ACCOUNTS
of chemical research

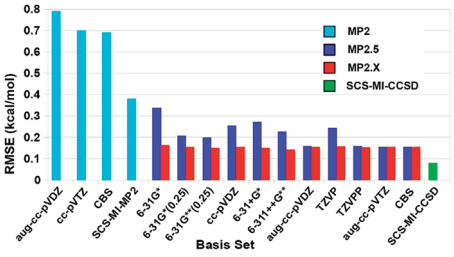
Calculations on Noncovalent Interactions and Databases of Benchmark Interaction Energies

PAVEL HOBZA

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 166 10 Prague, Czech Republic, Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacky University, 771 46 Olomouc, Czech Republic, and Department of Chemistry, Pohang University of Science and Technology, San 31, Hyojadong, Namgu, Pohang 790-784, Korea

RECEIVED ON OCTOBER 5, 2011

CONSPECTUS



Although covalent interactions determine the primary structure of a molecule, the noncovalent interactions are responsible for the tertiary and quaternary structure of a molecule and create the fascinating world of the 3D architectures of biomacromolecules. For example, the double helical structure of DNA is of fundamental importance for the function of DNA: it allows it to store and transfer genetic information. To fulfill this role, the structure is rigid to maintain the double helix with a proper positioning of the complementary base, and floppy to allow for its opening. Very strong covalent interactions cannot fulfill both of these criteria, but noncovalent interactions, which are about 2 orders of magnitude weaker, can. This Account highlights the recent advances in the field of the design of novel wave function theory (WFT) methods applicable to noncovalent complexes ranging in size from less than 100 atoms, for which highly accurate ab initio methods are available, up to extended ones (several thousands atoms), which are the domain of semiempirical QM (SQM) methods.

Accurate interaction energies for noncovalent complexes are generated by the coupled-cluster technique, taking single- and double-electron excitations iteratively and triple-electron excitation perturbatively with a complete basis set description (CCSD(T)/CBS). The procedure provides interaction energies with high accuracy (error less than 1 kcal/mol). Because the method is computationally demanding, its application is limited to complexes smaller than 30 atoms. But researchers would also like to use computational methods to determine these interaction energies accurately for larger biological and nanoscale structures. Standard QM methods such as MP2, MP3, CCSD, or DFT fail to describe various types of noncovalent systems (H-bonded, stacked, dispersion-controlled, etc.) with comparable accuracy. Therefore, novel methods are needed that have been parametrized toward noncovalent interactions, and existing benchmark data sets represent an important tool for the development of new methods providing reliable characteristics of noncovalent clusters.

Our laboratory developed the first suitable data set of CCSD(T)/CBS interaction energies and geometries of various noncovalent complexes, called S22. Since its publication in 2006, it has frequently been applied in parametrization and/or verification of various wave function and density functional techniques. During the intense use of this data set, several inconsistencies emerged, such as the insufficient accuracy of the CCSD(T) correction term or its unbalanced character, which has triggered the introduction of a new, broader, and more accurate data set called the S66 data set. It contains not only 66 CCSD(T)/CBS interaction energies determined in the equilibrium geometries but also 1056 interaction energies calculated at the same level for nonequilibrium geometries. The S22 and S66 data sets have been used for the verification of various WFT methods, and the lowest RMSE (S66, in kcal/mol) was found for the recently introduced SCS-MI-CCSD/CBS (0.08), MP2.5/CBS (0.16), MP2.X/6-31G* (0.27), and SCS-MI-MP2/CBS (0.38) methods. Because of their computational economy, the MP2.5 and MP2.X/6-31G* methods can be recommended for highly accurate calculations of large complexes with up to 100 atoms.

The evaluation of SQM methods was based only on the S22 data set, and because some of these methods have been parametrized toward the same data set, the respective results should be taken with caution. For really extended complexes such as protein–ligand systems, only the SMQ methods are applicable. After adding the corrections to the dispersion energy and H-bonding, several methods exhibit surprisingly low RMSE (even below 0.5 kcal/mol). Among the various SMQ methods, the PM6-DH2 can be recommended because of its computational efficiency and it can be used for optimization (which is not the case for other SQM methods). The PM6-DH2 is the base of our novel scoring function used in in silico drug design.

