JCTC Journal of Chemical Theory and Computation

Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set

Lucie Gráfová,[†] Michal Pitoňák,*,[‡,¶] Jan Řezáč,[‡] and Pavel Hobza*,[‡,§]

Department of Computing and Control Engineering, Faculty of Chemical Engineering, Institute of Chemical Technology, Technická 5, CZ-166 28 Prague, Czech Republic, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i. and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, 166 10 Prague 6, Czech Republic, Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava 4, Slovak Republic, and Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic

Abstract: In this paper, an extension of the S22 data set of Jurecka et al. (Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* 2006, *8*, 1985.), the data set of benchmark CCSD(T)/CBS interaction energies of twenty-two noncovalent complexes in equilibrium geometries, is presented. The S22 data set has been extended by including the stretched (one shortened and three elongated) complex geometries of the S22 data set along the main noncovalent interaction coordinate. The goal of this work is to assess the accuracy of the popular wave function methods (MP2-, MP3- and, CCSD-based) and density functional methods (with and without empirical correction for the dispersion energy) for noncovalent complexes based on a statistical evaluation not only in equilibrium, but also in nonequilibrium geometries. The results obtained in this work provide information on whether an accurate and balanced description of the different interaction types and complex geometry distortions can be expected from the tested methods. This information has an important implication in the calculation of large molecular complexes, where the number of distant interacting molecular fragments, often in far from equilibrium geometries, increases rapidly with the system size. The best performing WFT methods were found to be the SCS-CCSD (spin-component scaled CCSD, according to Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* 2008, *128*, 124111), MP2C (dispersion-corrected MP2, according to Hesselmann, A. *J. Chem. Phys.* 2008, *128*, 144112), and MP2.5 (scaled MP3, according to Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* 2009, *10*, 282.). Since none of the DFT methods fulfilled the required statistical criteria proposed in this work, they cannot be generally recommended for large-scale calculations. The DFT methods still have the potential to deliver accurate results for large molecules, but most likely on the basis of an error cancellation.

1. Introduction

The benchmark interaction energies of noncovalent complexes are extremely valuable and even indispensable data for the development and testing of new approximate methods designed to treat these applications. Of the several established

* To whom correspondence should be addressed: E-mail: pitonak@fns.uniba.sk (M.P.); pavel.hobza@uochb.cas.cz (P.H.).

[†] Department of Computing and Control Engineering, Faculty of Chemical Engineering, Institute of Chemical Technology, Technická 5.

[‡] Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i. and Center for Biomolecules and Complex Molecular Systems.

[¶] Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University.

[§] Department of Physical Chemistry, Palacký University.

2366 *J. Chem. Theory Comput., Vol. 6, No. 8, 2010*

Gráfová et al.

series of high-quality interaction energy databases, the S22 data set[1] of the Hobza group seems to be the most extensively used. In the vast majority of cases, wave function theory methods (WFT) (SCS(MI)-MP2,[2] MP2.5,[3] MP2C,[4,5] etc.), density functional theory (DFT) methods (DFT-D,[6,7] M06−2X,[8] etc.), or semiempirical methods[9−11] designed for the calculation of noncovalent interaction were either optimized toward the best performance on this data set or their performance was at least tested on the S22.[12]

Since the S22[1] was introduced in 2006, extensive research in the field of noncovalent interactions has revealed several limitations of this data set. The first of these drawbacks, the resolution of which is not the goal of this paper, originates from the selection of the benchmark method. In S22, it is the coupled-cluster singles, doubles, and perturbative triples, close to the complete basis set limit (CCSD(T)/CBS). The inaccuracies of the CCSD(T), taken as an approximation to the full configuration interaction (FCI), discussed in several papers,[13−16] are clearly not the major source of the errors. It is the incompleteness of the basis set that leads to deviations up to 0.5−0.6 kcal/mol (S22 CCSD(T)/CBS interaction energies range from −20.65 to −0.53 kcal/mol), but they are typically smaller, 0.1−0.3 kcal/mol, according to estimates from refs 16−18, confirmed by a complete reinvestigation of the S22 conducted by Sherrill.[19]

The limitation, which is of the main concern in this work, is related to the geometry of complexes in the S22. Without a single exception, the geometries of the S22 complexes correspond to the geometries of the global minima of a particular complex, in almost all cases obtained by the Møller−Plesset second-order (MP2) counterpoise-corrected gradient optimization in Dunning's cc-pVTZ basis set.[20] On the one hand, it is natural to insist on the best performance of a particular method on equilibrium complex geometries, but for real-life applications, this is certainly not sufficient. The structure of biomacromolecules is determined by the interaction of their building blocks. In the case of DNA and RNA, it is by nucleic acid bases, whereas in the case of proteins by amino acids. Only the neighboring pairs are roughly in an equilibrium geometry, while all of the other pairs are far from being in equilibrium. The so-called "clashes" in DNA,[21] or the tremendous variety of mutual amino acid residue orientations found in proteins[22] could serve as rather apparent examples. The nearest pairs obviously play the most important role, but all the other pairs contribute non-negligibly as well. The thermal effects can also cause temporary deviations of the geometry away from the equilibrium even for close contacts. Generally speaking, most of the noncovalent interactions prevalent in biological systems, in terms of distinct interaction types, are rather indirectional, or at least, have a very flat potential energy surface. The only exceptions might be the electrostatic-energy-dominated interactions, like the hydrogen or halogen bond, which have fairly strict geometrical constraints.

Interactions between nonequilibrium structures inherently take place in the calculation of large (e.g., of about hundreds of atoms and more) molecules. Presuming that the local interacting neighborhood is arranged in the "close-to-equilibrium" geometry, the middle- to long-range interaction of distant molecular fragments happens between significantly more disordered structures. Despite the fact that the contribution to the inter- or intramolecular stabilization from such interactions is much smaller in absolute values, their number increases quickly. Thus, the eventual accumulation or cancellation of errors because of such interactions certainly has to be taken into consideration. So far, these effects have not been explored enough for us to be able to estimate their magnitude accurately. A good example of such a case is, for instance, the PBE-D[23] calculations of the interaction of the hydrogen molecule with π-systems of increasing size, starting from benzene through naphthalene to the graphene surface;[24] the accuracy is enhanced upon π-system growth because of error cancellation between the inaccuracies in the equilibrium (the hydrogen molecule with the closest sp^2 carbon π-system) and the accumulated asymptotic interaction from the distant graphene surface. In even larger molecules or other large (complex) molecular systems, the mutual orientation of distant interacting fragments is expected to be almost random. Thus, an ill asymptotic behavior of a method can be expected to cancel itself out. This may not, however, be true for highly organized molecular systems (molecular crystals for instance), where the asymptotic failure of a method may accumulate and eventually propagate into an artificial (de)stabilization in a particular direction.

Finally, quickly evolving and nowadays quite feasible "on-the-fly" ab initio molecular dynamics sample large areas of the potential energy surface, far from the equilibrium. By virtue of all the arguments and examples above, it is thus clear that a database of equilibrium structures cannot properly facilitate the parametrization and testing of any new, approximate method suitable for such applications. The inclusion of nonequilibrium structures is clearly important, but certain effects, like error accumulation or cancellation in large-scale molecular complexes, cannot be simply extrapolated from their performance on small dimers. Nevertheless, the unphysical asymptotic performance indicates that the results obtained for large molecular systems should be accepted with caution, if possible validated with respect to another reliable theoretical or experimental source.

