

(58) Faver, J. C.; Benson, M. L.; He, X. A.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J Chem Theory Comput* **2011**, *7*, 790-797.

(59) Rezac, J.; Dubecky, M.; Jurecka, P.; Hobza, P. Extensions and applications of the A24 data set of accurate interaction energies. *Phys Chem Chem Phys* **2015**, *17*, 19268-19277.

(60) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys Chem Chem Phys* **2006**, *8*, 1985-1993.

(61) Grafova, L.; Pitonak, M.; Rezac, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *J Chem Theory Comput* **2010**, *6*, 2365-2376.

(62) Rezac, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J Chem Theory Comput* **2012**, *8*, 4285-4292.

(63) Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2012**, *2*, 254-272.

(64) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A., III. Evaluation of B3LYP, X3LYP, and M06-Class Density Functionals for Predicting the Binding Energies of Neutral, Protonated, and Deprotonated Water Clusters. *J Chem Theory Comput* **2009**, *5*, 1016-1026.

(65) Ramirez, F.; Hadad, C. Z.; Guerra, D.; David, J.; Restrepo, A. Structural studies of the water pentamer. *Chem Phys Lett* **2011**, *507*, 229-233.

(66) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. Benchmark database on isolated small peptides containing an aromatic side chain: comparison between wave function and density functional theory methods and empirical force field. *Phys Chem Chem Phys* **2008**, *10*, 2747-2757.

(67) Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389*, 457-484.

(68) Van Voorhis, T.; Kowalczyk, T.; Kaduk, B.; Wang, L. P.; Cheng, C. L.; Wu, Q.: The Diabatic Picture of Electron Transfer, Reaction Barriers, and Molecular Dynamics. In *Annual Review of Physical Chemistry, Vol 61*; Leone, S. R., Cremer, P. S., Groves, J. T., Johnson, M. A., Richmond, G., Eds.; Annual Review of Physical Chemistry; Annual Reviews: Palo Alto, 2010; Vol. 61; pp 149-170.

(69) Stuchebrukhov, A. Tunneling Time and the Breakdown of Born-Oppenheimer Approximation. *J. Phys. Chem. B* **2016**, *120*, 1408-1417.

(70) Schwenke, D. W. Towards accurate ab initio predictions of the vibrational spectrum of methane. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.* **2002**, *58*, 849-861.

(71) Rezac, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J Chem Theory Comput* **2013**, *9*, 2151-2155.

(72) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-d group 16-18 elements. *J. Chem. Phys.* **2003**, *119*, 11113-11123.

(73) Ahuja, R.; Blomqvist, A.; Larsson, P.; Pyykko, P.; Zaleski-Ejgierd, P. Relativity and the Lead-Acid Battery. *Phys. Rev. Lett.* **2011**, *106*, 4.

(74) Lowdin, P. O. QUANTUM THEORY OF MANY-PARTICLE SYSTEMS .3. EXTENSION OF THE HARTREE-FOCK SCHEME TO INCLUDE DEGENERATE SYSTEMS AND CORRELATION EFFECTS. *Physical Review* **1955**, *97*, 1509-1520.

(75) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N-2, and H2O. *Chem Phys Lett* **1998**, *286*, 243-252.

(76) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639-9646.

(77) Peterson, K. A.; Dunning, T. H. Accurate correlation consistent basis sets for molecular core-valence correlation effects: The second row atoms Al-Ar, and the first row atoms B-Ne revisited. *J. Chem. Phys.* **2002**, *117*, 10548-10560.

(78) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618.

(79) Sedlak, R.; Riley, K. E.; Rezac, J.; Pitonak, M.; Hobza, P. MP2.5 and MP2.X: Approaching CCSD(T) Quality Description of Noncovalent Interaction at the Cost of a Single CCSD Iteration. *Chemphyschem* **2013**, *14*, 698-707.

(80) Woon, D. E.; Dunning, T. H. GAUSSIAN-BASIS SETS FOR USE IN CORRELATED MOLECULAR CALCULATIONS .4. CALCULATION OF STATIC ELECTRICAL RESPONSE PROPERTIES. *J. Chem. Phys.* **1994**, *100*, 2975-2988.

(81) Harihara.Pc; Pople, J. A. INFLUENCE OF POLARIZATION FUNCTIONS ON MOLECULAR-ORBITAL HYDROGENATION ENERGIES. *Theoretica Chimica Acta* **1973**, *28*, 213-222.