Published on the Web 01/06/2012 www.pubs.acs.org/accounts
10.1021/ar200255p © 2012 American Chemical Society

Introduction

While covalent interactions determine the primary structure of a molecule, the noncovalent interactions (NI) are responsible for the tertiary and quaternary structure of a molecule and thus create the fascinating world of the 3D architectures of biomacromolecules. The primary structure, determined by bond lengths and valence angles, is similar for all combinations of chemical elements. On the other hand, macromolecular architectures are characterized by their variety. The structure of biomacromolecules evidently represents a key property of a system, and to understand the function of biomacromolecules one should first understand their structures. The double helical structure of DNA is of fundamental importance for the function of DNA; it allows it to store and transfer genetic information. To fulfill this role, the structure should be on the one hand rigid (to keep the double helix with a proper positioning of the complementary bases) and on the other hand floppy to allow for its opening. Clearly, very strong covalent interactions cannot fulfill this, whereas NI that are about 2 orders of magnitude weaker can. (Interaction strength per atom pair in covalent bond is about 100 kcal/mol while that in noncovalent bond is about 1 kcal/mol, or less.)

Another important difference between covalent and NI concerns their origin. In the case of the former, the interaction origin is unique: it is an overlap between the orbitals of interacting subsystems. Consequently, all covalent bonds are governed by interatomic overlap and some variety comes from the different electronegativities of the interacting atoms. On the other hand, the wide variability of noncovalent structures is due to the variety of noncovalent contributions codetermining the total interaction. The electrostatic, induction, and dispersion energies have different origins and different dependences on the distance. The concerted action of all these interactions exhibiting different geometrical dependences is important for the structure and function of biomacromolecules. As an example, it is possible to mention the errorless closing of DNA double helical structure after its designed or spontaneous opening. Evidently, not despite but because of its weakness, NI play a key role in biodisciplines. We have focused on the biomolecular interactions, but the same is true for NI in nanostructures.

This Account highlights the recent advances in the field of the design of novel wave function theory (WFT) methods applicable to noncovalent complexes of different sizes, from small ones (up to 100 atoms), for which highly accurate ab initio methods are available, up to extended ones (several thousands atoms), which are the domain of semiempirical QM (SQM) methods.

Why Use Computational Techniques for Evaluating NI?

Different experimental techniques provide information on various properties, but no single technique yields a complete description of a system. The use of experimental values is also limited by the lack of a sufficiently large set of consistent data which are important for the parametrization and/or verification of novel computational techniques (see below). Our recent book[1] describes the theoretical and experimental investigations of NI, and the reader can find more details there on the advantages and disadvantages of both approaches. On the other hand, by solving the Schrödinger equation one obtains, besides basic information such as the structure and energy, any other property generated from the knowledge of the wave function. A theoretical treatment thus relatively easily provides a consistent description of the various types of noncovalent complexes which can form a benchmark database covering all of the important bonding motifs. It must, however, be mentioned that only the highest level of theoretical calculations yields satisfactory results; the lower-level techniques can provide misleading results and subsequently also misleading conclusions. Another important advantage of the theoretical approach is that it provides valuable information on the origin of stabilization. Understanding the nature of stabilization of, for example, a protein−ligand complex aids in the design of a more powerful ligand exhibiting larger binding free energy and consequently also greater pharmacological activity.

Which theoretical approach is the most suitable for the computation of NI? Extensive evidence has been collected in the past few years[2,3] showing that a coupled cluster (CC) treatment represents the method of choice. The significant feature of coupled-cluster theory (as against, e.g., DFT) is that it is systematically improvable upon the inclusion of a higher excitation operator, as shown below:

$$\mathrm{CCSD} < \mathrm{CCSD(T)} < \mathrm{CCSDT} < \mathrm{CCSDT(Q)} < \mathrm{CCSDTQ} < \mathrm{FCI}$$

The CCSD(T) method, taking single (S) and double (D) electron excitations iteratively and triple (T) electron excitation perturbatively with a complete basis set description (CCSD(T)/CBS), provides a highly accurate description of various types of noncovalent complexes.