The last issue concerning the S22 database, which in our opinion deserves reinvestigation, is related to the type of interactions represented in the S22. The complexes in the S22 are divided into three groups, that is, hydrogen-bonded, dispersion-dominated and "mixed-character" interactions. This division, carried out "intuitively" by the authors of ref 1, is clear in about half of the cases (i.e., the hydrogen-bonded complexes, obviously), but in some cases it is not so clear-cut (e.g., the ethene dimer being "dispersion-dominated", whereas the ethene···ethine complexes are categorized as "mixed-character" complexes). Any statistical evaluation of performance on a method that is based on this categorization may thus lead to dubious, "general" conclusions. In this work, we will propose a more physical division, based exclusively on the highly accurate interaction energy decomposition obtained by the DFT-SAPT method.[25] In so doing, we propose a, still rough, division into the electrostatics-dominated, dispersion-dominated and "mixed-character" complexes. As discussed below, this division is still not

refined enough to distinguish even qualitatively different interaction types within the S22 (like the dispersion-dominated $\pi-\pi$ stacking interaction in the benzene dimer and the "aliphatic" dispersion in the methane dimer), but it will serve the purpose of this work better than the former intuitive division. Besides these issues, we are aware of other drawbacks of the S22 database, like its having only a few representatives of dispersion-dominated interaction other than $\pi-\pi$ stacking, electrostatics-dominated complexes being represented solely by the H-bonded complexes, the large span of the complex interactions energies (about 20 kcal/mol) leading to the eventual favoring of complexes with large interaction energies over those with small interaction energies (if the absolute, rather than relative, interaction energies are used for parametrization and testing purposes), and last of all, the fact that certain interaction types (like the cation$\cdots\pi$, halogen-bond, etc.) are not covered by the S22 at all. Such improvements are clearly beyond the scope of this article and are the subject of ongoing research in our laboratory.

In this work, we will propose the so-called "S22 × 5" database, which is essentially an extension of the S22 from ref 1. for complexes in nonequilibrium geometries. For each complex from the original S22, the S22 × 5 database contains four additional stretched structures (i.e., ×5), one (by 10%) shortened, and three elongated (by 20%, 50%, and 100%) structures along the "interaction coordinate", generated in a systematic way to respect the character of the interaction. For details, please see section 2.1. For each such structure, we have calculated an estimate of the CCSD(T)/CBS interaction energy by following the identical methodology as used in the original S22[1] (for details, again, please see section 2.1). Despite the fact that, since the date of the publication of the original S22 paper, the computational possibilities have allowed for the use of more extended basis sets (needed mostly for the CCSD(T)), we have decided to follow the original methodology for several reasons. First of all, even though the CCSD(T) calculations (being the bottleneck) are feasible nowadays in larger basis sets, calculating five different distortions for each of the twenty-two complexes would be quite time-consuming. Second, the error bars of the present methodology (stabilization energies being overestimated by about 0.5 kcal/mol, ranging in absolute values from −20.65 to −0.53 kcal/mol), are fairly accurately established either from the series of our benchmark papers on the benzene dimer,[16] uracil dimer,[17] or methyl-adenine$\cdots$methyl-thymine,[18] or by other authors.[19,26]

In parallel with our research, Molnar et al.[27] has published a similar study on the potential-energy curves obtained by a variation of the distances of the centers of mass of the monomers for twenty of the complexes from the S22. The authors of that study, however, focused more on the performance of the methods applicable to truly large-scale calculations, like the MP2 and local MP2 (with a special focus on the effect of the basis set superposition error (BSSE)) in small basis sets, DFT methods without non-local exchange and semiempirical methods. Contrary to their aims, the focus of this work is the testing of a broader range of methods, scaling as N^3-N^6 with the system's size, applied close to the complete basis set limit. The conclusions presented here should thus draw a broader picture about the "limit" accuracy of the methods, perhaps suggesting the best candidate in terms of computational demands vs accuracy ratio, without providing too much space for error cancellation arising from the basis set unsaturation or BSSE. In a sense, the paper of Molnar et al. thus can be considered complementary rather than competing.

2. Methodology

2.1. Methods and Geometries. The nonequilibrium geometries were prepared by shifting one of the monomers from the equilibrium complex geometry, without modifying the intramolecular coordinates. Both the axis along which the molecule is shifted (i.e., the interaction coordinate) and the distance from which the relative displacements 0.9, 1.0 (i.e., the original S22 geometry), 1.2, 1.5, and 2.0 were calculated are defined by selected atoms or the centers of mass of groups of atoms in each of the monomers. For instance, in hydrogen-bonded complexes, these are only the hydrogen atom involved in (one of) the hydrogen bonds of a monomer and its acceptor in the other monomer. This choice of the interaction coordinate ensures the consistent treatment of the different types of interactions.

As already mentioned, the estimates of the CCSD(T)/CBS interaction energy were calculated analogously to ref 1, that is, MP2/CBS being a two-point "$1/X^3$" extrapolation according to Halkier et al.[28] utilizing the MP2 interaction energies in the(aug)-cc-pVXZ and (aug)-cc-pV(X+1)Z (X = T, Q, 5, depending on the complex size) basis sets, augmented by the ΔCCSD (T) higher-order correlation term from (aug)-cc-pVXZ (X = D, T, T′, Q; T′ being the cc-pVTZ basis set with f-functions removed and all but the most diffuse d-function removed). The only exception was the stacked indole$\cdots$benzene complex, in the case of which the aug-cc-pVDZ was used for calculation of the ΔCCSD(T) term because the originally used cc-pVDZ basis set was found to be strongly unsaturated.[5] The MP2C, MP2.5, and SCS-CCSD calculations were conducted using the same methodology, that is, the ΔX (X being MP2C, etc.), calculated in the same basis set as the ΔCCSD(T), combined with the MP2/CBS. The spin-component scaled MP2 methods, SCS-MP2,[29] SOS-MP2,[30,31] SOS(MI)-MP2,[2] and SC-S(MI)-MP2,[2] were calculated in cc-pVTZ, except for the smallest complexes (the ammonia dimer, water dimer, methane dimer, and ethene dimer), in the case of which the cc-pVQZ basis set was used. The basis sets used for the DFT calculations (TZVP, TZVPP, and QZVP, taken from TUR-BOMOLE library;[32] LP = 6-311++G(3df,3pd)[33] and aDZ = aug-cc-pVDZ) are explicitly shown in the relevant tables. The frozen core approximation was applied. PBE0AC exchange-correlation functional with density fitting was used for the DFT-SAPT calculations. The basis sets used for calculation of the DFT-SAPT interaction energies and the differences (shift) between the vertical ionization potential (IP) and the highest occupied molecular orbital (HOMO) energy are specified in Table 2.

The MP2 and the DFT calculations (except for the M06-2X) were carried out using the TURBOMOLE code;[34] the

Table 1. Definition of the (Relative) Errors Used for the Statistical Evaluation of the Interaction Energies[a]

abreviation	error	equation
RMS(R)E	root-mean-square (relative)	$\sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(x_i)^2}$
M(R)E	mean (relative)	$\dfrac{1}{N}\sum_{i=1}^{N}x_i$
MA(R)E	mean absolute (relative)	$\dfrac{1}{N}\sum_{i=1}^{N}|x_i|$
MAX(R)	maximum absolute (relative)	$\max_i|x_i|$

[a] In the case of "relative" quantities, $x_i = 100[(E_{CCSD(T)} - E_{method})/E_{CCSD(T)}]$, the error is expressed as a percentage; otherwise, $x_i = E_{CCSD(T)} - E_{method}$, where index "*i*" indicates the molecular complex and "method" stands for the name of the tested method (MP2, SCS-MP2, etc.).

M06-2X calculations were done in Gaussian;[35] and the DFT-SAPT, SCS-MP2 (and all SOS and (MI) variants), MP2C, MP2.5, and CCSD(T) calculations were carried out using the MOLPRO[36] package. The SCS-CCSD and part of the CCSD(T) calculations were run using the MOLCAS[37] suite of programs.

2.2. Statistics. For the purpose of the statistical evaluation of the method's performance, we used root-mean-square error (RMSE), mean error (ME), mean absolute error (MAE), and the maximum absolute error (MAX), applied on the difference of the interaction energy obtained by a particular method and the CCSD(T)/CBS interaction energy. Analogously, we have used the corresponding "relative" quantities, root-mean-square relative error (RMSRE), mean relative error (MRE), mean absolute relative error (MARE), and the maximum absolute relative error (MAXR), that is, applied to the percentual error of the interaction energy, as obtained by a particular method and the CCSD(T)/CBS. For more details on the definition of these quantities, please see Table 1. The reason for using the relative errors in statistics is straightforward. The S22 itself, as already mentioned, has a large span of the absolute values of the interaction energies, and in addition, the interaction energies change significantly upon the stretching of the complex geometries. In so doing, all of the complexes in all of the geometries, regardless of their absolute interaction energy, contribute equally to the statistics.