(82) Hehre, W. J.; Ditchfield, R.; Pople, J. A. SELF-CONSISTENT MOLECULAR-ORBITAL METHODS .12. FURTHER EXTENSIONS OF GAUSSIAN-TYPE BASIS SETS FOR USE IN MOLECULAR-ORBITAL STUDIES OF ORGANIC-MOLECULES. *J. Chem. Phys.* **1972**, *56*, 2257-+.

(83) Kroonbatenburg, L. M. J.; Vanduijneveldt, F. B. THE USE OF A MOMENT-OPTIMIZED DZP BASIS SET FOR DESCRIBING THE INTERACTION IN THE WATER DIMER. *Journal of Molecular Structure-Theochem* **1985**, *22*, 185-199.

(84) Grimme, S. Improved second-order Moller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* **2003**, *118*, 9095-9102.

(85) Grimme, S.; Goerigk, L.; Fink, R. F. Spin-component-scaled electron correlation methods. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2012**, *2*, 886-906.

(86) Distasio, R. A., Jr.; Head-Gordon, M. Optimized spin-component scaled second-order Moller-Plesset perturbation theory for intermolecular interaction energies. *Mol. Phys.* **2007**, *105*, 1073-1083.

(87) Hill, J. G.; Platts, J. A. Spin-component scaling methods for weak and stacking interactions. *J Chem Theory Comput* **2007**, *3*, 80-85.

(88) Kutzelnigg, W.; Klopper, W. WAVE-FUNCTIONS WITH TERMS LINEAR IN THE INTERELECTRONIC COORDINATES TO TAKE CARE OF THE CORRELATION CUSP .1. GENERAL-THEORY. *J. Chem. Phys.* **1991**, *94*, 1985-2001.

(89) Ten-No, S. Initiation of explicitly correlated Slater-type geminal theory. *Chem Phys Lett* **2004**, *398*, 56-61.

(90) May, A. J.; Manby, F. R. An explicitly correlated second order Moller-Plesset theory using a frozen Gaussian geminal. *J. Chem. Phys.* **2004**, *121*, 4479-4485.

(91) Ten-no, S.; Noga, J. Explicitly correlated electronic structure theory from R12/F12 ansatze. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2012**, *2*, 114-125.

(92) Cizek, J. ON CORRELATION PROBLEM IN ATOMIC AND MOLECULAR SYSTEMS . CALCULATION OF WAVEFUNCTION COMPONENTS IN URSELL-TYPE EXPANSION USING QUANTUM-FIELD THEORETICAL METHODS. *J. Chem. Phys.* **1966**, *45*, 4256-+.

(93) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Headgordon, M. A 5TH-ORDER PERTURBATION COMPARISON OF ELECTRON CORRELATION THEORIES. *Chem Phys Lett* **1989**, *157*, 479-483.

(94) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. Improvement of the coupled-cluster singles and doubles method via scaling same- and opposite-spin components of the double excitation correlation energy. *J. Chem. Phys.* **2008**, *128*.

(95) Pitonak, M.; Rezac, J.; Hobza, P. Spin-component scaled coupled-clusters singles and doubles optimized towards calculation of noncovalent interactions. *Phys Chem Chem Phys* **2010**, *12*, 9611-9614.

(96) Dubecky, M. QUANTUM MONTE CARLO FOR NONCOVALENT INTERACTIONS: A TUTORIAL REVIEW. *Acta Physica Slovaca* **2014**, *64*, 501-575.

(97) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. GENERAL ATOMIC AND MOLECULAR ELECTRONIC-STRUCTURE SYSTEM. *J. Comput. Chem.* **1993**, *14*, 1347-1363.

(98) Wagner, L. K.; Bajdich, M.; Mitas, L. QWalk: A quantum Monte Carlo program for electronic structure. *Journal of Computational Physics* **2009**, *228*, 3390-3404.

(99) Burkatzki, M.; Filippi, C.; Dolg, M. Energy-consistent pseudopotentials for quantum monte carlo calculations. *J. Chem. Phys.* **2007**, *126*.

(100) Dubecky, M.; Jurecka, P.; Derian, R.; Hobza, P.; Otyepka, M.; Mitas, L. Quantum Monte Carlo Methods Describe Noncovalent Interactions with Subchemical Accuracy. *J Chem Theory Comput* **2013**, *9*, 4287-4292.