TABLE 1. CBS Interaction Energies (in kcal/mol) for the S22 Data Set[a]

no. complex	Δ*E*(MP2)	Δ*E*(CCSD(T))	geometry	ref 5
hydrogen-bonded complexes (7)				
1 $(NH_3)_2$ (C_{2h})	−3.20 (QZ→5Z)	−3.17 (qz)	CCSD(T)/QZ	−3.17
2 $(H_2O)_2$ (C_S)	−5.03 (QZ→5Z)	−5.02 (qz)	CCSD(T)/QZ	−5.02
3 formic acid dimer (C_{2h})	−18.60 (QZ→5Z)	−18.61 (tz)	CCSD(T)/TZ	−18.80
5 uracil dimer (C_{2h})	−20.61 (TZ→QZ)	−20.65 (tz-fd)	MP2/TZ- CP	−20.69
6 2P···2AP (C_1)	−17.37 (TZ→QZ)	−16.71 (tz-fd)	MP2/TZ-CP	−17.00
7 A···TWC (C_1)	−16.54 (TZ→QZ)	−16.37 (dz)	MP2/TZ-CP	−16.74
complexes with a predominant dispersion contribution (8)				
8 $(CH_4)_2$ (D_{3d})	−0.51(QZ →5Z)	−0.53 (qz)	CCSD(T)/TZ	−0.53
9 $(C_2H_4)_2$ (D_{2d})	−1.62(QZ →5Z)	−1.51 (qz)	CCSD(T)/QZ	−1.50
10 benzene···CH_4 (C_3)	−1.86(QZ →5Z)	−1.50 (tz-fd)	MP2/TZ-CP	−1.45
11 benzene dimer (C_{2h})	−4.95 (aT→aQ)	−2.73 (adz)	MP2/TZ-CP	−2.62
12 pyrazine dimer (C_s)	−6.90 (aT→aQ)	−4.42 (tz-fd)	MP2/TZ-CP	−4.20
13 uracil dimer (C_2)	−11.39 (TZ→QZ)	−10.12 (tz-fd)	MP2/TZ-CP	−9.74
14 indole···benzene (C_1)	−8.12 (TZ→QZ)	−5.22 (dz)	MP2/TZ-CP	−4.59
15 A···T stack (C_1)	−14.93(TZ →QZ)	−12.23 (dz)	MP2/TZ-CP	−11.66
mixed complexes (7)				
16 ethene···ethine (C_{2v})	−1.69 (QZ→5Z)	−1.53 (tz)	CCSD(T)/QZ	−1.51
17 benzene···H_2O (C_S)	−3.61 (QZ→5Z)	−3.28 (tz-fd)	MP2/TZ-CP	−3.29
18 benzene···NH_3 (C_S)	−2.72 (QZ→5Z)	−2.35 (tz-fd)	MP2/TZ-CP	−2.32
19 benzene···hcn (C_S)	−5.16 (aT→aQ)	−4.46 (tz-fd)	MP2/TZ-CP	−4.55
20 benzene dimer (C_{2v})	−3.62 (aT→aQ)	−2.74 (adz)	MP2/TZ-CP	−2.71
21 indole···benzene (T) (C_1)	−7.03 (TZ→QZ)	−5.73 (dz)	MP2/TZ-CP	−5.62
22 phenol dimer (C_1)	−7.76 (TZ→QZ)	−7.05 (tz-fd)	MP2/TZ-CP	−7.09

[a]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. In the modified cc-pVTZ set (tz-fd), one set of f- and one set of d-functions have been removed (only the more diffuse d-function has been kept) and the hydrogen basis set has been modified analogically. The abbreviations 2P, 2AP, A, and T stand for 2-pyridoxine, 2-aminopyridine, adenine, and thymine, respectively. Last column contains recalculated CCSD(T) interaction energies (aT →aQ; aD →aT) from ref **5**.