To analyze several, qualitatively different, aspects of the method performance, we have proposed five criteria. These criteria will be further referred to as "the requirements" (i.e., RQ), RQ_1-RQ_5, because the method will be considered to "score" for a particular requirement, whether the underlying criteria are below the predefined threshold or fall into the predefined error bars.

- RQ_1: The MARE for all of the interaction types and geometry distortions is below a certain threshold. This is the simplest measure of the general accuracy of a method for all the database structures on a relative scale.

It is important to point out that the accuracy on a "relative scale" may seem to be stricter than accuracy on an absolute scale because of the contribution from all of the complexes (even with the smallest interaction energy). One can argue, however, that a 50% relative error on an interaction energy of 0.05 kcal/mol is irrelevant, but, in our opinion, the relative scale is the only transferable measure for complexes of different sizes and interaction strengths.

- RQ_2: The MARE of each category of interaction type (i.e., electrostatics-dominated, dispersion-dominated and mixed-character) for a particular geometry distortion should be similar. This criterion measures the uniform accuracy of a method for different interaction types for a particular geometry distortion. "Uniform accuracy" means that it should provide similarly large mean absolute relative errors for all of the interaction types.
- RQ_3: The MARE of each geometry distortion (i.e., 1.0, 1.2, 1.5, and 2.0) of a particular interaction type should be similar. This criterion measures the uniform accuracy of a method for different geometrical distortions for a particular interaction type.
- RQ_4: Close RMSRE and MARE values for a particular interaction type and geometry distortion. This criterion measures the uniform accuracy for a particular interaction type and geometry distortion. The difference between the RMSRE and MARE is very sensitive to the presence of a value that is clearly outlying from the average MARE. Please note that this requirement can be fulfilled not only if the method performs equally well but also if it performs equally poorly for the whole subgroup.
- RQ_5: The MRE for all the interaction types and geometry distortions should be close to zero. This criterion measures the overall balance of the description, thus also a tendency to over- or underestimate the interaction energy.

For each requirement, we have introduced two ad hoc thresholds, that is, strict and lenient, which in our opinion reflect the typical expectations from an accurate method for noncovalent interaction calculation. For RQ_1, it is a MARE below 10% for a strict and 20% for a lenient threshold; for RQ_2 and RQ_3, it is a MARE within ±3% (±6%) in a strict (lenient) threshold; for RQ_4, the RMSRE being less than 3% (6%) larger than the MARE in the strict (lenient) threshold; and finally, for the RQ_5, it is the MRE being within ±3% (±6%) off from zero for the strict (lenient) threshold.

Since it is more natural to think in absolute rather than relative errors, we can demonstrate the impact of these thresholds for a model complex with an interaction energy of, for instance, −20 kcal/mol. In this case, applying the strict thresholds, the fulfillment of RQ_1 leads to an mean absolute (unsigned) error of 2 kcal/mol, the fulfillment of RQ_3 means that the mean absolute error on each geometry distortion (including all or the individual interaction types) differs from the others by less than 0.6 kcal/mol, etc.

Table 2. Distinction between the Complexes As Interaction Types, E, Electrostatics-Dominated, D, Dispersion-Dominated, M, Mixed-Character, According to the DFT-SAPT Analysis[a]

	DFT-SAPT	complex	geometry distortion				
			0.9	1.0	1.2	1.5	2.0
H-bonded	aQZ (QZVP)	ammonia dimer	E	E	E	E	E
	aQZ (QZVP)	water dimer	E	E	E	E	E
	aQZ (QZVP)	formic acid dimer	E	E	E	E	E
	aQZ (QZVP)	formamide dimer	E	E	E	E	E
	aTZ (QZVP)	2-PO···0.2-PY	E	E	E	E	E
	aDZ (TZVP)	uracil dimer H	E	E	E	E	E
	aDZ (TZVP)	adenine···thymine WC	E	E	E	E	E
dispersion dominated	aQZ (QZVP)	methane dimer	D	D	D	D	D
	aQZ (QZVP)	ethene dimer	D	D	D	D	M
	aTZ (QZVP)	benzene···methane	D	D	D	D	D
	aTZ (QZVP)	benzene dimer S	D	D	D	D	M
	aTZ (QZVP)	pyrazine dimer	D	D	D	M	M
	aTZ (QZVP)	uracil dimer S	E	M	M	M	M
	aDZ (TZVP)	indole···benzene S	D	D	D	D	D
	aDZ (TZVP)	adenine···thymine S	M	M	M	M	M
Mixed character	aQZ (QZVP)	ethene···ethyne	M	M	M	M	M
	aQZ (QZVP)	benzene···water	E	M	M	M	M
	aTZ (QZVP)	benzene···ammonia	M	M	M	M	M
	aTZ (QZVP)	benzene···HCN	E	E	M	M	M
	aTZ (QZVP)	benzene dimer T	M	D	D	D	M
	aTZ (TZVP)	indole···benzene T	M	M	M	M	M
	aTZ (QZVP)	phenole dimer	M	M	M	M	M
	count	electrostatics	10	8	7	7	7
		dispersion	6	7	7	6	3
		mixed	6	7	8	9	12

[a] The DFT-SAPT column shows the basis sets used for the DFT-SAPT and HOMO energy calculations and for calculations of the vertical IP (in parentheses). aXZ stands for Dunning's aug-cc-pVXZ basis sets, and TZVP[40] and QZVP[41] stand for basis sets of Ahlrichs and coworkers.

3. Results and Discussion

3.1. Interaction Types. DFT-SAPT interaction energy decomposition was used to evaluate the character of the interaction for each complex and geometry distortion. As already mentioned, in this work we considered three distinct interaction types: electrostatic-energy-dominated, dispersion-dominated and the mixed-character. A complex in a particular geometry is assigned to, for instance, the electrostatics-dominated type, if (the absolute value of) the electrostatic component (i.e., $E^{(1)}$) of the DFT-SAPT interaction energy is at least twice as large as the dispersion component ($E^{(2)}_{disp}$ + $E^{(2)}_{disp-exchange}$), and vise versa for the dispersion-dominated category. The complexes for which neither of these is true were assigned to the mixed-character interaction type group. The result of the SAPT-DFT analysis, along with the specification of the basis sets used for the DFT-SAPT and the monomer DFT shifts employed, are shown in Table 2. The arrangement of rows in the table is the same as in the original S22 paper. As clearly seen from the column corresponding to the geometry distortion 1.0, that is, the equilibrium S22 geometries, the number of the complexes in each category, 8/7/7, for electrostatics (E)/dispersion (D)/mixed (M), is almost identical to the count of hydrogen-bonded/dispersion/mixed in the original paper, that is, 7/8/7, but they are organized slightly differently (e.g., stacked uracil dimer and adenine···thymine dimer being "mixed" now, but T-shaped benzene dimer being, on the contrary, dispersion-dominated). However, upon complex dissociation, this ratio changes dramatically and in the distance elongated by 100%, that is, 2.0, the ratio is 7/3/12. This rather

anticipated consequence of the different asymptotic scaling of various interaction energy components, however, strongly influences the statistical evaluation of the method's accuracy because of the reorganization of the complexes between the different interaction type categories.

As will be analyzed below in more detail, the performance of almost all of the WFT/DFT methods for the 0.9 distortion is much worse when compared to the equilibrium and elongated structures. Including this distortion would thus completely contaminate the simultaneous statistics over all of the geometry distortions. That is the reason that in all of the following tables, except for those where it is clearly stated, the 0.9 geometry distortion has been excluded from the statistics. This finding is surprising and at the same time alarming. Since these distorted geometries may occur in biomacromolecules (e.g., tight clashes in DNA), we are going to investigate this feature in more detail in the near future.

Because of space limitation, we have selected only a few statistical markers, some of which are presented only in a qualitative way (e.g., "yes/no" instead of an exact value, etc.). Exact numerical values and several other statistical evaluations can be found in the Supporting Information.

