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†

Petr Jurečka,[a] Jiří Šponer,[ab] Jiří Černý[a] and Pavel Hobza*[a]

Received 4th January 2006, Accepted 15th February 2006
First published as an Advance Article on the web 7th March 2006
DOI: 10.1039/b600027d

MP2 and CCSD(T) complete basis set (CBS) limit interaction energies and geometries for more than 100 DNA base pairs, amino acid pairs and model complexes are for the first time presented together. Extrapolation to the CBS limit is done by using two-point extrapolation methods and different basis sets (aug-cc-pVDZ – aug-cc-pVTZ, aug-cc-pVTZ – aug-cc-pVQZ, cc-pVTZ – cc-pVQZ) are utilized. The CCSD(T) correction term, determined as a difference between CCSD(T) and MP2 interaction energies, is evaluated with smaller basis sets (6-31G** and cc-pVDZ). Two sets of complex geometries were used, optimized or experimental ones. The JSCH-2005 benchmark set, which is now available to the chemical community, can be used for testing lower-level computational methods. For the first screening the smaller training set (S22) containing 22 model complexes can be recommended. In this case larger basis sets were used for extrapolation to the CBS limit and also CCSD(T) and counterpoise-corrected MP2 optimized geometries were sometimes adopted.

1 Introduction

The structure of biomacromolecules like DNA, RNA and proteins is determined by noncovalent interactions among the building blocks which are DNA and RNA bases and amino acids. The respective building blocks are electroneutral or charged and various energy terms contribute to their overall stabilization. Among them hydrogen (H-) bonding and electrostatic contributions were expected to be the most important.[1] H-bonding is highly specific and directional and this type of interaction is thus responsible for the stabilization of biomacromolecules but also for such important phenomena as the transfer of genetic code *via* formation of complementary DNA base pairs. Electrostatic interaction is especially important in proteins, and ion pairs or salt bridges play an important role in stabilizing selected protein structures. Induction and charge-transfer terms do not play a decisive role but should be properly considered. Stacking interactions are non-specific, their origin is entirely different from mainly electrostatic H-bonding and for a long time they were believed to be much weaker than the H-bonding. There was a good reason for this—London dispersion energy, which forms a dominant part of stacking, was considered a rather weak and exotic force stabilizing, *e.g.*, rare gas dimers. Only recent calculations demonstrated that stabilization energies of stacked DNA base pairs as well as stacked amino acids can be surprisingly large and almost reaching stabilization of H-bonding. It becomes evident now that stacking is limited not only to the aromatic systems but it is important also for interaction of aromatic systems with other delocalized π-electron systems like peptide bond[2] or even between two systems with delocalized electrons. Evidently, stacking plays a much wider role in biology and thus represents one of the key problems of today's science.

To prove the role of various energy contributions in extended biomacromolecules, for which only simple empirical potential calculations are feasible, is difficult. The one way to solve this problem is to fragment a biomacromolecule into smaller representative components which are tractable at a high quantum chemical level. The situation with both dominant biomacromolecules, DNA and proteins, is favorable since they consist of characteristic building blocks (nucleic acid bases and amino acids) interacting *via* noncovalent interactions. We can thus easily construct smaller complexes (*e.g.* H-bonded and stacked DNA base pairs) possessing characteristic stabilizing contributions. The problem is that experimental data on these and similar complexes allowing the extraction of the relative importance of H-bonding, stacking and other energy contributions are missing and the only possibility to address this problem is to use advanced quantum chemical procedures. Because of the different origin of all these interactions, calculations should be performed at a high theoretical level to prevent the traditional problems of quantum and computational chemistry such as the size of the AO basis set used or the portion of correlation energy covered.

The significant progress in quantum and computational chemistry made in recent years allows us to estimate interaction energies of extended complexes with more than 24 atoms (the size of the benzene dimer) at the CCSD(T) level with

[a] *Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, Flemingovo náměstí 2, 166 10 Prague 6, Czech Republic. E-mail: pavel.hobza@uochb.cas.cz*
[b] *Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic*

† Electronic supplementary information (ESI) available: Geometry parameters of all structures (3-D rotatable figures). See DOI: 10.1039/b600027d

infinite AO basis set.[3,4] The infinite basis set calculations are realized by extrapolation to the complete basis set (CBS) limit.[5,6] The CCSD(T) method is accurate enough and represents a compromise in accuracy and economy of the calculations. We have shown recently[7] that the CCSDT and CCSD(T) interaction energies of several model complexes (H-bonded and stacked ones) agree very well. The computational time for the former method is prohibitively large and the respective calculations can be thus performed for small and highly symmetric complexes only. On the other hand, the calculations with the CCSD(T)/infinite basis set can be realized even for extended complexes and these calculations represent the new benchmark level. The existence of the benchmark data for a broad set of extended complexes is extremely important since this set can be used for the testing of new computational procedures. It is clear that progress in material and biological sciences requires the application of accurate computational methods allowing us to tackle systems with thousands of atoms at both static and dynamic levels. Since 2003 when we presented our first paper on accurate stabilization energies of DNA base pairs we have published several papers with accurate stabilization energies of H-bonded and stacked DNA and RNA base pairs[8–13] as well as amino acid pairs.[2] We systematically used the same computational philosophy that now allows us to collect all these data and publish the accurate interaction energies for the whole set. There is a good reason to do it—we were frequently asked by our colleagues from different computational laboratories to provide our accurate stabilization energies for H-bonded and stacked pairs. We are aware that each benchmark set requires an acronym; we used the initials of our family names and named the set as JSCH. Since the set will be extended in the future we named the present one as JSCH-2005.

We expect that the JSCH-2005 benchmark set will be used mainly for estimating the accuracy of less reliable theoretical approaches such as empirical force fields, semiempirical methods, or density functional based methods. In some cases, working with the whole set counting over 100 complexes may be impractical. Therefore, we decided to separate a smaller set of 22 mostly small complexes, which could be conveniently used as a training set (Set 22, S22). The remaining part of our benchmark database can then serve as a realistic validation set of the "real life" molecules. We believe that our S22 set will manage to represent non-covalent interactions in biological molecules in a balanced way and that it will help to design and test fast computational tools for biologically oriented applications.

2 Methods

Complete basis set limit calculations

Since the total interaction energy is constructed as a sum of the Hartree–Fock (HF) and correlated (COR) interaction energies, the extrapolation to the CBS limit can be done separately for both components. The reason for the separated extrapolation is the fact that the HF interaction energy converges with respect to the one-electron basis set already for relatively small basis sets while the correlation interaction energy converges to its CBS limit unsatisfactorily slow. Several extrapolation schemes were suggested in the literature, for instance those of Helgaker *et al.*[14,15] and Truhlar[16] (eqn (1) and (2)) and others.[17,18]

$$E_X^{\mathrm{HF}} = E^{\mathrm{HF}}{}_{\mathrm{CBS}} + A\exp(-\alpha X) \text{ and } E_X^{\mathrm{corr}} = E^{\mathrm{corr}}{}_{\mathrm{CBS}} + BX^{-3} \quad (1)$$

$$E_X{}^{\mathrm{HF}} = E_{\mathrm{CBS}}{}^{\mathrm{HF}} + BX^{-\alpha} \text{ and } E_X{}^{\mathrm{corr}} = E_{\mathrm{CBS}}{}^{\mathrm{corr}} + BX^{-\beta}. \quad (2)$$

We have chosen the scheme of Helgaker *et al.*, in which E_X and E_{CBS} are energies for the basis set with the largest angular momentum X and for the complete basis set respectively, and α is a parameter fitted in the original work. The two point extrapolation form is preferable as it was shown that the inclusion of an additional lower quality basis set results often spoils the quality of the fit, especially in case of the smallest basis sets like cc-pVDZ.[15] The most problematic are the stacked clusters for which the double-ζ basis sets yield strongly underestimated stabilization energies and first reasonable results are obtained with aug-cc-pVDZ (or similar) basis set. The extrapolation can only be performed if systematically improved AO basis sets are applied. Throughout this study we used the Dunning's AO basis sets and both augmented as well as non-augmented ones were applied. The HF and correlation interaction energies were corrected for the basis set superposition error (BSSE)[19] and extrapolation was applied to the total energies as well as to the BSSE corrected subsystem energies. Frozen-core approximation was applied throughout this study.