Chemical (∼1 kcal/mol) or even subchemical (∼0.1 kcal/mol) accuracy can be obtained. (All of the characteristics of the various covalent bonds are accurately described at that level of theory; the description of covalent interactions represents a considerably easier task.) Another important feature of the method described is that it represents a genuine ab initio approach in the sense that no single parameter is adopted and all of the values needed are evaluated from the first principles. We should, however, pay a price for this. Despite the enormous progress of both hardware and software in the past decade, the use of the method is limited to systems with up to about 30 atoms. The treatment of larger systems with hundreds to thousands of atoms, that is, systems playing a role in bio- and nanostructures, requires the use of quite different techniques. What makes the situation difficult is the fact that the accuracy of these methods should be similar to that of accurate CCSD(T)/CBS calculations. Standard WFT and DFT methods cannot fulfill this requirement and the only chance is to use novel methods which are parametrized. Such a parametrization should be done toward a data set collecting the interaction energies and geometries of various types of noncovalent complexes evaluated at a sufficiently high level. The problem was that until very recently no such data set existed. A serious problem connected with the parametrization and/or scaling the interaction energy is connected with the lost of wave function information.

Data Sets of Benchmark Interaction Energies and Geometries

S22 Data Set. The first suitable data set, called S22, originated from our laboratory and appeared already in 2006.[4] Table 1 shows the CCSD(T)/CBS interaction energies and geometries of the complexes studied. The success of the S22 data set stems from the fact that it contains 22 carefully selected complexes (for the structures, see Figure 1), 7 H-bonded, 8 dispersion-controlled, and 7 mixed ones; further small, medium, as well as large complexes are covered. It is equally important that the S22 set spans a wide range of interaction strengths in order to represent a