(101) Hohenberg, P.; Kohn, W. INHOMOGENEOUS ELECTRON GAS. *Phys. Rev. B* **1964**, *136*, B864-+.

(102) Medvedev, M. G.; Bushmarinov, I. S.; Sun, J. W.; Perdew, J. P.; Lyssenko, K. A. Density functional theory is straying from the path toward the exact functional. *Science* **2017**, *355*, 49-+.

(103) Mardirossian, N.; Head-Gordon, M. omega B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **2016**, *144*, 23.

(104) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc. A-Math. Phys. Eng. Sci.* **2014**, *372*, 52.

(105) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 19.

(106) Becke, A. D.; Johnson, E. R. A density-functional model of the dispersion interaction. *J. Chem. Phys.* **2005**, *123*.

(107) Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions. *J. Chem. Phys.* **2005**, *123*.

(108) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the ospersion interaction. *J. Chem. Phys.* **2005**, *122*.

(109) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456-1465.

(110) Dobes, P.; Fanfrlik, J.; Rezac, J.; Otyepka, M.; Hobza, P. Transferable scoring function based on semiempirical quantum mechanical PM6-DH2 method: CDK2 with 15 structurally diverse inhibitors. *Journal of Computer-Aided Molecular Design* **2011**, *25*, 223-235.

(111) Fanfrlik, J.; Bronowska, A. K.; Rezac, J.; Prenosil, 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.

(112) Stewart, J. J. P. Application of the PM6 method to modeling proteins. *Journal of Molecular Modeling* **2009**, *15*, 765-805.

(113) Leverentz, H. R.; Qi, H. W.; Truhlar, D. G. Assessing the Accuracy of Density Functional and Semiempirical Wave Function Methods for Water Nanoparticles: Comparing Binding and Relative Energies of (H2O)(16) and (H2O)(17) to CCSD(T) Results. *J Chem Theory Comput* **2013**, *9*, 995-1006.

(114) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13*, 1173-1213.

(115) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods. *J Chem Theory Comput* **2010**, *6*, 344-352.

(116) Rezac, J. Cuby: An Integrative Framework for Computational Chemistry. *J. Comput. Chem.* **2016**, *37*, 1230-1237.

(117) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *J. Comput. Chem.* **2007**, *28*, 555-569.

(118) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A WELL-BEHAVED ELECTROSTATIC POTENTIAL BASED METHOD USING CHARGE RESTRAINTS FOR DERIVING ATOMIC CHARGES - THE RESP MODEL. *Journal of Physical Chemistry* **1993**, *97*, 10269-10280.

(119) Case, D. A. D., T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.;Luo, R.; Walker, R. C.;Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K.F.; Paesani, F.; Vanicek, J.; Liu, 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. (2010), AMBER version 11, University of California, San Francisco.

(120) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999-2012.

(121) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* **2010**, *78*, 1950-1958.

(122) Tidor, B.; Karplus, M. THE CONTRIBUTION OF VIBRATIONAL ENTROPY TO MOLECULAR ASSOCIATION - THE DIMERIZATION OF INSULIN. *Journal of Molecular Biology* **1994**, *238*, 405-414.

(123) Eckart, C. Some studies concerning rotating axes and polyatomic molecules. *Physical Review* **1935**, *47*, 552-558.

(124) Bowler, D. R.; Miyazaki, T. O(N) methods in electronic structure calculations. *Rep. Prog. Phys.* **2012**, *75*, 43.

(125) Schneider, W. B.; Bistoni, G.; Sparta, M.; Saitow, M.; Riplinger, C.; Auer, A. A.; Neese, F. Decomposition of Intermolecular Interaction Energies within the Local Pair Natural Orbital Coupled Cluster Framework. *J Chem Theory Comput* **2016**, *12*, 4778-4792.

(126) Fedorov, D. G.; Nagata, T.; Kitaura, K. Exploring chemistry with the fragment molecular orbital method. *Phys Chem Chem Phys* **2012**, *14*, 7562-7577.

(127) Neese, F.; Wennmohs, F.; Hansen, A. Efficient and accurate local approximations to coupled-electron pair approaches: An attempt to revive the pair natural orbital method. *J. Chem. Phys.* **2009**, *130*, 18.

(128) He, X.; Merz, K. M. Divide and Conquer Hartree-Fock Calculations on Proteins. *J Chem Theory Comput* **2010**, *6*, 405-411.