The question arises at which level the extrapolation should be performed. The choice of a method is determined by the fact that it is necessary to perform calculations at two subsequent levels and the only tractable combinations are aug-cc-pVDZ, aug-cc-pVTZ; aug-cc-pVTZ, aug-cc-pVQZ and cc-pVTZ,cc-pVQZ. In the case of the S22 set we have used also larger basis sets (up to cc-pV5Z) whenever possible. Since the systems considered are extended, it becomes evident that CCSD(T) calculations are above the possibilities of the present computer resources and extrapolation can only be performed at the MP2 level. The role of the higher-order correlation energy contributions can not be neglected and the CBS limit CCSD(T) interaction energies were determined using the following scheme:

$$\Delta E^{\mathrm{CCSD(T)}} = \Delta E^{\mathrm{MP2}}{}_{\mathrm{CBS}} + (\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}})|_{\mathrm{smallbasis\,set}}. \quad (3)$$

The use of eqn (3) is based on the assumption that the difference between the CCSD(T) and MP2 interaction energies ($\Delta E^{\mathrm{CCSD(T)}}$–$\Delta E^{\mathrm{MP2}}$) depends only negligibly on the basis set size and can thus be determined with small or medium basis set only. This assumption was shown to be valid and supporting results were obtained not only for model H-bonded[20] and stacked[13] clusters but recently even for H-bonded and stacked structures of the smallest NA base pair–uracil dimer.[21]

We have shown that extrapolation can be done only at three different levels where the first one (aug-cc-pVDZ → aug-cc-pVTZ) is the easiest. To prove its validity it is, however, necessary to compare it with higher-level extrapolation (aug-cc-pVTZ → aug-cc-pVQZ). The MP2/aug-cc-pVQZ calculations for clusters of size of DNA base pairs or amino acid pairs with present computer resources is difficult—if not

impractical—and these calculations become only feasible by using the approximate resolution of identity MP2 (RI-MP2) method[22,23] (and references therein). We explored recently the applicability of the RI-MP2 method for NA bases, base pairs[24] and larger DNA fragments and have shown that the method is capable of an accurate description of H-bonded and stacked NA base pairs. The RI-MP2 total as well as relative energies differ only marginally from those evaluated with the exact MP2 method, while the time saving is as large as one order of magnitude.

All the RI-MP2[22,23] calculations were carried out using the TURBOMOLE 5.6 program suite[25] using the aug-cc-pVXZ (X = D, T, Q) and cc-pVXZ (X = T, Q) basis sets[26,27] and standard (default) TURBOMOLE auxiliary basis sets.[23] The CCSD(T) calculations using 6-31G*(0.25), 6-31G**(0.25, 0.15) (numbers in parentheses give the value of d- and p-polarization functions used instead of standard functions with exponents 0.8 and 1.1, respectively), cc-pVDZ and aug-cc-pVDZ basis sets were performed with the MOLPRO 2002 suit of programs.[28]

Geometries

Two different sources of complex geometries were used. Firstly, the complex structure was optimized and gradient optimization was mostly performed at the RI-MP2 level using the cc-pVTZ [4s3p2d1f/3s2p1d] or very similar TZVPP [5s3p2d1f/3s2p1d] basis sets. The stabilization energy was then *a posteriori* corrected for the BSSE. Because the BSSE is determined for the subsystem geometry taken from a dimer (and not from optimization of an isolated monomer) it is necessary to remove so called deformation energy. This is the energy needed to deform the monomer from the optimized geometry to the geometry it possesses in the dimer. The deformation energy is always positive (repulsive). In several cases (see S22) theoretically more justified counterpoise-corrected gradient optimization was also used. A very important advantage of this procedure is the fact that the BSSE error does not influence the energy through the artificial deformation of the monomers. The smallest complexes were optimized by the CCSD(T) method (numerical gradients) using cc-pVTZ and cc-pVQZ basis sets without counterpoise correction. Besides the optimized geometries experimental (crystal and NMR solvent) geometries were also used and the single-point calculations were performed for the fixed geometry. In the cases when hydrogen atoms were missing their positions were determined by performing the DFT/B3LYP/6-31G** gradient optimization. In the case of experimental geometry the deformation energy is not defined.

Complexes considered

DNA and RNA base pairs. Geometries of complexes considered were either fully optimized or taken from experiment and abbreviation OG (optimized geometry) or EG (experimental geometry) is added to each structure.

Amino acid pairs. Experimental geometries of amino acid pairs were systematically used; neutral amino acids as well as ion pairs are included. The geometry of heavy atoms was taken from the experiment while the positions of the hydrogen atoms were optimized for each amino acid pair.

S22 set. S22 set consists of small to relatively large (30 atoms) complexes of common molecules containing only C, N, O and H, and single, double and triple bonds. Most typical noncovalent interactions, such as hydrogen bonds (XH···Y), dispersion interactions (stacked parallel, T-shaped), and mixed electrostatic-dispersion interactions are represented. A total of 22 complexes are divided into three subgroups: (i) hydrogen bonded complexes; (ii) complexes with predominant dispersion stabilization; (iii) mixed complexes in which electrostatic and dispersion contributions are similar in magnitude. All complexes are in their optimal geometries.

3 Results and discussion

DNA and RNA base pairs

First, the gas phase geometries will be considered. Table 1 shows a finding of key importance, namely that all stabilization energies‡ are large—much larger—than previously expected. The GC WC pair in various modifications possessing three H-bonds represents the strongest pair with stabilization energies reaching or even exceeding 29 kcal mol^{-1}. Other planar H-bonded nucleic acid pairs are less stable but even the weakest pair has a stabilization energy of about 10 kcal mol^{-1}. A surprising result concerns stability of the stacked pairs. They are less stable than their H-bonded counterparts but stabilization energies are frequently larger than 10 kcal mol^{-1} and in the strongest pairs it reaches 18 kcal mol^{-1}. Investigating the stacked structures, we conclude that they do not contain regular H-bonds. In several stacked structures the X-H···Y H-bonded contacts exist but they are never as linear as required for strong H-bonding. It is to be mentioned that the GC stacked pairs are about as stable as the AT planar pair containing two strong linear H-bonds. This finding changes the widely spread opinion about stabilization of DNA double helix where a dominant contribution was expected to originate in the H-bonding.