3.2. WFT Methods. In Tables 3−8, we present the statistical analysis of the most frequently used WFT methods for noncovalent interaction energy calculations (ordered according to scaling of the computation requirements with the system size): SOS-MP2, SOS(MI)-MP2 (scaling N^4), MP2, SCS-MP2, SCS(MI)-MP2, MP2C (scaling N^5), MP2.5 (scaling N^6, noniterative), and SCS-CCSD (scaling N^6, iterative). More methods can be found in the Supporting

Table 3. WFT Methods: A Statistical Evaluation of All of the S22 × 5 Complexes

method/basis	RQ$_1$[a]		RQ$_2$		RQ$_3$		RQ$_4$		RQ$_5$			score[b] (%)
	10%	20%	±3%	±6%	±3%	±6%	3%	6%	±3%	±6%		
MP2C	yes	yes	yes	yes	yes	yes	yes	yes	no	yes		90
SCS-CCSD	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	(−)[c]	90
MP2.5	yes	yes	no	yes	yes	yes	yes	yes	no	no	(+)	70
SCS(MI)-MP2	no	yes	no	no	no	yes	no	yes	no	no		30
SCS-MP2	no	no	no	no	no	no	yes	yes	no	no	(−)	20
SOS-MP2	no	no	no	no	no	no	no	yes	no	no	(−)	10
MP2	no	no	no	no	no	no	no	no	no	no	(+)	0
SOS(MI)-MP2	no	no	no	no	no	no	no	no	no	no	(−)	0

[a] RQ$_1$, a small (<10% or <20%) mean absolute relative error (MARE) for all complexes and geometry distortions; RQ$_2$, a small (<±3% or <±6%) MARE variation for all of the investigated complexes for a particular geometry distortion; RQ$_3$, a small (<±3% or <±6%) MARE variation within a group of complexes of a particular interaction type for all of the geometry distortions; RQ$_4$, a balanced description within a group of complexes of a particular interaction type and geometry distortion evaluated by the small (<±3% or <±6%) difference between the MARE and root-mean-square relative error (RMSRE); RQ$_5$, a small (<±3% or <±6%) deviation of the mean relative error (MRE) from zero, indicating systematic over- or understabilization of all of the complexes and geometry distortions. For a more detailed definition and reasoning of these restrictions, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (ten in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

Information. Allow us first to analyze the results shown in Tables 3−6. In these tables, we present the "score" of each method and the requirement with the strict and the lenient threshold for the whole of S22 × 5 (Table 3) and for each interaction type separately (Tables 4−6). For easier orientation in each table, a horizontal line has been drawn separating the methods that reached a score above 70% (if any), which in our opinion represent methods capable of proper description of noncovalent complexes. The score is calculated as the number of fulfilled requirements, both strict and lenient, divided by the total number of requirement/threshold columns, converted to percents. Methods having the same score are ordered alphabetically.

In Table 3, the statistics for all of the categories of complexes are shown. The best performing methods, which in the spirit of the motivation of this work are expected to perform well for the calculation of biomacromolecules, are the MP2C, MP2.5. and SCS-CCSD, scoring at least to 70%. The methods performing the best were MP2C and SCS-CCSD with a score of 90%, which fully satisfies all of the requirements (four requirements with the strict and one, RQ$_5$, with the lenient threshold).

The most demanding requirement (not only for the entire S22 × 5 but also for different interaction types separately) was clearly RQ$_5$, that is, the criterion measuring the overall balance of the description across all of the interaction types and geometrical distortions (in other words, the tendency to over- or underestimate the interaction energy). None of the methods fulfilled the strict threshold for this requirement; among the best scoring methods, the MP2.5 did not even meet the lenient one. In Table 3, in the RQ$_5$ column, there is a sign indicating whether the method has a tendency to overstabilize (+) or understabilize (−) the complexes systematically. The best performing methods did not show the same tendency of these characteristics. The SCS-CCSD had a tendency to underestimate the interaction energies systematically, unlike the MP2.5, which had a tendency to overestimate the interaction energies systematically. The MP2C provided unbalanced description without an equable error orientation.

What is slightly surprising is the performance of the SCS(MI)-MP2 method, which was parametrized for the S22 test set and did not fulfill any of the requirements with the strict threshold. The performance of the SOS-MP2 and especially SOS(MI)-MP2 (parametrized for the S22 test set as well), confirmed the inability of the sole opposite spin contribution to describe the noncovalent interactions correctly. The zero score of the MP2 method is no surprise because of its known strong overestimations of the $\pi-\pi$ stacked complexes, which clearly is in conflict with four of our five requirements. The SCS-MP2 provided a poor score (20%), but on the other hand, surprisingly fulfilled the RQ$_4$ with the strict threshold, which means that the method is capable of delivering balanced accuracy for all of the interaction types and geometrical distortions.

We should now analyze the performance of the WFT methods for the different interaction types separately. Please notice that the score is now evaluated according to four/eight requirements/thresholds, since RQ$_2$ is not applicable here.

Exceptionally good performance was found for the electrostatics-dominated complexes (see Table 4). Except for the SOS/SCS-MP2 and SOS(MI)-MP2, all of the methods reached a score clearly above 70% (four methods even had a score of 100%) and thus can be safely recommended for the calculation of electrostatics-dominated noncovalent interactions (e.g., H-bonded complexes). This is, however, not so surprising, since already the MP2 method is known to describe the electric properties of molecules (dipole moment, quadrupole moment, etc.) fairly accurately. All of the methods fulfill requirement RQ$_4$ even with the strict threshold (i.e., the criteria measuring the uniform accuracy for a particular interaction type and geometry distortion), which might be a consequence of the fact that the electrostatics-dominated complexes are represented solely (except for the benzene···HCN) by the same "chemical" type of interaction, that is, hydrogen bond.

The performance of the WFT methods on the dispersion-dominated complexes, Table 5, is the worst. The reason is that the complexes in this category are mostly of the $\pi-\pi$

Table 4. WFT Methods: A Statistical Evaluation of the Electrostatics-Dominated Complexes of the S22 × 5 Set

method/basis	RQ_1[a]		RQ_3		RQ_4		RQ_5			score[b] (%)
	10%	20%	±3%	±6%	3%	6%	±3%	±6%		
MP2	yes	yes	yes	yes	yes	yes	yes	yes		100
MP2.5	yes	yes	yes	yes	yes	yes	yes	yes		100
MP2C	yes	yes	yes	yes	yes	yes	yes	yes		100
SCS(MI)-MP2	yes	yes	yes	yes	yes	yes	yes	yes		100
SCS-CCSD	yes	yes	yes	yes	yes	yes	no	yes	(−)[c]	88
SCS-MP2	no	yes	no	yes	yes	yes	no	no	(−)	50
SOS(MI)-MP2	no	yes	no	yes	yes	yes	no	no	(−)	50
SOS-MP2	no	no	no	no	yes	yes	no	no	(−)	25

[a] For a brief definition of requests, RQ_1, RQ_3, ..., RQ_5, see the footnote of Table 3. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

Table 5. WFT Methods: A Statistical Evaluation of the Dispersion-Dominated Complexes of the S22 × 5 Set

method/basis	RQ_1[a]		RQ_3		RQ_4		RQ_5			score[b] [%]
	10%	20%	±3%	±6%	3%	6%	±3%	±6%		
SCS-CCSD	yes	yes	yes	yes	yes	yes	yes	yes		100
MP2C	yes	yes	yes	yes	yes	yes	no	yes		88
MP2.5	yes	yes	yes	yes	yes	yes	no	no	(+)[c]	75
SCS(MI)-MP2	no	yes	no	yes	no	yes	no	no	(−)	38
SCS-MP2	no	no	no	no	yes	yes	no	no	(−)	25
SOS-MP2	no	no	no	no	no	yes	no	no		13
MP2	no	no	no	no	no	no	no	no	(+)	0
SOS(MI)-MP2	no	no	no	no	no	no	no	no		0

[a] For a brief definition of requests RQ_1, RQ_3, ..., RQ_5, see the footnote of Table 3. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