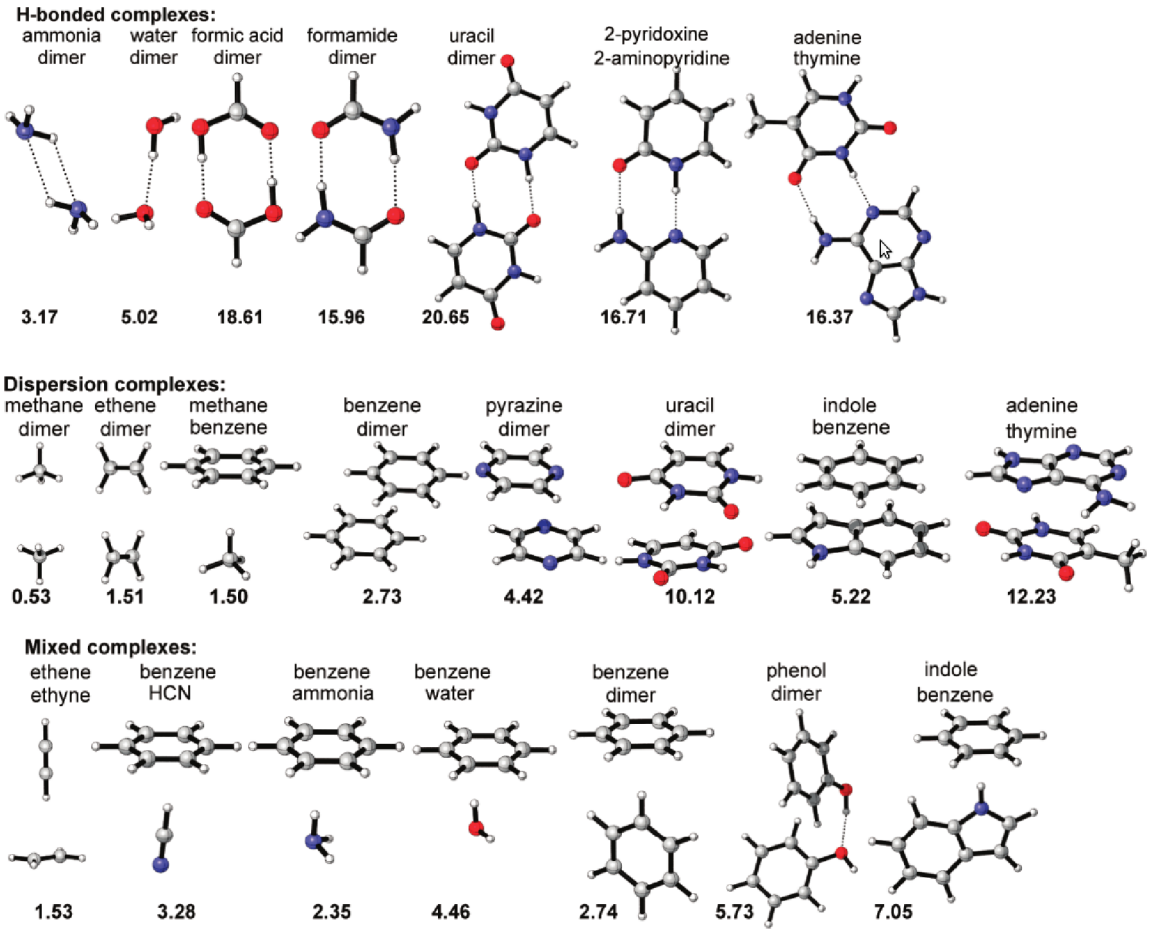


FIGURE 1. Optimized structures and stabilization energies (in kcal/mol) of 22 complexes from the S22.

variety of NI. The benchmark CCSD(T)/CBS interaction energies were determined as follows:

$$\Delta E^{\mathrm{CCSD(T)}}{}_{\mathrm{CBS}} = \Delta E^{\mathrm{MP2}}{}_{\mathrm{CBS}} + \Delta\mathrm{CCSD(T)} \qquad (1)$$

While the first term was extrapolated using the aug-cc-pVDZ → aug-cc-pVTZ and also higher (and more accurate) schemes, the determination of the ΔCCSD(T) term (defined as ($\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}}$)) was not unique. Because of its smaller dependence on the basis-set size (unlike the MP2 term), it could be determined directly with a relatively small basis set. For the largest complexes of the S22 set, their values were calculated with the DZ (or similar) basis set. For the smaller complexes, we adopted a more accurate procedure based on extrapolation. The unequal evaluation of the ΔCCSD(T) term represents the most serious problem of the data set. Another problem concerns the fact that the database was not well balanced and the aromatic interactions clearly dominated over the aliphatic ones. Further, the equilibrium geometries were determined mostly "only" at the MP2/cc-pVTZ counterpoise-corrected level. The final potential problem concerned the fact that all of the entries in the S22 refer to equilibrium distances while the whole dissociation curves might provide a better base for verification and/or parametrization of the new computational techniques. Consequently, several authors have attempted to improve the S22 data set, first by recalculating the interaction energies at a higher computational level.[5,6] The recalculated interaction energies, based on the basis-set extrapolated coupled-cluster values, differ only slightly from the original values (a mean absolute relative error of ~2%), but the error for some complexes was higher (see Table 1). And second, they tried by going beyond equilibrium geometries.[7,8]

S66 Data Set. The above-mentioned points triggered the preparation of a data set of a new generation which removes all of the problems and inaccuracies discussed and also allows for further systematic extension. In 2011, we

TABLE 2. CCSD(T)/CBS Interaction Energies (in kcal/mol) for the S66 Data Set[a]