(129) Nishizawa, H.; Nishimura, Y.; Kobayashi, M.; Irle, S.; Nakai, H. Three pillars for achieving quantum mechanical molecular dynamics simulations of huge systems: Divide-and-conquer, density-functional tight-binding, and massively parallel computation. *J. Comput. Chem.* **2016**, *37*, 1983-1992.

(130) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chemical Reviews* **2012**, *112*, 632-672.

(131) Suarez, E.; Diaz, N.; Suarez, D. Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide. *J Chem Theory Comput* **2009**, *5*, 1667-1679.

(132) Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599-3605.

(133) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. New advance in computational chemistry: Full quantum mechanical ab initio computation of streptavidin-biotin interaction energy. *J. Phys. Chem. B* **2003**, *107*, 12039-12041.

(134) Roothaan, C. C. J. NEW DEVELOPMENTS IN MOLECULAR ORBITAL THEORY. *Rev. Mod. Phys.* **1951**, *23*, 69-89.

(135) Huang, L.; Massa, L.; Karle, J. The kernel energy method of quantum mechanical approximation carried to fourth-order terms. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 1849-1854.

(136) Phipps, M. J. S.; Fox, T.; Tautermann, C. S.; Skylaris, C. K. Energy decomposition analysis approaches and their evaluation on prototypical protein-drug interaction patterns. *Chem. Soc. Rev.* **2015**, *44*, 3177-3211.

(137) Hirao, H. Energy Decomposition Analysis of the Protein Environmental Effect: The Case of Cytochrome P450cam Compound I. *Chem. Lett.* **2011**, *40*, 1179-1181.

(138) Wu, Q.; Ayers, P. W.; Zhang, Y. K. Density-based energy decomposition analysis for intermolecular interactions with variationally determined intermediate state energies. *J. Chem. Phys.* **2009**, *131*, 8.

(139) Li, L.; Li, C.; Zhang, Z.; Alexov, E. On the 'Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *J Chem Theory Comput* **2013**, *9*, 2126-2136.

(140) Kato, M.; Pisliakov, A. V.; Warshel, A. The barrier for proton transport in aquaporins as a challenge for electrostatic models: The role of protein relaxation in mutational calculations. *Proteins-Structure Function and Bioinformatics* **2006**, *64*, 829-844.

(141) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chemical Reviews* **2005**, *105*, 2999-3093.

(142) Mennucci, B. Continuum Solvation Models: What Else Can We Learn from Them? *Journal of Physical Chemistry Letters* **2010**, *1*, 1666-1674.

(143) Mennucci, B. Modeling environment effects on spectroscopies through QM/classical models. *Phys Chem Chem Phys* **2013**, *15*, 6583-6594.

(144) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808-813.

(145) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society* **2008**, *130*, 2817-2831.

(146) Schrödinger Release 2015-1: WaterMap, version 2.2, Schrödinger, LLC, New York, NY, 2015.

(147) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling* **2013**, *19*, 1-32.

(148) Rezac, J.; Hobza, P. Extrapolation and Scaling of the DFT-SAPT Interaction Energies toward the Basis Set Limit. *J Chem Theory Comput* **2011**, *7*, 685-689.

(149) Boys, S. F.; Bernardi, F. CALCULATION OF SMALL MOLECULAR INTERACTIONS BY DIFFERENCES OF SEPARATE TOTAL ENERGIES - SOME PROCEDURES WITH REDUCED ERRORS. *Mol. Phys.* **1970**, *19*, 553-&.

(150) MOPAC2012, J. J. P. S., Stewart Computational Chemistry, Colorado Spring. CO, USA, http://OpenMOPAC.net (2012).

(151) MOLPRO, v., a package of ab initio programs, H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, Y. Liu, A. W. Lloyd, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, and M. Wang, see http://www.molpro.net.

(152) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **2003**, *91*, 4.

(153) Brunk, E.; Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chemical Reviews* **2015**, *115*, 6217-6263.

(154) Polly, R.; Werner, H. J.; Manby, F. R.; Knowles, P. J. Fast Hartree-Fock theory using local density fitting approximations. *Mol. Phys.* **2004**, *102*, 2311-2321.

(155) Rezac, J.; Huang, Y. H.; Hobza, P.; Beran, G. J. O. Benchmark Calculations of Three-Body Intermolecular Interactions and the Performance of Low-Cost Electronic Structure Methods. *J Chem Theory Comput* **2015**, *11*, 3065-3079.