Extrapolation to the CBS limit is essential for both types of interactions and increases the stabilization energy significantly. Passing from aug-cc-pVDZ to aug-cc-pVTZ basis set is connected with relatively important increase of 5 and 7% for H-bonding and stacking, respectively (in absolute values 1.0 and 0.6 kcal mol^{-1}, see Table 1). Passing from aug-cc-pVTZ to very large aug-cc-pVQZ is still connected with some stabilization energy increase which is on average 45 and 33% of the increase for the previous pair of basis sets for H-bonded and stacked complexes, respectively. Extrapolation from the latter pair of basis sets yields systematically larger stabilization energies than that from the previous pair and the average difference is about 0.5 kcal mol^{-1} for both H-bonds and stacks. We can conclude that the extrapolation from the aug-cc-pVDZ to aug-cc-pVTZ basis sets is feasible even for relatively large complexes and yields a significant improvement of the stabilization energy in comparison with the aug-

‡ Stabilization energy (E_{tot} in the tables) includes deformation of the monomers, whereas interaction energy (E_{int}) does not.

Table 1 Interaction energies of DNA base pairs (kcal mol^{-1}). Abbreviations used in the first column are the same as in original references: (A, T, C, G, U, I, F – adenine, thymine, cytosine, guanine, uracil, inosine, difluorotoluene; m, t, o – methyl-, thio-, oxo-; WC, pl – Watson–Crick, planar; s, st, S – stacked and is, IS – interstrand)

Hydrogen-bonded DNA base pairs

Complex	Geom.[a]	Ref.	aD[b]	aT[b]	aQ[b]	aD→aT Helgaker[c]	aD→aT Truhlar[c]	aT→aQ Helgaker[c]	E_{def}[d]	ΔCCSD(T)	E_{int}[e]	E_{tot}
G···C WC	OG	9	−28.68	−30.44	−31.13	−31.15	—	−31.59	3.36	−0.47	−32.06	−28.80
mG···mC WC	OG	9	−28.31	−30.04	−30.68	−30.74	—	−31.11	3.16	−0.48	−31.59	−28.50
A···T WC	OG	9	−14.81	−16.05	−16.53	−16.56	—	−16.85	1.43	−0.01	−16.86	−15.43
mA···mT H	OG	9	−16.19	−17.34	−17.90	−17.81	—	−18.29	1.89	0.14	−18.16	−16.27
8oG···C WC pl	OG	10	−30.20	−31.60	−32.30	−32.10	−33.20	−32.90	3.90	−0.40	−33.30	−29.40
I···C WC pl	OG	10	−22.80	−23.80	−24.30	−24.20	−24.90	−24.70	2.20	−0.20	−24.90	−22.70
G···U wobble	OG	10	−17.40	−18.40	−18.70	−18.80	−19.00	−18.80	3.00	−0.30	−19.10	−16.10
CCH+	OG	10	−49.20	−50.60	−51.30	−51.90	−52.10	−51.30	4.90	−0.10	−51.40	−46.50
U···U Calcutta pl	OG	10	−9.40	−9.80	−10.10	−10.00	−10.30	−10.20	0.50	−0.10	−10.30	−9.80
U···U pl	OG	10	−12.50	−13.20	—	−13.50	−13.80	—	1.10	−0.20	−13.70	−12.60
6tG···C WC pl	OG	10	−27.60	−29.00	—	−29.50	−25.90	—	4.00	—	−29.50	−25.50
A···4tU WC	OG	10	−13.80	−14.10	—	−14.20	−14.4	—	1.00	—	−14.20	−13.20
2-aminoA···T	OG	10	−18.00	−19.00	—	−19.50	−19.8	—	1.90	—	−19.50	−17.60
2-aminoA···T pl	OG	10	−18.20	−19.30	—	−19.70	−20.0	—	2.40	—	−19.70	−17.30
A···F	OG	10	−5.10	−5.20	—	−5.20	−5.4	—	0.30	—	−5.20	−4.90
G···4tU	OG	10	−16.60	−17.50	—	−17.80	−18.1	—	1.90	—	−17.80	−15.90
G···2tU	OG	10	−15.10	−16.20	—	−16.60	−16.9	—	2.00	—	−16.60	−14.60
A···C pl	OG	10	−16.60	−17.30	—	−17.60	−17.8	—	1.70	—	−17.60	−15.90
G···G pl	OG	10	−20.30	−21.00	—	−21.30	−21.4	—	2.90	—	−21.30	−18.40
G···6tG pl	OG	10	−20.30	−21.40	—	−21.80	−22.0	—	2.80	—	−21.80	−19.00
6tG···G pl	OG	10	−21.40	−22.40	—	−22.70	−23.0	—	3.10	—	−22.70	−19.60
G···A 1	OG	10	−18.30	−19.10	—	−19.40	−19.7	—	1.90	—	−19.40	−17.50
G···A 1 pl	OG	10	−17.60	−18.60	—	−18.90	−19.1	—	2.80	—	−18.90	−16.10
G···A 2	OG	10	−13.20	−14.00	—	−14.40	−14.8	—	3.50	—	−14.40	−10.90
G···A 2 pl	OG	10	−11.80	−12.50	—	−12.80	−12.9	—	2.30	—	−12.80	−10.50
G···A 3	OG	10	−17.80	−18.50	—	−18.80	−19.0	—	2.00	—	−18.80	−16.80
G···A 4	OG	10	−12.40	−13.20	—	−13.50	−13.6	—	1.40	—	−13.50	−12.10
A···A 1 pl	OG	10	−13.60	−14.20	—	−14.50	−14.6	—	1.40	—	−14.50	−13.10
A···A 2 pl	OG	10	−12.90	−13.40	—	−13.70	−13.8	—	1.40	—	−13.70	−12.30
A···A 3 pl	OG	10	−11.50	−12.00	—	−12.20	−12.3	—	1.30	—	−12.20	−10.90
8oG···G	OG	10	−21.50	−22.40	—	−22.80	−23.1	—	3.20	—	−22.80	−19.60
2tU···2tU pl	OG	10	−11.20	−12.20	—	−12.60	−13.0	—	1.00	—	−12.60	−11.60
A···T WC	EG	12	−14.80	−15.90	—	−16.40	—	—	—	0.00	−16.40	—
G···C WC*	EG	12	−32.70	−34.60	—	−35.40	—	—	—	−0.40	−35.80	—
A···T WC	EG	12	−16.60	−17.70	—	−18.20	—	—	—	−0.20	−18.40	—
G···A HB	EG	12	−10.20	−11.40	—	−11.90	—	—	—	0.60	−11.30	—
C···G WC	EG	12	−27.70	−29.50	—	−30.30	—	—	—	−0.40	−30.70	—
G···C WC	EG	12	−28.40	−30.20	—	−31.00	—	—	—	−0.40	−31.40	—
Average – aQZ data[f]			−24.11	−25.34	−25.88	−25.92		−26.19				
Average-all			−19.16	−20.20		−20.64				−0.20	−20.79	