HB	$-\Delta E$	dispersion	$-\Delta E$	others	$-\Delta E$
water···water	4.92/5.01	Ben···Ben ($\pi\cdots\pi$)	2.82/2.72	Ben···Ben (T)	2.88/2.83
water···MeOH	5.59/5.70	Pyr···Pyr ($\pi\cdots\pi$)	3.90/3.80	Pyr···Pyr (T)	3.54/3.51
water···MeNH2	6.91/7.04	Ur···Ur($\pi\cdots\pi$)	9.83/9.75	Ben···Pyr (T)	3.33/3.29
water···peptide	8.10/8.22	Ben···Pyr ($\pi\cdots\pi$)	3.44/3.34	Ben···ethyne (CH$\cdots\pi$)	2.87/2.86
MeOH···MeOH	5.76/5.85	Ben···Ur ($\pi\cdots\pi$)	5.71/5.59	ethyne···ethyne (T)	1.52/1.54
MeOH···MeNH2	7.55/7.67	Pyr···Ur ($\pi\cdots\pi$)	6.82/6.70	Ben···AcOH (OH$\cdots\pi$)	4.71/4.73
MeOH···peptide	8.23/8.34	Ben···ethene	1.43/1.36	Ben···AcNH2 (NH$\cdots\pi$)	4.36/4.40
MeOH···water	5.01/5.09	Ur···ethene	3.38/3.33	Ben···water (OH$\cdots\pi$)	3.28/3.29
MeNH2···MeOH	3.06/3.11	Ur···ethyne	3.74/3.69	Ben···MeOH (OH$\cdots\pi$)	4.19/4.17
MeNH2···MeNH2	4.16/4.22	Pyr···ethene	1.87/1.80	Ben···MeNH2 (NH$\cdots\pi$)	3.23/3.20
MeNH2···peptide	5.42/5.48	pentane···pentane	3.78/3.76	Ben···peptide (NH$\cdots\pi$)	5.28/5.26
MeNH2···water	7.27/7.40	Neopen···pentane	2.61/2.60	Pyr···Pyr (CH···N)	4.15/4.24
peptide···MeOH	6.19/6.28	Neopen···Neopen	1.78/1.76	ethyne···water (CH···O)	2.85/2.93
peptide···MeNH2	7.45/7.56	Cyclopen···Neopen	2.40/2.40	ethyne···AcOH (OH$\cdots\pi$)	4.87/4.97
peptide···peptide	8.63/8.72	Cyclopen···Cyclopen	3.00/2.99	pentane···AcOH	2.91/2.91
peptide···water	5.12/5.20	Ben···Cyclopen	3.58/3.51	pentane···AcNH2	3.53/3.53
uracil···uracil	17.18/17.45	Ben···Neopen	2.90/2.85	Ben···AcOH	3.80/3.75
water···pyridine	6.86/6.97	Ur···pentane	4.85/4.81	peptide···ethene	3.00/3.00
MeOH···pyridine	7.41/7.51	Ur···Cyclopen	4.14/4.09	Pyr···ethyne	3.99/4.10
AcOH···AcOH	19.09/19.41	Ur···Neopen	3.71/3.69	MeNH2···Pyr	3.97/3.97
AcNH2···AcNH2	16.26/16.52	ethene···pentane	2.01/1.99		
AcOH···uracil	19.49/19.78	ethyne···pentane	1.75/1.75		
AcNH2···uracil	19.19/19.47	peptide···pentane	4.26/4.26		

[a]In all cases, MP2 correlation energy was extrapolated from aTZ and aTQ basis sets while the upper and lower values differ by the ΔCCSD(T) correction determined at the aDZ (upper) or extrapolated from the haDZ and haTZ basis sets (lower). The abbreviations Ben, Pyr, Ur, Neopen, and Cyclopen stand for benzene, pyridine, uracil, neopentane, and cyclopentane, respectively.

published a better balanced data set named S66,[9] which contains 66 complexes formed by combining 14 monomers in various configurations. The following monomers, representing important motifs in NI, were considered: acetic acid, acetamide, benzene, cyclopentane, ethane, ethyne, neopetane, n-pentane, methylamine, methanol, *N*-methylacetamide, pyridine, uracil, and water. Evidently, only first-row elements and uncharged subsystems were considered. The straightforward construction of the data set, however, allows its easy extension.

The CCSD(T)/CBS interaction energies were again determined using eq 1, where the MP2 term was extrapolated from the aug-cc-pVTZ and aug-cc-pVQZ basis sets and the ΔCCSD(T) correction was calculated with the aug-cc-pVDZ basis set. All of the interaction energies were corrected for the basis-set superposition error using the counterpoise scheme. The complex geometry was determined at the counterpoise-corrected MP2/cc-pVTZ level. Table 2 presents the CCSD(T)/CBS interaction energies for 66 complexes of the S66 data set (23 H-bonded, 23 dispersion-controlled and 20 others) in the equilibrium geometries, and the respective structures can be found in the original paper.