(156) Lee, C. T.; Yang, W. T.; Parr, R. G. DEVELOPMENT OF THE COLLE-SALVETTI CORRELATION-ENERGY FORMULA INTO A FUNCTIONAL OF THE ELECTRON-DENSITY. *Phys. Rev. B* **1988**, *37*, 785-789.

(157) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787-1799.

(158) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865-3868.

(159) Becke, A. D. DENSITY-FUNCTIONAL THERMOCHEMISTRY .3. THE ROLE OF EXACT EXCHANGE. *J. Chem. Phys.* **1993**, *98*, 5648-5652.

(160) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158-6170.

(161) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 4.

(162) Andzelm, J.; Huzinaga, S.; Klobukowski, M.; Radzio, E. MODEL POTENTIAL STUDY OF THE INTERACTIONS IN AR2, KR2 AND XE2 DIMERS. *Mol. Phys.* **1984**, *52*, 1495-1513.

(163) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **2005**, *7*, 3297-3305.

(164) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. OPTIMIZATION OF GAUSSIAN-TYPE BASIS-SETS FOR LOCAL SPIN-DENSITY FUNCTIONAL CALCULATIONS .1. BORON THROUGH NEON, OPTIMIZATION TECHNIQUE AND VALIDATION. *Can. J. Chem.-Rev. Can. Chim.* **1992**, *70*, 560-571.

(165) Frisch, M. J.; Pople, J. A.; Binkley, J. S. SELF-CONSISTENT MOLECULAR-ORBITAL METHODS .25. SUPPLEMENTARY FUNCTIONS FOR GAUSSIAN-BASIS SETS. *J. Chem. Phys.* **1984**, *80*, 3265-3269.

(166) Hehre, W. J.; Stewart, R. F.; Pople, J. A. SELF-CONSISTENT MOLECULAR-ORBITAL METHODS .I. USE OF GAUSSIAN EXPANSIONS OF SLATER-TYPE ATOMIC ORBITALS. *J. Chem. Phys.* **1969**, *51*, 2657-+.

(167) 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.

(168) **Basis set Exchange. https://bse.pnl.gov/bse/portal (accessed Feb 07, 2017).**

(169) TURBOMOLE V7.0 2015. a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

(170) Fanfrlík, J., et al. Competition between Halogen, Hydrogen and Dihydrogen Bonding in Brominated Carboranes. *ChemPhysChem* **2016**, *17*, 3373–3376.

(171) Jones, S.; Shanahan, H. P.; Berman, H. M.; Thornton, J. M. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Research* **2003**, *31*, 7189-7198.

(172) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research* **2001**, *29*, 2860-2874.

(173) Rai, S. K.; Singh, P.; Kumar, R.; Tewari, A. K.; Hostas, J.; Gnanasekaran, R.; Hobza, P. Experimental and Theoretical Study for the Assessment of the Conformational Stability of Polymethylene-Bridged Heteroaromatic Dimers: A Case of Unprecedented Folding. *Cryst Growth Des* **2016**, *16*, 1176-1180.

(174) Jakubec, D.; Hostas, J.; Laskowski, R. A.; Hobza, P.; Vondrasek, J. Large-Scale Quantitative Assessment of Binding Preferences in Protein-Nucleic Acid Complexes. *J Chem Theory Comput* **2015**, *11*, 1939-1948.

(175) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. USE OF APPROXIMATE INTEGRALS IN ABINITIO THEORY - AN APPLICATION IN MP2 ENERGY CALCULATIONS. *Chem Phys Lett* **1993**, *208*, 359-363.

(176) Riley, K. E.; Rezac, J.; Hobza, P. MP2.X: a generalized MP2.5 method that produces improved binding energies with smaller basis sets. *Phys Chem Chem Phys* **2011**, *13*, 21121-21125.

(177) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*.

(178) 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.

(179) Oswald, C.; Smits, S. H. J.; Hoing, M.; Sohn-Bosser, L.; Dupont, L.; Le Rudulier, D.; Schmitt, L.; Bremer, E. Crystal Structures of the Choline/Acetylcholine Substrate-binding Protein ChoX from Sinorhizobium meliloti in the Liganded and Unliganded-Closed States. *Journal of Biological Chemistry* **2008**, *283*, 32848-32859.