Interstrand base pairs

Complex	Geom.	Ref.	aD[a]	aT[a]	aD→aT Helgaker[b]	aD→aT Truhlar[b]	ΔCCSD(T)	E_{int}[d]
GG0/3.36 CGis036	EG	8	−3.41	−3.54	−3.59	−3.61	−0.10	−3.68
GG0/3.36 GCis036	EG	8	−4.64	−4.68	−4.70	−4.70	−0.12	−4.82
AA20/3.05 ATis2005	EG	8	−2.05	−2.23	−2.31	−2.37	−0.03	−2.34
AA20/3.05 TAis2005	EG	8	−2.01	−2.18	−2.25	−2.30	0.09	−2.16
GC0/3.25 C//Cis	EG	8	2.89	2.88	2.87	2.87	0.22	3.09
GC0/3.25 G//Gis	EG	8	1.48	1.35	1.29	1.24	0.63	1.93
CG0/3.19 G//Gis	EG	8	−4.03	−4.56	−4.79	−4.94	0.88	−3.91
CG0/3.19 C//Cis	EG	8	1.16	1.12	1.10	1.08	0.14	1.24
GA10/3.15 A//Cis	EG	8	−0.18	−0.27	−0.31	−0.33	0.00	−0.31
GA10/3.15 T//Gis	EG	8	0.49	0.41	0.38	0.36	0.19	0.58
AG08/3.19 T//Gis	EG	8	−0.19	−0.45	−0.56	−0.63	0.09	−0.47
AG08/3.19 A//Cis	EG	8	−0.11	−0.21	−0.26	−0.29	0.08	−0.18
TG03.19 A//Gis	EG	8	−4.32	−4.72	−4.88	−4.99	0.66	−4.22
TG03.19 T//Cis	EG	8	−1.12	−1.14	−1.15	−1.16	0.00	−1.15
GT10/3.15 T//Cis	EG	8	0.22	0.21	0.20	0.20	0.10	0.30
GT10/3.15 A//Gis	EG	8	−3.99	−4.20	−4.29	−4.34	0.23	−4.06
AT10/3.26 T//Tis	EG	8	0.80	0.75	0.73	0.72	0.16	0.88
AT10/3.26 A//Ais	EG	8	−0.87	−1.01	−1.07	−1.12	0.15	−0.92
TA08/3.16 A//Ais	EG	8	−1.87	−2.24	−2.40	−2.50	0.85	−1.55
TA08/3.16 T//Tis	EG	8	0.60	0.58	0.57	0.56	0.14	0.70
AA0/3.24 A//Tis	EG	8	−1.71	−1.75	−1.77	−1.79	0.07	−1.71
AA0/3.24 T//Ais	EG	8	−1.31	−1.37	−1.39	−1.40	0.09	−1.30
A···A IS	EG	12	−0.60	−0.80	−0.90	—	0.20	−0.70
T···T IS	EG	12	0.90	0.80	0.80	—	0.20	1.00

Table 1 (*continued*)

G···G IS	EG	12	−4.90	−5.40	−5.60	—	1.10	−4.50
C···C IS	EG	12	1.40	1.30	1.30	—	0.10	1.40
A···G IS	EG	12	−4.70	−4.90	−5.00	—	0.20	−4.80
T···C IS	EG	12	−0.20	−0.20	−0.20	—	0.10	−0.10
C···A IS	EG	12	−2.70	−2.90	−3.10	—	0.10	−3.00
G···G IS	EG	12	−4.90	−5.00	−5.10	—	−0.10	−5.20
G···G IS	EG	12	0.40	0.10	0.00	—	0.80	0.80
C···C IS	EG	12	2.90	2.90	2.90	—	0.20	3.10
Average			−1.02	−1.17	−1.30		0.24	−1.00

Stacked base pairs

Complex	Geom.	Ref.	aD[a]	aT[a]	aQ[a]	aD→aT Helgaker[b]	aD→aT Truhlar[b]	aT→aQ Helgaker[b]	E_{def}[c]	ΔCCSD(T)	E_{int}[d]	E_{tot}
G···C S	OG	9	−18.53	−19.97	−20.48	−20.57	—	−20.84	2.04	1.82	−19.02	−16.90
mG···mC S	OG	9	−20.04	−21.64	−22.35	−22.30	—	−22.78	2.24	2.43	−20.35	−18.00
A···T S	OG	9	−13.13	−14.37	−14.78	−14.90	—	−15.08	0.67	2.77	−12.30	−11.64
mA···mT S	OG	9	−15.78	−17.31	−17.80	−17.95	—	−18.14	1.58	3.57	−14.57	−13.10
CC1	OG	11	1.56	0.87	—	0.58	—	—	—	1.87	2.45	—
CC2	OG	11	−4.48	−5.14	—	−5.40	—	—	—	1.55	−3.85	—
CC3	OG	11	−9.05	−9.72	—	−9.99	—	—	—	1.11	−8.88	—
CC4	OG	11	−10.15	−10.78	—	−11.03	—	—	—	1.11	−9.92	—
CC5	OG	11	−0.60	−1.26	—	−1.52	—	—	—	1.84	0.32	—
CC6	OG	11	−0.27	−0.95	—	−1.24	—	—	—	1.88	0.64	—
CC7	OG	11	−1.46	−1.83	—	−1.98	—	—	—	1.00	−0.98	—
CC8	OG	11	−9.07	−9.58	—	−9.81	—	—	—	0.71	−9.10	—
CC9	OG	11	−9.38	−10.03	—	−10.31	—	—	—	1.20	−9.11	—
CC10	OG	11	−8.42	−9.11	—	−9.39	—	—	—	1.12	−8.27	—
CC11	OG	11	−9.44	−10.05	—	−10.29	—	—	—	0.86	−9.43	—
CC12	OG	11	−7.20	−7.45	—	−7.55	—	—	—	0.12	−7.43	—
CC13	OG	11	−8.45	−9.37	—	−9.71	—	—	—	0.91	−8.80	—
CC14	OG	11	−9.02	−9.71	—	−10.00	—	—	—	0.89	−9.11	—
AAst	EG	8	−10.50	−11.10	—	−11.40	—	—	—	2.82	−8.58	—
GGst	EG	8	−13.80	−14.50	—	−14.80	—	—	—	2.13	−12.67	—
ACst	EG	8	−11.30	−11.90	—	−12.20	—	—	—	1.98	−10.22	—
GAst	EG	8	−13.20	−13.80	—	−14.00	—	—	—	2.62	−11.38	—
CCst	EG	8	−10.10	−10.70	—	−11.00	—	—	—	0.98	−10.02	—
AUst	EG	8	−10.70	−11.30	—	−11.60	—	—	—	1.81	−9.79	—
GCst	EG	8	−11.10	−11.70	—	−12.00	—	—	—	1.40	−10.60	—
CUst	EG	8	−10.20	−10.80	—	−11.10	—	—	—	0.68	−10.42	—
UUst	EG	8	−7.70	−8.30	—	−8.50	—	—	—	1.04	−7.46	—
GUst	EG	8	−12.40	−13.10	—	−13.40	—	—	—	1.31	−12.09	—
GG0/3.36 GGs036	EG	8	−4.90	−5.42	—	−5.64	−5.77	—	—	2.10	−3.54	—
GG0/3.36 CCs036	EG	8	−2.13	−2.52	—	−2.67	−2.76	—	—	1.05	−1.62	—
AA20/3.05 AAs2005	EG	8	−7.66	−8.33	—	−8.60	−8.76	—	—	2.54	−6.06	—
AA20/3.05 TTs2005	EG	8	−4.61	−5.22	—	−5.47	−5.63	—	—	1.29	−4.18	—
GC0/3.25 G//Cs	EG	8	−11.21	−11.93	—	−12.22	−12.39	—	—	1.42	−10.80	—
CG0/3.19 G//Cs	EG	8	−8.05	−8.43	—	−8.57	−8.62	—	—	0.69	−7.88	—
GA10/3.15 A//Gs	EG	8	−10.51	−11.29	—	−11.60	−11.79	—	—	2.47	−9.14	—
GA10/3.15 T//Cs	EG	8	−5.06	−5.55	—	−5.75	−5.87	—	—	1.05	−4.69	—
AG08/3.19 A//Gs	EG	8	−8.87	−9.30	—	−9.46	−9.55	—	—	1.89	−7.58	—
AG08/3.19 T//Cs	EG	8	−6.22	−6.64	—	−6.81	−6.90	—	—	0.73	−6.07	—
TG03.19 T//Gs	EG	8	−5.92	−6.34	—	−6.50	−6.59	—	—	0.83	−5.67	—
TG03.19 A//Cs	EG	8	−5.46	−5.94	—	−6.12	−6.21	—	—	1.16	−4.96	—
GT10/3.15 T//Gs	EG	8	−5.62	−6.42	—	−6.74	−6.94	—	—	1.78	−4.96	—
GT10/3.15 A//Cs	EG	8	−6.26	−6.84	—	−7.07	−7.21	—	—	1.63	−5.44	—
AT10/3.26 A//Ts	EG	8	−7.39	−7.99	—	−8.24	−8.39	—	—	1.60	−6.64	—
TA08/3.16 A//Ts	EG	8	−6.47	−6.91	—	−7.09	−7.19	—	—	1.02	−6.07	—
AA0/3.24 A//As	EG	8	−7.74	−8.35	—	−8.59	−8.74	—	—	2.34	−6.25	—
AA0/3.24 T//Ts	EG	8	−4.13	−4.85	—	−5.15	−5.32	—	—	1.29	−3.86	—
A···T S	EG	12	−9.20	−9.80	—	−10.10	—	—	—	2.00	−8.10	—
G···C S	EG	12	−8.10	−8.30	—	−8.30	—	—	—	0.40	−7.90	—
A···C S	EG	12	−7.70	−8.10	—	−8.30	—	—	—	1.60	−6.70	—
T···G S	EG	12	−7.00	−7.60	—	−7.90	—	—	—	1.70	−6.20	—
G···C S	EG	12	−8.30	−8.90	—	−9.20	—	—	—	1.50	−7.70	—
A···G S	EG	12	−7.80	−8.50	—	−8.80	—	—	—	2.30	−6.50	—
C···G S	EG	12	−12.90	−13.50	—	−13.80	—	—	—	1.40	−12.40	—
G···C S	EG	12	−11.80	−12.40	—	−12.70	—	—	—	1.10	−11.60	—
Average – aQZ data[f]			−16.87	−18.32	−18.85	−18.93		−19.21				—
Average – all			−8.42	−9.07		−9.35				1.53	−7.84	