Table 6. WFT Methods: Statistical Evaluation of the Mixed-Character Complexes of the S22 × 5 Set

method/basis	RQ_1[a]		RQ_3		RQ_4		RQ_5			score[b] (%)
	10%	20%	±3%	±6%	3%	6%	±3%	±6%		
MP2C	yes	yes	yes	yes	yes	yes	yes	yes	(+)[c]	100
SCS-CCSD	yes	yes	yes	yes	yes	yes	yes	yes	(−)	100
MP2.5	yes	yes	yes	yes	yes	yes	no	yes	(+)	88
SCS(MI)-MP2	yes	yes	yes	yes	yes	yes	no	yes		88
SOS(MI)-MP2	no	yes	yes	yes	yes	yes	no	no		63
MP2	no	yes	yes	yes	no	yes	no	no	(+)	50
SCS-MP2	no	no	no	no	yes	yes	no	no	(−)	25
SOS-MP2	no	no	no	no	yes	yes	no	no	(−)	25

[a] For a brief definition of requests, RQ_1, RQ_3, ..., RQ_5, see the footnote of Table 3. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

stacking type, which are known to be the most challenging to describe in terms of the level of the WFT sophistication. The performance of the methods essentially followed the performance for the whole S22 × 5 test set, perhaps with only minor differences. The methods capable of the proper description of dispersion-dominated complexes are the SCS-CCSD, MP2C, and MP2.5, scoring above 70%. The best-performing method was the SCS-CCSD, which fulfilled all of the requirements with the strict thresholds.

The description of mixed-character complexes (see Table 6) was almost as good as that of the electrostatics-dominated ones. The methods that scored above 70% and can thus be safely used for the proper description of this class of complexes were the MP2C, SCS-CCSD (both with a score of 100%), MP2.5, and SCS(MI)-MP2. The biggest difference was in the performance of the MP2 method, which satisfied

all the requirements with strict thresholds for the electrostatics-dominated complexes but now reached only a 50% score. Its most significant drawback was in its slightly unbalanced description upon the interaction-coordinate elongation (RQ_4) with a tendency to overstabilize these complexes, as indicated by requirement RQ_5.

Table 7 summarizes the absolute and relative ME, MAE, and MAX values for all of the interaction types and individual geometry distortions. In this table, the values of these statistical parameters are explicitly shown also for distortion 0.9. It is clear that the values for the 0.9 distortion do not follow the same trends as for other distortions, being almost all above 100% (or below −100%) on the relative scale. It is only the SCS-CCSD method that performs reasonably well for this geometry distortion. From the data in the MAE and MAX columns, it can be observed that these

Table 7. WFT Methods[a]

method/basis	0.9			1.0			1.2			1.5			2.0		
MP2	1.34	1.35	5.46	0.81	0.82	3.46	0.32	0.37	1.48	0.10	0.14	0.48	0.03	0.04	0.12
	(>−100)[b]	(>100)	(>100)	(−17.89)	(18.33)	(79.20)	(−10.61)	(11.09)	(44.55)	(−9.94)	(10.39)	(48.99)	(−13.48)	(13.91)	(>100)
SCS-MP2	−2.25	2.25	5.88	−1.63	1.63	3.78	−0.85	0.85	2.30	−0.33	0.33	1.11	−0.11	0.11	0.47
	(>100)	(>100)	(>100)	(27.00)	(27.00)	(52.90)	(17.59)	(17.59)	(35.49)	(14.10)	(14.1)	(30.38)	(12.77)	(12.77)	(36.59)
SOS-MP2	−3.29	3.29	8.97	−2.32	2.32	5.69	−1.17	1.17	2.77	−0.45	0.45	1.29	−0.14	0.14	0.52
	(>100)	(>100)	(>100)	(40.22)	(40.22)	(69.37)	(25.64)	(25.64)	(46.65)	(20.72)	(20.72)	(41.25)	(19.85)	(19.85)	(78.37)
SOS(MI)-MP2	−1.02	1.78	5.59	−0.92	1.34	4.53	−0.63	0.77	3.11	−0.32	0.36	1.71	−0.14	0.15	0.85
	(−67.94)	(>100)	(>100)	(7.86)	(18.68)	(45.82)	(7.83)	(12.89)	(23.8)	(7.05)	(11.39)	(23.25)	(3.60)	(16.00)	(90.35)
SCS(MI)-MP2	0.05	0.43	1.12	0.00	0.30	0.78	−0.01	0.13	0.45	0.03	0.06	0.25	0.02	0.03	0.11
	(−0.15)	(27.14)	(>100)	(2.57)	(7.98)	(41.66)	(1.03)	(4.88)	(25.63)	(−0.34)	(4.66)	(20.33)	(−1.29)	(5.68)	(18.83)
MP2C	0.04	0.30	1.31	0.06	0.19	0.83	0.02	0.10	0.40	−0.01	0.04	0.21	−0.01	0.02	0.13
	(4.80)	(20.46)	(>100)	(−2.82)	(4.11)	(11.19)	(−1.68)	(2.56)	(5.98)	(−0.88)	(1.92)	(4.36)	(0.09)	(1.77)	(4.89)
MP2.5	0.26	0.29	0.85	0.16	0.18	0.55	0.06	0.09	0.26	0.02	0.04	0.13	0.01	0.02	0.06
	(−30.29)	(32.12)	(>100)	(−3.72)	(4.45)	(16.21)	(−2.40)	(2.93)	(9.43)	(−2.56)	(3.00)	(9.4)	(−3.45)	(3.77)	(16.71)
SCS-CCSD	−0.25	0.27	1.05	−0.20	0.21	0.79	−0.12	0.12	0.47	−0.06	0.06	0.23	−0.02	0.02	0.09
	(1.52)	(4.71)	(10.84)	(1.71)	(2.42)	(4.31)	(1.5)	(1.86)	(3.61)	(1.36)	(1.73)	(3.77)	(1.20)	(1.75)	(5.22)

[a] The values in rows: in the first line of each row, ME, MAE, and MAX (in kcal/mol), in the second line, the relative quantities, MRE, MARE, and MAXR (in %) for all of the interaction types and particular geometry distortion. The absolute and relative MAX(R) in a particular table cell do not necessarily correspond to the same complex. [b] The values above 100% (or below −100%) are for simplification shown as ">100". The actual values can be found in the Supporting Information.

Table 8. WFT Methods[a]

method/CBS	$(CH_4)_2$	$(C_6H_6)_2$	Δ[b]
SCS-CCSD	1.95	1.34	0.61
SOS-MP2	48.88	47.52	1.36
MP2C	4.32	1.71	2.61
MP2.5	2.93	11.38	−8.45
SOS(MI)-MP2	20.00	28.67	−8.68
SCS-MP2	36.66	24.93	11.73
SCS(MI)-MP2	26.61	3.87	22.75
MP2	1.70	54.91	−53.22

[a] The averaged absolute relative errors (AARE, in %) over all (except 0.9) geometry distortions for the methane and benzene dimers. [b] AARE(methane dimer)−AARE(benzene dimer).

"absolute-value-based" errors are far less important as compared to our statistics based on RQ₁ to RQ₅. According to the results in Table 3, the SCS-CCSD is a much better performing method than the SCS(MI)-MP2, despite the fact that both have almost the same MAE and MAX values for almost all of the geometrical distortions.

The last note in this section concerns the incapability of our statistical model to distinguish between certain interaction types, namely the $\pi-\pi$ stacking and aliphatic dispersion-dominated complexes. In Table 8, the absolute relative errors (ARE) for the methane and benzene dimers, averaged over the distortions 1.0−2.0 are shown. The last column in the table shows the difference between these two numbers, thus reflecting the balance of the simultaneous description of these two complexes. The first three methods, that is, SCS-CCSD, SOS-MP2, and MP2C, deliver exceptionally balanced descriptions, with the difference being below 3%. In the case of the SOS-MP2 method, it is only a consequence of error compensation, because the average ARE for both complexes are almost 50% off. This method thus cannot be expected to deliver a both accurate and balanced description of the dispersion-dominated complexes. Nevertheless, the most curious performance is found for the MP2 method. It is notoriously known to overestimate the stability of the $\pi-\pi$ stacked complexes (by about 55% are for benzene dimer), but its excellent performance for the methane dimer (1.7% ARE, that is, the most accurate of all the tested methods) is indeed surprising.