(180) Zhang, G.; Mastalerz, M. Organic cage compounds - from shape-persistency to function. *Chem. Soc. Rev.* **2014**, *43*, 1934-1947.

(181) Ballester, P.; Shivanyuk, A.; Far, A. R.; Rebek, J. A synthetic receptor for choline and carnitine. *Journal of the American Chemical Society* **2002**, *124*, 14014-14016.

(182) Lagona, J.; Mukhopadhyay, P.; Chakrabarti, S.; Isaacs, L. The cucurbit n uril family. *Angew. Chem.-Int. Edit.* **2005**, *44*, 4844-4870.

(183) Wang, R. B.; Yuan, L. N.; Macartney, D. H. Inhibition of C(2)-H/D exchange of a bis(imidazolium) dication upon complexation with cucubit 7 uril. *Chemical Communications* **2006**, 2908-2910.

(184) Mock, W. L.; Shih, N. Y. HOST GUEST BINDING-CAPACITY OF CUCURBITURIL. *Journal of Organic Chemistry* **1983**, *48*, 3618-3619.

(185) Kim, C.; Agasti, S. S.; Zhu, Z. J.; Isaacs, L.; Rotello, V. M. Recognition-mediated activation of therapeutic gold nanoparticles inside living cells. *Nature Chemistry* **2010**, *2*, 962-966.

(186) Reczek, J. J.; Kennedy, A. A.; Halbert, B. T.; Urbach, A. R. Multivalent Recognition of Peptides by Modular Self-Assembled Receptors. *Journal of the American Chemical Society* **2009**, *131*, 2408-2415.

(187) Li, W.; Bockus, A. T.; Vinciguerra, B.; Isaacs, L.; Urbach, A. R. Predictive recognition of native proteins by cucurbit 7 uril in a complex mixture. *Chemical Communications* **2016**, *52*, 8537-8540.

(188) Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S. Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *Journal of Physical Chemistry Letters* **2014**, *5*, 4275-4284.

(189) Sure, R.; Antony, J.; Grimme, S. Blind Prediction of Binding Affinities for Charged Supramolecular Host-Guest Systems: Achievements and Shortcomings of DFT-D3. *J. Phys. Chem. B* **2014**, *118*, 3431-3440.

(190) Jensen, J. H. Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods. *Phys Chem Chem Phys* **2015**, *17*, 12441-12451.

(191) Cao, L.; Skalamera, D.; Zavalij, P. Y.; Hostas, J.; Hobza, P.; Mlinaric-Majerski, K.; Glaser, R.; Isaacs, L. Influence of hydrophobic residues on the binding of CB 7 toward diammonium ions of common ammonium center dot center dot center dot ammonium distance. *Org Biomol Chem* **2015**, *13*, 6249-6254.

(192) Fenley, A. T.; Henriksen, N. M.; Muddana, H. S.; Gilson, M. K. Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril-Guest Binding Enthalpies. *J Chem Theory Comput* **2014**, *10*, 4069-4078.

(193) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 host-guest blind prediction challenge: an overview. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 305-317.

(194) Caldararu, O.; Olsson, M. A.; Riplinger, C.; Neese, F.; Ryde, U. Binding free energies in the SAMPL5 octa-acid host-guest challenge calculated with DFT-D3 and CCSD(T). *Journal of Computer-Aided Molecular Design* **2017**, *31*, 87-106.

(195) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* **1996**, *14*, 33-38.

(196) Nau, W. M.; Florea, M.; Assaf, K. I. Deep Inside Cucurbiturils: Physical Properties and Volumes of their Inner Cavity Determine the Hydrophobic Driving Force for Host-Guest Complexation. *Israel Journal of Chemistry* **2011**, *51*, 559-577.

(197) Biedermann, F.; Uzunova, V. D.; Scherman, O. A.; Nau, W. M.; De Simone, A. Release of High-Energy Water as an Essential Driving Force for the High-Affinity Binding of Cucurbit n urils. *Journal of the American Chemical Society* **2012**, *134*, 15318-15323.

(198) Klamt, A.; Jonas, V. Treatment of the outlying charge in continuum solvation models. *J. Chem. Phys.* **1996**, *105*, 9972-9981.

(199) Lepsik, M.; Rezac, J.; Kolar, M.; Pecina, A.; Hobza, P.; Fanfrlik, J. The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design. *ChemPlusChem* **2013**, *78*, 921-931.