*The geometries of both GC WC pairs in Table 1 in ref. 12 are identical.[a] OG and EG mean optimized and experimental geometry. [b] Basis set: aug-cc-pVXZ. [c] The extrapolated interaction energy—see Methods. [d] Deformation energy of monomers. [e] E^{int} is the interaction energy, *i.e.*, it is not corrected for the deformation energy of monomers. [f] Average over the data calculated with the aug-cc-pVQZ basis set.

cc-pVDZ results. The respective CBS values are comparable to much more expensive CBS results obtained from aug-cc-pVTZ – aug-cc-pVQZ extrapolation. Helgaker[14] and Truhlar[16] extrapolation schemes differ only slightly and both provide similar average errors. The strong point of the former extrapolation scheme is the fact that it gives very good agreement with the results for the aug-cc-pVQZ basis set.

Inclusion of the ΔCCSD(T) correction term is important and generally cannot be neglected. In the case of planar H-bonded complexes, this term (sometimes stabilizing, sometimes destabilizing) is small. The largest absolute value of 0.66 kcal mol^{-1} was found for the 2-pyridoxine···2-aminopyridine complex while the average absolute value for all H-bonded complexes is −0.2 kcal mol^{-1}. Evidently, the ΔCCSD(T) correction term is small for this class of complexes and can in most cases be neglected. The complete opposite is the situation for stacked complexes where the ΔCCSD(T) correction is always important. Table 1 shows the largest ΔCCSD(T) correction term (3.57 kcal mol^{-1}) for the mAmT stacked pair. The average value for 53 stacked pairs amounts to 1.5 kcal mol^{-1} and on average it constitutes about 20% of the total stabilization. This is a surprisingly large value. It should be mentioned that we have never found any stacked nucleic acid base pair having stabilizing ΔCCSD(T) correction term and stacking is systematically connected with positive (repulsive) ΔCCSD(T) correction. Consequently, the ΔCCSD(T) term for stacked complexes cannot be neglected and should be included every time. This conclusion seriously complicates the calculations of stabilization energies for stacked nucleic acid base pairs since the determination of the ΔCCSD(T) correction term for such extended complexes (mostly without any symmetry element) is extremely time consuming. Only in the case of DNA base pairs can we clearly separate (planar) H-bonded and stacked which allows us to avoid time consuming calculations of the CCSD(T) term for the former complexes. Unfortunately, the situation in proteins is much more complicated (see later) as it is difficult to find such clearly defined geometry motifs. This means that in these cases the CCSD(T) calculations should be systematically applied.

Let us focus on the differences between the description of the H-bonded and stacked complexes in more detail now. First, we will compare the results for a very popular MP2/aug-cc-pVTZ combination with our CCSD(T)/CBS reference. While H-bonded complexes are underestimated on average by about 8%, stacked complexes are by about 7% overestimated at this level of theory (see Table 1). When we compare this 15% difference, corresponding to about 2.2 kcal mol^{-1}, with a typical span of the stabilities of different conformers in proteins (15 most stable conformers of phenylalanyl-glycyl-glycine are distributed in the energy span of 2.5 kcal mol^{-1}, see ref. 29), it is immediately obvious that if complexes or conformers with some dispersion contribution are to be ordered on the same energy scale, MP2/aug-cc-pVDZ method is insufficient and high level of theory is clearly unavoidable. Is this a consequence of the basis set incompleteness or rather of the higher order correlation correction? Comparing the MP2/CBS results with the CCSD(T)/CBS results we can see that the H-bonded interactions are underestimated by about 0.7% by the MP2 method while the stacked interactions are overestimated by about 19%. Therefore, a major part of the discrepancy between these two types of complexes originates in the higher order correlation contributions and the influence of the basis set size on the relative description of the H-bonded and stacked complexes is only secondary.

Experimental stabilization energies which can be used for verification of the theoretical procedures exist for a limited number of complexes only and, surprisingly, DNA base pairs belong to this class. Stabilization enthalpies of various DNA base pairs including the 9-methyl guanine···1-methyl cytosine and 9-methyl adenine···1-methyl thymine were obtained from the temperature dependence of the equilibrium constants measured at 323 and 381 K.[30] Comparison of theoretical and experimental data is not straightforward since at such high temperatures a mixture of various complexes can be found. Further, it is necessary to pass from the stabilization energy to the free energy of the interaction. Performing the molecular dynamics simulations at temperatures mentioned we found[31] that H-bonded structures were populated negligibly, while stacked structures were populated dominantly. Taking these populations into account we determined[9] the average stabilization enthalpy for both pairs (18.0 and 11.3 kcal mol^{-1}) which agreed fairly well with respective experimental data[30] (21.0 and 13.0 kcal mol^{-1}). The theoretical stabilization energies are thus still too small. In the original paper, we concluded that while the H-bonded stabilization energies were close to accurate data the stacking energies were too small by about 10%. However, a more careful analysis indicates that a large part of the discrepancy is probably caused by inaccurate populations taken from the MM/Quench study[31] (empirical force field somewhat underestimates population of energetically favorable hydrogen bonded structures). The stabilization energies of both the H-bonded and stacked complexes presented in Table 1 are thus probably rather accurate.