3.3. DFT Methods. A statistical assessment of the performance of the DFT methods is shown in Tables 9−14. We considered the following functionals (ordered according to scaling of the computation requirements with the system size): BLYP, PBE, TPSS (scaling N^3), M06-2X (scaling N^4), B2PLYP (scaling N^5), with and without empirical correction for the dispersion energy (if applicable). The M06-2X interaction energies were calculated both in "Fine" and "UltraFine" integration grid, as according to the GAUSSIAN options, with no noteworthy improvement, as could be hoped for according to ref 38.

When comparing the DFT results for all of S22 × 5, as is shown in Table 9, with the results for the WFT methods, the difference is striking. Only three methods had nonzero score, the best scoring method (PBE-D_G) achieved only 40%. None of the DFT methods can thus, according to our statistics, be clearly recommended for "black-box" application on the calculation of biomacromolecules. This does not necessarily mean that the use of the DFT methods is meaningless, but the possible results of the satisfactory accuracy obtained for large molecules are quite likely to be attributable to error cancellation.

In contrast with the generally poor performance of the DFT methods, the description of the electrostatics-dominated complexes (Table 10) is much better. According to our statistics, almost all tested DFT methods are capable of proper description of these complexes. Only three methods (B2PLYP-D, PBE- D_G, TPSS-D_G) have not reached the "critical" score of 70%. These results support the generally accepted fact that the DFT methods are suitable for the description of (at least small) H-bonded complexes. Clearly the most demanding requirement (not only for this category) was RQ₅ (like for the WFT methods), that is, the balanced treatment of all the geometry distortions. None of the methods fulfilled the strict threshold for this requirement. The best-performing methods did not have a similar tendency to over- or underestimate the stability of complexes (TPSS-D_J and PBE- D_J overestimated, B2PLYP and PBE underestimated and BLYP- D_J, BLYP-D_G and M06−2X provided unbalanced description without an equable error orientation).

Table 9. DFT Methods: A Statistical Evaluation of All the S22 × 5 Complexes

method/basis	RQ₁[a] 10%	20%	RQ₂ ±3%	±6%	RQ₃ ±3%	±6%	RQ₄ 3%	6%	RQ₅ ±3%	±6%		score[b] (%)
PBE-D$_G$/LP	no	yes	no	yes	no	yes	no	yes	no	no		40
BLYP-D$_J$/LP	no	no	no	no	no	yes	yes	yes	no	no	(−)	30
B2PLYP-D/TZVPP	no	no	no	no	no	no	no	yes	no	no		10
BLYP-D$_G$/LP	no	no	no	no	no	no	no	no	no	no		0
B2PLYP/TZVPP	no	no	no	no	no	no	no	no	no	no	(−)[c]	0
M06-2X/aDZ	no	no	no	no	no	no	no	no	no	no		0
PBE-D$_J$/LP	no	no	no	no	no	no	no	no	no	no	(+)	0
PBE/LP	no	no	no	no	no	no	no	no	no	no	(−)	0
TPSS-D$_G$/LP	no	no	no	no	no	no	no	no	no	no	(+)	0
TPSS-D$_J$/LP	no	no	no	no	no	no	no	no	no	no	(+)	0

[a] RQ₁, a small (<10% or <20%) mean absolute relative error (MARE) for all complexes and geometry distortions; RQ₂, a small (<±3% or <±6%) MARE variation for all of the investigated complexes for a particular geometry distortion; RQ₃, a small (< ± 3% or < ± 6%) MARE variation within a group of complexes of a particular interaction type for all of the geometry distortions; RQ₄, a balanced description within a group of complexes of a particular interaction type and geometry distortion evaluated by the small (<±3% or <±6%) difference between the MARE and root-mean-square relative error (RMSRE); RQ₅, a small (<±3% or <±6%) deviation of the mean relative error (MRE) from zero, indicating systematic over- or understabilization of all of the complexes and geometry distortions. For a more detailed definition and reasoning of these restrictions, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (ten in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

Table 10. DFT Methods: A Statistical Evaluation of the Electrostatics-Dominated Complexes of the S22 × 5 Set

method/basis	RQ₁[a]		RQ₃		RQ₄		RQ₅			score[b]
BLYP-D$_G$/LP	yes	yes	yes	yes	yes	yes	no	yes		88
M06-2X/aDZ	yes	yes	yes	yes	yes	yes	no	yes		88
TPSS-D$_J$/LP	yes	yes	yes	yes	yes	yes	no	yes	(+)[c]	88
BLYP-D$_J$/LP	yes	yes	yes	yes	yes	yes	no	no		75
B2PLYP/TZVPP	yes	yes	yes	yes	no	yes	no	yes	(−)	75
PBE/LP	yes	yes	yes	yes	yes	yes	no	no	(−)	75
PBE-D$_J$/LP	yes	yes	yes	yes	yes	yes	no	no	(+)	75
B2PLYP-D/TZVPP	no	yes	yes	yes	yes	yes	no	no	(−)	63
PBE-D$_G$/LP	no	yes	no	yes	yes	yes	no	no	(+)	50
TPSS-D$_G$/LP	no	yes	no	yes	yes	yes	no	no	(+)	50

[a] For a brief definition of requests RQ₁, RQ₃, ..., RQ₅, see the footnote of Table 9. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

Table 11. DFT Methods: A Statistical Evaluation of the Dispersion-Dominated Complexes of the S22 × 5 Set

method/basis	RQ₁[a]		RQ₃		RQ₄		RQ₅			score[b]
BLYP-D$_J$/LP	no	no	no	yes	yes	yes	no	no		38
PBE-D$_G$/LP	no	yes	no	yes	no	no	no	no	(+)	25
B2PLYP-D/TZVPP	no	no	no	no	no	yes	no	no	(−)	13
BLYP-D$_G$/LP	no	no	no	no	no	no	no	no		0
B2PLYP/TZVPP	no	no	no	no	no	no	no	no	(−)[c]	0
M06-2X/aDZ	no	no	no	no	no	no	no	no		0
PBE/LP	no	no	no	no	no	no	no	no	(−)	0
PBE-D$_J$/LP	no	no	no	no	no	no	no	no	(+)	0
TPSS-D$_G$/LP	no	no	no	no	no	no	no	no	(+)	0
TPSS-D$_J$/LP	no	no	no	no	no	no	no	no	(+)	0

[a] For a brief definition of requests RQ₁, RQ₃, ..., RQ₅, see the footnote of Table 9. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

According to the result shown in Table 11, it is clearly the dispersion-dominated complex category that dictates the

Table 12. DFT Methods: A Statistical Evaluation of the Mixed-Character Complexes of the S22 × 5 Set

method/basis	RQ₁[a]		RQ₃		RQ₄		RQ₅			score[b]
BLYP-D$_J$/LP	no	yes	no	yes	yes	yes	no	no	(−)[c]	50
B2PLYP-D/TZVPP	no	yes	no	yes	yes	yes	no	no		50
PBE-D$_J$/LP	no	yes	no	yes	yes	yes	no	no	(+)	50
BLYP-D$_G$/LP	no	yes	no	yes	no	yes	no	no		38
PBE-D$_G$/LP	no	yes	no	no	no	yes	no	no	(+)	25
TPSS-D$_J$/LP	no	yes	no	no	no	yes	no	no	(+)	25
B2PLYP/TZVPP	no	no	no	yes	no	no	no	no	(−)	13
TPSS-D$_G$/LP	no	no	no	no	no	yes	no	no	(+)	13
M06-2X/aDZ	no	no	no	no	no	no	no	no		0
PBE/LP	no	no	no	no	no	no	no	no	(−)	0

[a] For a brief definition of requests RQ₁, RQ₃, ..., RQ₅, see the footnote of Table 9. For a more detailed definition and reasoning, see section 2. [b] The score is defined as the number of positive (yes) matches of a particular (sub)requirement, divided by the total number of requirements applied (eight in this case). [c] (+) means overstabilization; (−) means understabilization. If no symbol is provided, the method could not be unambiguously assigned to any category.

overall substandard performance of the DFT methods. Only three DFT methods have a nonzero score. The most likely reason for the failure of, at least, the empirically dispersion-corrected DFT methods in delivering a balanced and accurate description of noncovalent complexes upon the interaction coordinate stretching originates from the fundamental limitation of the damping function to compensate equally well for the doublecounting of the dispersion energy at various geometry distortions.