By performing higher-level calculations, it is possible to estimate an error at the current computational level. The computational scheme (eq 1) can be improved in different ways, but the most significant is the estimation of the CCSD(T)/CBS interaction energy from direct extrapolations of the CCSD(T) energies. For the 10 smallest complexes, the CCSD(T)/CBS interaction energies were obtained by direct extrapolation of the CCSD(T) energies from the aug-cc-pVTZ and aug-cc-pVQZ basis sets. For the selected test set, the original S66 method gives an average error value of 1.2% with the largest error being 2.5%, and for the whole S66 set we expect errors below 3%. The small test set also allows the investigation of different schemes for the determination of the ΔCCSD(T) term, which is critical for the overall accuracy. Passing from the original aug-cc-pVDZ to the larger aug-cc-pVTZ basis set, the RMSE is reduced from 0.080 to 0.020 kcal/mol, and even better results (0.009 kcal/mol) were obtained when this term was extrapolated from the aug-cc-pVDZ/cc-pVDZ and aug-cc-pVTZ/cc-pVTZ basis sets (the upper and lower basis sets refer to the non-hydrogen and hydrogen atoms, respectively, and in Table 2 we used abbreviations haDZ and haTZ, respectively). When this more advanced level was used for the whole S66 set (see Table 2), negligible errors (an average unsigned error of 0.08 kcal/mol (1.5%) and RMSE of 0.10 kcal/mol) were obtained.[10] This is a clear message that the theoretical level used for the original S66 data set is accurate enough and could be safely used for the future extension of the data set. The categorical imperative is to use exactly the same protocol as was adopted for the original S66. We are aware of the fact that although the S66 is large, it did not cover all of the important noncovalent motifs and its subsequent extension is crucial. The possible extension concerns, for example, complexes

TABLE 3. RMSE (in kcal/mol) of Interaction Energies Determined by the Selected Method with Respect to the Benchmark Interaction Energies on the Original S22 and S66 Data Sets (in both cases the interaction energies were systematically corrected for the basis set superposition error); S66 Values Are in Parentheses

method	RMSE	method	RMSE
MP2	0.94 (0.69)	CCSD	0.42 (0.70)
SCS-MP2	0.58 (0.87)	SCS-CCSD	0.29 (0.25)
SCS-MI-MP2	0.26 (0.38)	SCS-MI-CCSD	0.17 (0.08)
MP2.5	0.22 (0.16)	MP2C	(0.71)
MP3	0.67 (0.62)		

containing halogen atoms, atoms of the second period, or charged systems.

The discussion in the previous paragraph was limited to the evaluation of the accuracy of the CCSD(T)/CBS technique. But what is the role of other limitations, especially of higher electron excitations? Passing to more accurate CCSDT (triple excitations are now solved iteratively like the single and double excitations) and CCSD(TQ) (besides triple excitations also quadruple excitations are considered) methods has been connected with only a marginal improvement (below 0.1 kcal/mol) in the stabilization energy.[11,12] The CCSD(T) thus stands as the "golden-standard" method because of its outstanding accuracy for the computational cost ratio.

Besides 66 equilibrium geometries, also 528 nonequilibrium geometries covering the stretching variations along the dissociation curves were included where interaction energies were determined at the same level.[9] For each dissociation curve, a new minimum was extrapolated at the CCSD(T)/CBS level. It means that besides counterpoise-corrected MP2/cc-pVTZ equilibrium geometries also higher-level CCSD(T)/CBS equilibrium geometries were generated. Finally,[10] we added another 526 nonequilibrium geometries covering the angular variations. The complete S66 data set contains 1122 benchmark interaction energies and thus represents the ideal tool for the testing and/or parametrization of novel computational techniques.

Performance of Selected WFT Methods Evaluated on the Basis of the S22 and S66 Data Sets

The S22 data set was intensely used for the parametrization and testing of new computational techniques and Table 3 shows the RMSE for the various methods. Several methods shown in Table 3 were, however, parametrized toward the S22 data set, which calls the results presented into question. The only correct way is to use different data sets for the parametrization and validation, and this was strictly followed in the case of S66 data set. We will limit ourselves to WFT methods only since a detailed study on the performance of DFT methods is under preparation. The respective RMSE for selected methods are collected in Table 3 and are given in parentheses (for details of obtaining the CBS limit see Table 1 and paragraph "S66 data set"). Evidently, the RMSE based on the S22 and S66 data sets are similar, and we will briefly discuss only the more reliable values based on the S66 data set.