Now we will discuss the DNA base pairs in experimental geometries taken from X-ray or NMR measurements.[12] Table 1 shows that stabilization energies of GC WC and AT WC pairs agree fairly well with the stabilization energies determined for the gas-phase optimized geometries, except in a few cases where the distortion of the electron density due to inaccurate X-ray geometry has caused small deviations (GC WC). This is understandable in light of the fact that the experimental geometries agree relatively well with the theoretical data. On the other hand, for the stacked base pairs the stabilization energies calculated with the experimental geometries are smaller than these for the optimized geometries, but are still rather large. For example, for the AT rich double helix both H-bonding and stacking contribute almost equally. In the DNA crystal H-bonded and stacked motifs are supplemented by interstrand contacts which are as numerous as H-bonded and stacked ones. Investigating Table 1, we found that they are characterized with smaller stabilization energies than H-bonded ones but they are definitely not negligible. Sometimes, interstrand contacts are connected with repulsion. Extrapolation for aug-cc-pVDZ and aug-cc-pVTZ is connected with average stabilization energy increase of 0.05 kcal mol^{-1}

(5%). Not surprisingly, the same extrapolation for pair having repulsive interaction leads to reduction of repulsion. The CCSD(T) term is smaller than for previous (stacked) complexes and it is mostly repulsive (on average +0.2 kcal mol^{-1}).

Amino acid pairs

Twelve neutral amino acid pairs containing phenylalanine were taken from the hydrophobic core of small protein Rubredoxin[2] and their stabilization energies are collected in Table 2. Contrary to the DNA base pairs discussed in the previous paragraph, the amino acid pairs do not correspond either to H-bonded or to stacked arrangement and mostly they do not contain a strong H-bond. Stabilization energies were smaller than these of DNA base pairs but were still rather large. The largest one (8.2 kcal mol^{-1}) corresponds to the phenylalanine···peptide bond complex where a peptide bond is modeled by *N*-methylformamide. The stabilization energy of the motif (without any H-bond) is surprisingly high and sheds new light on the role of peptide bonds in stabilization of protein structures. Here the extrapolation to the CBS limit was done only for aug-cc-pVDZ and aug-cc-pVTZ basis sets and for the strongest pairs the CBS limits differ from the aug-cc-pVDZ stabilization energy by about 1 kcal mol^{-1}. The ΔCCSD(T) correction term is systematically small (roughly about 10% of CBS stabilization energy) and is always positive (destabilizing).

Stabilization energies of the last six complexes in Table 2 are very large and in one case even exceed 100 kcal mol^{-1}. This is understandable in light of the fact that these complexes correspond to ion pairs. The extrapolation to the CBS limit is, however, not negligible and three times leads to stabilization increase and three times to stabilization decrease. The CCSD(T) term is systematically very small and can be safely neglected, which is probably true for all salt bridges.

Training set of noncovalent interactions

Increasing interest in correct description of the noncovalent interactions brings a plethora of new methods and solutions of the problem, mainly in the field of the ever-more popular density functional theory (DFT). Typically, a new method is parameterised and tested on a small number of hydrogen bonded systems and a few (sometimes only one) dispersion bonded complexes. Often only an application of this method to a realistic problem reveals deficiencies which were not apparent in the original work. On the other hand, it may be difficult for a promising method to gain the credit of users if it is not tested appropriately. In our opinion, one of the most important reasons for this unsatisfying situation is lack of a reliable reference set of data to quickly assess the quality of a newly designed method.

Here we present a set of weakly bonded molecular complexes (see Table 3), which we believe can provide reasonable assessment of performance of a tested method. For the sake of systematic comparison of different methods for a particular type of interactions we divided our set into three subsets, hydrogen bonded, complexes with predominant dispersion interaction (for the sake of simplicity referred to as "dispersion bonded" in the following text) and mixed complexes (this categorization is based on our unpublished SAPT calculations). We are certainly aware that every such classification is to some extent arbitrary. For instance, hydrogen bonded complexes are predominantly electrostatic but also dispersion contribution is not negligible. Also in our "dispersion bonded" complexes dipolar interaction (AT stack, pyrazine dimer stack, ...) or multipolar interactions (benzene dimers, ethene dimer, ...) contribute. For more rigorous classification we refer the reader to, *e.g.*, ref. 32.

When working with a set of just a few molecules it is of the utmost importance that they represent all characteristic types

Table 2 Interaction energies of amino acid pairs (kcal mol^{-1}). Abbreviations used in the first column are the same as in the original reference and represent conventional one-letter symbols of amino acids while the number stand for position of respective residuum in protein. The last six complexes contain also the PDB code of the protein (in parentheses)

Complex	Ref.	aD[a]	aT[a]	aD→aT Helgaker[b]	ΔCCSD(T)	E_{int}[c] (Best)
F30-K46	2	−3.10	−3.30	−3.40	0.30	−3.10
F30-L33	2	−4.90	−5.30	−5.50	0.50	−5.00
F30-Y13	2	−4.20	−4.40	−4.50	0.60	−3.90
F30-F49	2	−3.10	−3.30	−3.30	—	−3.30
F30-Y4	2	−6.50	−6.80	−7.00	—	−7.00
F49-C39	2	−1.70	−2.00	−2.10	—	−2.10
F49-C6	2	−4.40	−4.80	−5.00	—	−5.00
F49-K46	2	−4.00	−4.60	−4.80	—	−4.80
F49-V5	2	−5.60	−6.40	−6.70	—	−6.70
F49-Y37	2	−2.30	−2.40	−2.50	—	−2.50
F49-Y4	2	−2.70	−3.00	−3.10	—	−3.10
F49-PB (Y4-V5)	2	−3.00	−3.10	−3.20	0.40	−2.80
F49-PB (V5−C6)	2	−7.90	−8.50	−8.80	0.60	−8.20
E47-K6 (1IU5)	41	−80.21	−80.61	−80.78	0.05	−80.73
E49-K6 (1BQ9)	41	−115.39	−114.11	−113.57	0.22	−113.35
E54-K2 (1SMM)	41	−109.83	−94.52	−88.06	−0.23	−88.29
E50-K30 (1BRF)	41	−58.47	−59.71	−60.23	−0.13	−60.36
E50-K52 (1BRF)	41	−99.35	−97.78	−97.12	−0.02	−97.14
E49-K6 (1BRF)	41	−72.48	−73.72	−74.24	—	−74.24

[a] Basis set: aug-cc-pVXZ. [b] The extrapolated interaction energy — see Methods. [c] E_{int} is the interaction energy, *i.e.*, it is not corrected for the deformation energy of monomers.