The accuracy of description of the mixed-character complexes (see Table 12) is, again, significantly better than that of all the S22 × 5 and the dispersion-dominated complexes but is still clearly inferior to the accuracy of the WFT methods. However, none of the methods reached the score of 70%.

Table 13 presents, analogously to Table 7 for the WFT methods, the absolute and relative ME, MAE, and MAX values for all of the interaction types and individual geometry distortions. It is again clear that the WFT methods provide much better accuracy than the DFT methods. For instance, unlike the WFT methods, only one of the DFT methods

Table 13. DFT Methods[a]

method/basis	0.9			1.0			1.2			1.5			2.0		
TPSS-D$_G$/LP	1.78	1.78	4.2	1.04	1.04	2.21	0.43	0.43	1.18	0.28	0.28	0.75	0.08	0.08	0.26
	(>−100)[b]	(>100)	(>100)	(−21.49)	(21.49)	(40.84)	(−12.67)	(12.67)	(48.86)	(−18.69)	(18.69)	(58.33)	(−19.34)	(19.34)	(>100)
TPSS-D$_J$/LP	−0.09	0.63	2.67	0.19	0.32	1.17	0.25	0.27	0.68	0.22	0.22	0.53	0.08	0.08	0.26
	(−6.26)	(32.29)	(>100)	(−5.90)	(8.61)	(34.70)	(−10.69)	(10.98)	(54.61)	(−18.05)	(18.05)	(64.10)	(−18.93)	(18.93)	(>100)
M06-2X/aDZ	0.82	0.85	2.59	0.55	0.6	1.95	0.00	0.20	0.75	−0.09	0.17	0.56	−0.01	0.05	0.15
	(−80.58)	(80.73)	(>100)	(−15.86)	(16.18)	(61.03)	(−3.64)	(7.38)	(60.32)	(0.94)	(12.29)	(36.51)	(16.5)	(21.00)	(>100)
BLYP-D$_G$/LP	0.38	0.42	1.05	0.25	0.32	0.86	−0.03	0.20	0.53	0.00	0.13	0.51	−0.02	0.07	0.16
	(−31.91)	(33.31)	(>100)	(−2.84)	(7.17)	(31.96)	(7.12)	(8.81)	(50.18)	(6.33)	(8.93)	(31.92)	(−9.50)	(17.28)	(>100)
BLYP-D$_J$/LP	−0.25	0.50	1.54	−0.47	0.61	2.29	−0.37	0.48	1.57	−0.17	0.21	0.51	−0.07	0.09	0.27
	(44.74)	(53.94)	(>100)	(12.80)	(15.13)	(54.64)	(17.46)	(18.73)	(66.25)	(16.37)	(16.78)	(45.73)	(2.51)	(13.23)	(50.76)
PBE/LP	−2.92	2.99	13.45	−2.21	2.24	9.58	−1.13	1.15	4.21	−0.42	0.43	1.20	−0.14	0.14	0.40
	(>100)	(>100)	(>100)	(46.78)	(47.32)	(>100)	(31.04)	(31.48)	(81.64)	(27.38)	(27.82)	(>100)	(39.80)	(39.80)	(>100)
PBE-D$_G$/LP	1.23	1.23	2.66	0.81	0.81	1.97	0.38	0.39	1.28	0.21	0.21	0.90	0.03	0.04	0.25
	(−81.83)	(81.83)	(>100)	(−18.32)	(18.34)	(39.11)	(−10.55)	(10.89)	(38.21)	(−10.04)	(10.04)	(31.23)	(−3.42)	(4.54)	(20.38)
PBE-D$_J$/LP	−0.63	1.19	6.32	0.06	0.44	2.05	0.47	0.47	1.00	0.32	0.32	0.68	0.08	0.08	0.34
	(74.14)	(>100)	(>100)	(−5.30)	(10.16)	(66.90)	(−17.56)	(17.56)	(71.15)	(−21.10)	(21.10)	(65.25)	(−17.45)	(17.45)	(95.17)
B2PLYP/TZVPP	−2.31	2.36	9.24	−1.67	1.7	6.27	−0.88	0.92	2.92	−0.37	0.39	1.03	−0.09	0.1	0.26
	(>100)	(>100)	(>100)	(40.31)	(41.07)	(>100)	(30.24)	(31.12)	(95.18)	(31.83)	(33.01)	(>100)	(33.84)	(34.81)	(>100)
B2PLYP-D /TZVPP	0.73	0.73	1.67	0.55	0.56	1.32	0.22	0.31	0.94	0.09	0.17	0.66	0.03	0.05	0.22
	(−27.96)	(28.64)	(>100)	(−7.43)	(9.42)	(23.67)	(−0.26)	(8.11)	(35.32)	(4.39)	(10.00)	(40.06)	(1.87)	(5.92)	(18.08)

[a] The values in rows: in the first line of each row, ME, MAE, and MAX (in kcal/mol); in the second line, the relative quantities MRE, MARE, and MAXR (in %) for all of the interaction types and particular geometry distortions. The absolute and relative MAX(R) in a particular table cell do not necessarily correspond to the same complex. [b] The values above 100% (or below −100%) are for simplification shown as ">100". The actual values can be found in the Supporting Information.

Table 14. DFT Methods[a]

method/basis	$(CH_4)_2$	$(C_6H_6)_2$	Δ[b]
B2PLYP/TZVPP	95.78	94.55	1.23
M06-2X/aDZ	47.75	39.67	8.08
BLYP-D$_J$/LP	48.79	33.04	15.75
B2PLYP-D/TZVPP	26.39	10.47	15.91
TPSS-D$_G$/LP	18.33	34.91	−16.59
PBE-D$_G$/LP	29.77	9.5	20.27
BLYP-D$_G$/LP	39.67	18.33	21.33
TPSS-D$_J$/LP	51.37	29.81	21.56
PBE-D$_J$/LP	63.51	30.47	33.04
PBE/LP	63.69	105.92	−42.23

[a] The averaged absolute relative errors (AARE, in %) over all (except "0.9") geometry distortions for the methane and benzene dimers. [b] AARE(methane dimer)−AARE(benzene dimer).

Table 15. DFT Methods: A Comparison of the Overall Score of the Counterpoise-Corrected (CP) and Uncorrected (NCP) Calculations for the Whole S22 × 5 Set

method/basis	electrostatic		dispersion		mixed		all	
	CP	NCP	CP	NCP	CP	NCP	CP	NCP
BLYP-D$_G$/LP	63	38	0	0	25	38	0	0
BLYP-D$_J$/LP	75	50	25	25	50	50	10	10
B2PLYP/TZVPP	63	75	0	0	0	13	0	0
B2PLYP-D/TZVPP	88	63	25	0	63	50	20	10
M06-2X/aDZ	88	88	0	0	0	0	0	0
PBE/LP	50	50	0	0	0	0	0	0
PBE-D$_G$/LP	38	38	13	25	50	13	10	0
PBE-D$_J$/LP	50	25	25	0	50	38	10	0
TPSS-D$_G$/LP	63	50	13	0	63	13	10	0
TPSS-D$_J$/LP	88	88	13	0	88	25	10	0

(BLYP-D$_G$) had a MAX value below 1 kcal/mol (0.86 kcal/mol) at the equilibrium distance (1.0), and even in the 100% stretched geometry the mean absolute error is about 0.3 kcal/mol. On the other hand, the MAE error of, for example, TPSS-D$_J$ or B2PLYP-D at the equilibrium distance is only 0.32 and 0.56 kcal/mol, which again demonstrates the contrast between the conclusions one can draw from the statistics used in this work and a "simplified", "MAE in equilibrium geometry" assessment.