(200) Muddana, H. S.; Gilson, M. K. Calculation of Host-Guest Binding Affinities Using a Quantum-Mechanical Energy Model. *J Chem Theory Comput* **2012**, *8*, 2023-2033.

(201) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463-1473.

List of Attached Publications

Some of the work described here has been presented in the following articles or has been recently submitted as manuscripts:

- **Attachment A:**
  Hostaš, J.; Řezáč, J.; Hobza, P.: "On the Performance of the Semiempirical Quantum Mechanical PM6 and PM7 Methods for Noncovalent Interactions", *Chem. Phys. Lett.* **2013**, 568, 161-166.

- **Attachment B:**
  Hostaš, J.; Řezáč J.: Accurate DFT-D3 Calculations in Small Basis Sets, *submitted*.

- **Attachment C:**
  Jakubec, D.; **Hostaš**, J.; Laskowski, R. A.; Hobza, P.; Vondrášek, J.: "Large-Scale Quantitative Assessment of Bindign Preferences in Protein-Nucleic Acid Complexes", *J. Chem. Theory Comput.* **2015**, 11, 1939-1948.

- **Attachment D:**
  Hostaš, J.; Jakubec, D.; Laskowski, R. A.; Gnanasekaran, R.; Řezáč, J.; Vondrášek, J.; Hobza P.: "Representative Amino Acid Side-Chain Interactions in Protein-DNA Complexes: A Comparison of Highly Accurate Correlated Ab Initio Quantum Mechanical Calculations and Efficient Aprroaches for Applications to Large Systems", *J. Chem. Theory Comput.* **2015**, 11, 4086-4092.

- **Attachment E:**
  Jia, Ch.; Zuo, W., Yang, X.-J.; Chen, Y.; Yang, D.; Cao, L.; Custelcean, R.; **Hostaš**, J.; Hobza, P.; Glaser, R.; Wu, B.: Highly Selective Binding of Choline by a Phosphate-Coordination-Assembled Triple Helicate Featuring an Aromatic Cage that Mimics ChoX Protein, *submitted*.

- **Attachment F:**

  Cao, L.; Škalamera, Đ.; Zavalij, P. Y.; **Hostaš**, J.; Hobza, P.; Mlinarić-Majerski, K.; Glaser, R.; Isaacs L.: "Influence of hydrophobic residues on the binding of CB[7] toward diammonium ions of common ammonium···ammonium distance", *Org. Biomol. Chem.* **2015**, 13, 6249.

- **Attachment G:**

  Hostaš, J.; Sigwalt, D; Šekutor, M.; Ajani, H.; Dubecký, M.; Řezáč, J.; Zavalij, P. Y.; Cao, L.; Wohlschlager, C.; Mlinarić-Majerski, K.; Isaacs, L.; Glaser, R.; Hobza, P.: "A Nexus between Theory and Experiment: Non-empirical Quantum Mechanical Computational Methodology Applied to Cucurbit[n]uril•Guest Binding Interactions", *Chem. Eur. J.* **2016**, 22, 17226-17238, cover page.

- **Attachment H:**

  Sigwalt, D.; Šekutor, M.; Cao, L., Zavalij, P.Y.; **Hostaš**, J.; Ajani, H.; Hobza, P.; Mlinarić-Majerski, K.; Glaser, R.; Isaacs, L.: Unraveling the Structure-Affinity Relationship Between Cucurbit[n]urils (n = 7, 8) and Cationic Diamondoids, *J. Am. Chem. Soc.* **2017**, DOI: 10.1021/jacs.7b00056.

The following related papers are not included in the thesis:

- **Attachment I:**

  Kolář, M.; **Hostaš**, J.; Hobza, P.: "The strength and directionality of a halogen bond are co-determined by the magnitude and size of the sigma-hole", *Phys. Chem. Chem. Phys.* **2014**, 16, 9987-9996.

- **Attachment J:**

  Rai, S. K.; Singh, P.; Kumar, R; Tewari, A. K.; **Hostaš**, J.; Gnanasekaran, R.; Hobza P.: "Experimental and Theoretical Study of the Assesment of the Conformational Stability of Polymethylene-Bridged Heteroaromatic Dimers: A case of Unprecedented Folding", *Cryst. Growth Des.* **2016**, 16, 1179-1180.

Attached Publications