Table 3 Interaction energies (in kcal mol^{-1}) for model complexes (set S22). Deformation energy of monomers is not included

No.	Complex (symmetry)	$\Delta E^{MP2}{}_1{}^a$	$\Delta E^{MP2}{}_2{}^a$	$\Delta E^{MP2}{}_{CBS}{}^a$	$E^{int}{}_{CCSD(T)/CBS}{}^b$	Geometry^c
Hydrogen bonded complexes (7)						
1	$(NH_3)_2$ (C_{2h})	−3.02 (QZ)	−3.10 (5Z)	−3.20	**−3.17 (qz)**	**CCSD(T)/QZ**
2	$(H_2O)_2$ (C_s)	−4.75 (QZ)	−4.89 (5Z)	−5.03	**−5.02 (qz)**	**CCSD(T)/QZ**
3	Formic acid dimer (C_{2h})	−17.88 (QZ)	−18.23 (5Z)	−18.60	−18.61 (tz)	CCSD(T)/TZ
4	Formamide dimer (C_{2h})	−15.19 (QZ)	−15.52 (5Z)	−15.86	−15.96 (tz)	CCSD(T)/TZ
5	Uracil dimer (C_{2h})	−19.90 (TZ)	−20.28 (QZ)	−20.61	−20.65 (tz−fd)	MP2/TZ-CP
6	2-pyridoxine · 2-aminopyridine (C_1)	−15.91 (TZ)	−16.77 (QZ)	−17.37	−16.71 (tz-fd)	MP2/TZ-CP
7	Adenine · thymine WC (C_1)	−14.92 (TZ)	−15.89 (QZ)	−16.54	−16.37 (dz)	MP2/TZ-CP
Complexes with predominant dispersion contribution (8)						
8	$(CH_4)_2$ (D_{3d})	−0.42 (QZ)	−0.46 (5Z)	−0.51	−0.53 (qz)	CCSD(T)/TZ
9	$(C_2H_4)_2$ (D_{2d})	−1.43 (QZ)	−1.57 (5Z)	−1.62	−1.51 (qz)	CCSD(T)/QZ
10	Benzene · CH_4 (C_3)	−1.66 (QZ)	−1.75 (5Z)	−1.86	−1.50 (tz−fd)	MP2/TZ-CP
11	Benzene dimer (C_{2h})	−4.70 (aT)	−4.85 (aQ)	−4.95	−2.73 (adz)	MP2/TZ-CP
12	Pyrazine dimer (C_s)	−6.56 (aT)	−6.76 (aQ)	−6.90	−4.42 (tz-fd)	MP2/TZ-CP
13	Uracil dimer (C_2)	−10.63 (TZ)	−10.99 (QZ)	−11.39	−10.12 (tz-fd)	MP2/TZ-CP
14	Indole · benzene (C_1)	−6.44 (TZ)	−7.42 (QZ)	−8.12	−5.22 (dz)	MP2/TZ-CP
15	Adenine · thymine stack (C_1)	−12.30 (TZ)	−13.83 (QZ)	−14.93	−12.23 (dz)	MP2/TZ-CP
Mixed complexes (7)						
16	Ethene · ethine (C_{2v})	−1.57 (QZ)	−1.62 (5Z)	−1.69	−1.53 (tz)	CCSD(T)/QZ
17	Benzene · H_2O (C_s)	−3.28 (QZ)	−3.43 (5Z)	−3.61	−3.28 (tz-fd)	MP2/TZ-CP
18	Benzene · NH_3 (C_s)	−2.44 (QZ)	−2.57 (5Z)	−2.72	−2.35 (tz-fd)	MP2/TZ-CP
19	Benzene · HCN (C_s)	−4.92 (aT)	−5.06 (aQ)	−5.16	−4.46 (tz-fd)	MP2/TZ-CP
20	Benzene dimer (C_{2v})	−3.46 (aT)	−3.55 (aQ)	−3.62	−2.74 (adz)	MP2/TZ-CP
21	Indole · benzene T-shape (C_1)	−6.16 (TZ)	−6.65 (QZ)	−7.03	−5.73 (dz)	MP2/TZ-CP
22	Phenol dimer (C_1)	−6.71 (TZ)	−7.33 (QZ)	−7.76	−7.05 (tz-fd)	MP2/TZ-CP

[a] MP2 energy for a smaller basis set (1), larger basis set (2), and the extrapolated value (CBS); the basis set abbreviations TZ, aTZ, QZ, aQZ and 5Z (in brackets) stand for cc-pVTZ, aug-cc-pVTZ, cc-pVQZ, aug-cc-pVQZ and cc-pV5Z, respectively. [b] Interaction energy (deformation energy of monomers is not included) $E^{int}{}_{CCSD(T)/CBS}$ = MP2 CBS energy +ΔCCSD(T) term (see eqn (3)). The basis sets used to evaluate ΔCCSD(T) term dz, tz, qz and tz-fd (in brackets) stand for cc-pVDZ, cc-pVTZ, cc-pVQZ and a modified cc-pVTZ, respectively. In the modified cc-pVTZ set one set of f- and one set of d-functions were removed (only the more diffuse d function was kept) and the hydrogen basis set was modified analogically. [c] A method and a basis set used for optimization; full gradient optimizations with analytical (MP2) or numerical (CCSD(T)) gradients. Basis set abbreviations as in, aTZ-CP means that counterpoise corrected gradient optimization was applied.

of non-covalent interactions as well as possible. Ideally, both absolute and relative strengths of hydrogen bonded and stacked interactions should be similar to the average values found in the systems of interest—*e.g.*, the nucleic acids or proteins. In the S22 set, the dispersion bonded complexes are as numerous as hydrogen bonded ones, but they contribute to the sum of the stabilization energies by less than 40%, which is in line with our previous calculations on DNA[12] and nucleic acids.[2] A sufficient amount of them is also necessary for a reliable assessment of the DFT based methods (with current density functionals, the DFT methods perform reasonably well for hydrogen bonded complexes,[33] but fail for stacks).[34] It is also important that the set spans a wide range of interaction strengths in order to represent the diversity of interactions in macromolecules (note that we use the word "represent" here in a different sense from the one used in the works of Truhlar group).[32,35] In each of the above mentioned subgroups the stability of the complexes ranges between 3 and 20, 0.5 and 15, and 1.2 and 8 kcal mol^{-1}, respectively. The number of complexes, 22, is, again, arbitrary. For practical reasons, a smaller number would be certainly advantageous, however, smaller sets could became imbalanced regarding the relative amount of different interaction energy contributions or due to under-representation of larger molecules (see the discussion on the size of the dispersion bonded model complex above).

Our data are of high level quality and should be close to their CCSD(T)/CBS limit. Interaction energies were calculated using the same techniques as described above, but with larger basis sets both for the MP2 and CCSD(T) calculations. This guarantees significantly improved accuracy of the results. We have also used better quality geometries (for instance, H_2O dimer: CCSD(T)/cc-pVQZ, AT stack: counterpoise-corrected gradient optimization at the MP2/cc-pVTZ level), which accounts for some portion of differences between Table 3 and 1). While good E_{int} estimates of the smallest complexes are available from different sources (see, *e.g.*, similar sets of Zhao and Truhlar[36–38] or Grimme[39] and references therein), our results for the larger complexes represent the most reliable estimates up to date. Just these relatively large complexes constitute the main difference with respect to the typical training sets.