As mentioned in the WFT section, the accuracy for qualitatively different types of dispersion-dominated complexes cannot be extracted from the statistics used throughout this work. The performance of the DFT methods on the methane and benzene dimers is shown in Table 14. The most balanced treatment is, because of error cancellation, provided by the B2PLYP method. If only those methods that describe at least one of the complexes with an error of below 20% are considered, it is the B2PLYP-D method that delivers the most balanced results, Δ being about 16%.

The final note in this section is dedicated to the importance of the counterpoise corrections for the BSSE. For the WFT methods, it is generally accepted that at least in a small (or medium-sized) basis set, counterpoise correction for the BSSE leads to interaction energies closer to the CBS limit than the use of plain ones. Because of the inherent local nature of the electron-correlation description in the DFT methods, convergence toward the CBS is much faster as compared to the WFT methods. Despite the undoubted validity of this statement, it is often overrated and the DFT methods are frequently referred to as the "BSSE-free". Table 15 shows the score of the DFT methods in counterpoise-corrected (CP) and uncorrected (NCP) calculations for the whole of S22 × 5 and each interaction type separately. Please note, that BLYP and PBE are now calculated in smaller, TZVP basis set, as often recommend in the literature. It is important to say that for the relevant cases, that is, when the method can be used for the description of noncovalent interactions at all (according to our statistics, that is, the score being above 70%), the CP enhances the accuracy (in two cases the performance was unaffected, that is, M06-2X and TPSS-D$_J$, for electrostatics-dominated complexes, in one case it is even worse, that is, B2PLYP, again for the electrostatics-dominated complexes) when compared to the NCP results. Furthermore, if the CP is used with the TPSS-D$_J$ method in the calculation of mixed complexes, its score exceeds 70%, thus making it the only DFT method suitable for the

description of complexes in this category (again, according to our statistics). When dispersion corrected PBE and BLYP CP/NCP results in larger, for example LP, basis set are compared, less profound and less systematic trends are observed (see the Supporting Information). Benefits from using CP with larger basis sets in DFT methods with empirical dispersion correction are less clear because of the parametric nature of the correction.

4. Conclusions

A comparison of the performance of several WFT and DFT methods with respect to the CCSD(T)/CBS benchmark interaction energies not only in the equilibrium but also in stretched complex geometries (one shortened and three elongated in the direction of interaction coordinate) was carried out to investigate the general applicability of the approximate methods for large-scale systems. Spatially separated interacting molecular fragments, typically in non-equilibrium geometries, coexist in abundance with the "equilibrium" close contacts in biomolecules. With respect to this fact, any computational method potentially used for the calculation of large molecules must be capable of a balanced simultaneous description of noncovalent interactions in all of the purposeful conformations.

The statistical measures, proposed and used in this paper, were designed to assess the method's general accuracy, balanced treatment within the categories of interaction types (electrostatic- or dispersion-energy-dominated or the "mixed-character" complexes), as well as the balanced treatment of different complex-geometry distortions. The distinction between the interaction types is based on a DFT-SAPT analysis of the interaction energy, carried out for each complex and geometry distortion separately, thus being completely independent of the interaction type's definition on the basis of "chemical intuition".

Of the family of the WFT methods, the most potent approaches in this respect proved to be the MP2C, MP2.5, and SCS-CCSD. Of the DFT family, none of the methods tested in this work fulfilled the, rather strict, statistical criteria applied, thus being significantly inferior (at least) to the three WFT methods mentioned above. This does not necessary mean that the use of the DFT methods (obviously only those covering the dispersion energy) would be meaningless, but the possible satisfactorily accurate results obtained for the large molecules are quite likely to be obtained on the basis of error cancellation.

The importance of the counterpoise corrections in the DFT calculations was shown. For the relevant cases, that is, when the method is, according to our statistics, suitable for the description of noncovalent interactions, the counterpoise correction enhances both the balance and accuracy of the description. With increase of the basis set size, the role of the counterpoise correction becomes less important and furthermore, in the case of the DFT methods with empirical dispersion correction even less systematic.

Refined statistics on the dispersion-dominated complexes revealed that the balanced treatment of the aliphatic and $\pi-\pi$ stacked complexes is a challenging problem and that some methods capable of delivering excellent accuracy for one of these interaction types, for example, MP2 for aliphatic or SCS(MI)-MP2 for $\pi-\pi$ stacked complexes, may fail for the other interaction type.

The S22 × 5 geometries and interaction energies will be made available at www.begdb.com[39] at the time of the publication of this work.

Supporting Information Available: Complete statistics for the tested WFT and DFT methods (extended of those not shown in the paper), S22×5 CCSD(T)/CBS interaction energies and detailed information on the basis sets used. Geometries of all S22×5 complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(2) DiStasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073.

(3) Pitoňák, M.; Neogrady, P.; Černý, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282.

(4) Hesselmann, A. *J. Chem. Phys.* **2008**, *128*, 144112.

(5) Pitoňák, M.; Heβelmann, A. *J. Chem. Theory Comput.* **2010**, *6*, 168.

(6) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.

(7) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

(8) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(9) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.

(10) Řezáč, J.; Fanfrlík, J.; Salahub, D. R.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.

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

(12) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* [Online early access]. DOI: 10.1021/cr1000173. Published Online: May 20, 2010.

2376 *J. Chem. Theory Comput., Vol. 6, No. 8, 2010*

Gráfová et al.

(13) Pittner, J.; Hobza, P. *Chem. Phys. Lett.* **2004**, *390*, 496.

(14) Hopkin, B. W.; Tschumper, G. S. *J. Phys. Chem. A* **2004**, *108*, 2941.

(15) Crittenden, D. L. *J. Phys. Chem. A* **2009**, *113*, 1663.

(16) Pitoňák, M.; Neogrády, P.; Řezáč, J.; Jurečka, P.; Urban, M.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 1829.

(17) Pitoňák, M.; Riley, K. E.; Neogrády, P.; Hobza, P. *ChemPhysChem* **2008**, *9*, 1636.

(18) Pitoňák, M.; Janowski, T.; Neogrády, P.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1761.

(19) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2010**, *132*, 144104.

(20) Dunning, T. H. J. *J. Chem. Phys.* **1989**, *90*, 1007.

(21) Morgado, C. A.; Jurečka, P.; Svozil, D.; Hobza, P.; Sponer, J. *J. Chem. Theory Comput.* **2009**, *5*, 1524.

(22) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982.

(23) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.

(24) Rubeš, M.; Bludský, O. *ChemPhysChem* **2009**, *10*, 1868.

(25) Heßelmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

(26) Marchetti, O.; Werner, H. *J. Phys. Chem. A* **2009**, *113*, 11580.

(27) Molnar, L. F.; He, X.; Wang, B.; Merz, K. M. *J. Chem. Phys.* **2009**, *131*, 065102.

(28) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

(29) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

(30) Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(31) Aquilante, F.; Pedersen, T. B. *Chem. Phys. Lett.* **2007**, *449*, 354.

(32) The basis sets are available from the TURBOMOLE homepage via the FTP server button (in the subdirectories basen, jbasen, and cbasen). See http://www.turbomole.com (accessed Jun 15, 2010).

(33) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(34) *TURBOMOLE*, V5.9 2009; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH: Karlsruhe, Germany, 1989−2007; TURBOMOLE GmbH: Karlsruhe, Germany, 2007−2010; available from http://www.turbomole.com (accessed Jun 15, 2010).

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

(36) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Schütz, M.; Celani, P.; Korona, T.; Manby, F. R.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. MOLPRO, a package of ab initio programs, version 2006.1; Birmingham, UK, 2006; see http://www.molpro.net (accessed Jun 15, 2010).

(37) Aquilante, F.; Vico, L. D.; Ferre, N.; Malmqvist, P.-Å.; Neogrady, P.; Pedersen, T.; Pitoňák, M.; Reiner, M.; Roos, B. O.; Serrano-Andres, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224.

(38) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; DiLabio, G. A. *J. Chem. Phys.* **2009**, *131*, 03411.

(39) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261.

(40) Schaefer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(41) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.