A practical question may be asked—is it really necessary to include complexes as large as adenine···thymine in the assessment set; confining the set to smaller molecules would save computer time. We believe it is advisable for the following reason: the rare gas dimers as typical models of dispersion bonded complexes appear to be rather poor representatives for dispersion interactions in large molecules. For instance, PW91 functional, which overestimates interaction in He_2, Ne_2, and Ar_2 by several hundreds of percent[40] turned out to substantially underestimate interaction energies in the large molecules (−1.8 kcal mol^{-1} for AT stack with TZVP basis set compared to an accurate value of −12.2 kcal mol^{-1}). Also, the density functionals which perform acceptably for rare gas dimers fail for larger molecules.[34] According to our calculations, similar

discrepancies can be found also comparing rare gas dimers results with results for $(CH_4)_2$, $(C_2H_4)_2$ and benzene dimer. In our experience, this misbehavior, namely that the rare gas atoms dimers are poor models for larger dispersion bonded complexes, is systematic within the current LDA and GGA functionals. In the light of these data, the need for the larger molecules in the assessment tests for especially the DFT based methods is obvious.

4 Conclusions

MP2 and CCSD(T) CBS interaction energies and geometries for more than 100 DNA base pairs, amino acid pairs and model complexes are for the first time presented together. The JSCH-2005 benchmark set, which is now available to the chemical community, can be used for testing of lower-level computational methods. For the first screening the smaller training set (S22) containing 22 smaller model complexes can be recommended.

Analysis of this extended set of data showed that very reasonable estimates of the complete basis set interaction energies in DNA and proteins can be obtained employing a two point extrapolation scheme with a pair of computationally accessible basis sets aug-cc-pVDZ and aug-cc-pVTZ. However, MP2 level of theory is insufficient and whenever significant dispersion contribution is expected a correction for higher order correlation effects must be applied.

When relative stabilities of different complexes or conformers are to be compared, the MP2 results must always be corrected for the higher order correlation effects whenever non-negligible dispersion contribution is expected. In such cases the higher order correlation effects affect ordering on the energy scale more than the basis set size if at least aug-cc-pVDZ basis set is used.

Acknowledgements

This work was supported by grants from the Grant Agency of the Czech Republic, Grant Agency of the Academy of Sciences of the Czech Republic and MŠMT of the Czech Republic (203/05/0009, A400550510, and LC512); further it was part of the research project Z4 055 0506.

References

1 K. Muller-Dethlefs and P. Hobza, *Chem. Rev.*, 2000, **100**, 143–167.
2 J. Vondrášek, L. Bendová, V. Klusák and P. Hobza, *J. Am. Chem. Soc.*, 2005, **127**, 2615–2619.
3 S. Tsuzuki, K. Honda, T. Uchimaru and M. Mikami, *J. Chem. Phys.*, 2005, **122**, 144323–144331.
4 M. O. Sinnokrot and C. D. Sherrill, *J. Am. Chem. Soc.*, 2004, **126**, 7690–7697.
5 S. Tsuzuki, K. Honda, T. Uchimaru, M. Mikami and K. Tanabe, *J. Am. Chem. Soc.*, 2000, **122**, 3746–3753.
6 P. Hobza and J. Šponer, *Chem. Rev.*, 1999, **99**, 3247–3276.
7 J. Pittner and P. Hobza, *Chem. Phys. Lett.*, 2004, **390**, 496–499.
8 A. Perez, J. Šponer, P. Jurečka, P. Hobza, F. J. Luque and M. Orozco, *Chem.-Eur. J.*, 2005, **11**, 5062–5066.
9 P. Jurečka and P. Hobza, *J. Am. Chem. Soc.*, 2003, **125**, 15608–15613.
10 J. Šponer, P. Jurečka and P. Hobza, *J. Am. Chem. Soc.*, 2004, **126**, 10142–10151.
11 P. Jurečka, J. Šponer and P. Hobza, *J. Phys. Chem. B*, 2004, **108**, 5466–5471.
12 I. Dąbkowska, H. V. Gonzalez, P. Jurečka and P. Hobza, *J. Phys. Chem. A*, 2005, **109**, 1131–1136.
13 P. Hobza and J. Šponer, *J. Am. Chem. Soc.*, 2002, **124**, 11802–11808.
14 A. Halkier, T. Helgaker, P. Jorgensen, W. Klopper and J. Olsen, *Chem. Phys. Lett.*, 1999, **302**, 437–446.
15 A. Halkier, T. Helgaker, P. Jorgensen, W. Klopper, H. Koch, J. Olsen and A. K. Wilson, *Chem. Phys. Lett.*, 1998, **286**, 243–252.
16 D. G. Truhlar, *Chem. Phys. Lett.*, 1998, **294**, 45–48.
17 P. L. Fast, M. L. Sanchez and D. G. Truhlar, *J. Chem. Phys.*, 2000, **113**, 3931–3931.
18 P. L. Fast, M. L. Sanchez and D. G. Truhlar, *J. Chem. Phys.*, 1999, **111**, 2921–2926.
19 S. F. Boys and F. Bernardi, *Mol. Phys.*, 2002, **100**, 65–73.
20 P. Jurečka and P. Hobza, *Chem. Phys. Lett.*, 2002, **365**, 89–94.
21 I. Dąbkowska, P. Jurečka and P. Hobza, *J. Chem. Phys.*, 2005, **122**, 204322–204329.
22 F. Weigend and M. Haser, *Theor. Chem. Acc.*, 1997, **97**, 331–340.
23 F. Weigend, M. Haser, H. Patzelt and R. Ahlrichs, *Chem. Phys. Lett.*, 1998, **294**, 143–152.
24 P. Jurečka, P. Nachtigall and P. Hobza, *Phys. Chem. Chem. Phys.*, 2001, **3**, 4578–4582.
25 R. Ahlrichs, M. Bar, M. Haser, H. Horn and C. Kolmel, *Chem. Phys. Lett.*, 1989, **162**, 165–169.
26 T. H. Dunning, *J. Chem. Phys.*, 1989, **90**, 1007–1023.
27 R. A. Kendall, T. H. Dunning and R. J. Harrison, *J. Chem. Phys.*, 1992, **96**, 6796–6806.
28 Molpro, version 2002.6, a package of *ab initio* programs H.-J. Werner, P. J. Knowles, R. Lindh, M. Schütz, P. Celani, T. Korona, F. R. Manby, G. Rauhut, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, A. W. Lloyd, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, P. Palmieri, R. Pitzer, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni and T. Thorsteinsson, 2003, see http://www.molpro.net.
29 D. Reha, H. Valdes, J. Vondrasek, P. Hobza, A. Abu-Riziq, B. Crews and M. S. de Vries, *Chem.-Eur. J.*, 2005, **11**, 6803–6817.
30 I. K. Yanson, A. B. Teplitsky and L. F. Sukhodub, *Biopolymers*, 1979, **18**, 1149–1170.
31 M. Kabeláč and P. Hobza, *J. Phys. Chem. B*, 2001, **105**, 5804–5817.
32 B. J. Lynch and D. G. Truhlar, *J. Phys. Chem. A*, 2003, **107**, 8996–8999.
33 S. Sirois, E. I. Proynov, D. T. Nguyen and D. R. Salahub, *J. Chem. Phys.*, 1997, **107**, 6770–6781.
34 J. Černý and P. Hobza, *Phys. Chem. Chem. Phys.*, 2005, **7**, 1624–1626.
35 N. E. Schultz, Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 4388–4403.
36 Y. Zhao and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2005, **7**, 2701–2705.
37 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 5656–5667.
38 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 6624–6627.
39 S. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463–1473.
40 Y. K. Zhang, W. Pan and W. T. Yang, *J. Chem. Phys.*, 1997, **107**, 7921–7925.
41 D. Horinek and P. Hobza, in preparation.