The error produced by the MP2 method is large, and when analyzing errors in single subgroups we find that it is mainly due to dispersion-controlled complexes (29%), where the stabilization energy is strongly overestimated. H-bonded complexes are described considerably better, and the error is negligible (2%). When passing to smaller basis sets, we surprisingly found that the RMSE (in kcal/mol) remains comparable. The most balanced description (comparable error for all three subclasses) was obtained with the cc-pVTZ basis (RMSE = 0.70) set. It should be mentioned that this level was recommended earlier for the determination of complex geometries.[13] The MP2/cc-pVTZ level thus provides relatively sound interaction energies and geometries for various types of noncovalent complexes. The error is rather large (see Table 3), but the method does not contain any parameter and could be used for extended complexes.

Is there any explanation for the failure of the MP2 method? As suggested in ref 14, it is caused by the fact that the MP2 interaction energies contain only the uncoupled Hartree–Fock dispersion energy, which is strongly overestimated. When this "incorrect" dispersion energy is replaced by the "correct" one (obtained from, e.g., the TD-DFT calculations), the problem can be solved (MP2C method).[14] The RMSE of the MP2C is, however, larger than that of MP2 and arises from the unbalanced description of the H-bonded complexes on the one hand and dispersion-controlled and other complexes on the other hand, where the method failed. This is a surprising result, because it was expected that this method could be an alternative to the MP2.

Another possibility how to solve the problems with the MP2 method is to use spin-component scaled MP2 (SCS-MP2) method[15] based on a separate scaling of the same and opposite components of the correlation energy. As can be seen from Table 3, the overall performance of the method is worse than that of the MP2. The method removes the strong overestimation of the stacking energies but unfortunately also removes the strong point of the MP2 method, its excellent performance for H-bonding. Consequently, there is no advantage to preferring the SCS-MP2 method over the plain MP2 one. There is an explanation of

the "failure" of this method and it is the way of parametrization; it was parametrized toward reaction rates. When parametrization was based on NI (S22 data set), the performance of SCS-MI-MP2 method[16] was considerably improved (RMSE = 0.38 kcal/mol) and the method is superior among all of the MP2-based methods.

The overall accuracy and balance of the MP3/CBS is only slightly better than those of the MP2/CBS but still worse than those of the cheaper SCS-MI-MP2 method. The MP2 and MP3 describe different types of noncovalent complexes differently. While H-bonding energies are described comparably well by both methods, the $\pi-\pi$ stacking energies are overestimated by the MP2 and underestimated by the MP3. Interestingly, both errors were comparable in magnitude. It was thus natural to introduce a new method, called MP2.5,[17] which corrects the CBS MP2 interaction energy using a scaled (with the scaling factor being 0.5) third-order MP3 correlation correction term. Table 3 shows a considerably better performance than the MP2 and MP3 methods as well as all of the MP2-based methods including the SCS-MI-MP2. The MP2.5 also provides a balanced description of all three subclasses with slightly larger errors for dispersion-bound complexes. The method is more computationally intensive than any of the MP2-based methods, but the difference is not dramatic.

The error of the CCSD method is comparable to that of the MP2 method. This is a surprising finding in light of the fact that the CCSD covers a considerably larger portion of the correlation energy than the MP2 method and is also much more CPU intensive. Considerable improvement is, however, achieved when passing to the SCS-CCSD method.[18] The most accurate method from Table 3 is the SCS-MI-CCSD/CBS method[19] with negligible RMSE for the whole set as well as for the subsets. The method was parametrized toward NI, specifically the S22 data set. The computation time for all of the CCSD based methods is the same and is by more than one order magnitude greater than that for the MP2.5.

The previous discussion concerns only the equilibrium geometries for the 66 complexes of the S66 data set. Practically the same picture about the relative performance of all the methods discussed was obtained when besides the equilibrium geometries also 528 nonequilibrium geometries (forming the S66 $\times$ 8 set) were considered. When investigating the relative errors, we found that they are systematically larger (by 100−200%) for distances shorter than equilibrium. We pointed out this problem already in our previous study,[8] where we further showed that it is critical especially for all of the DFT-based methods. In order to explore the whole potential energy surface fully, we finally added also the angular dependences obtained by the rotation of the monomer in the complex. The resulting S66a8 set contains an additional 528 points for which interaction energies were determined consistently with the S66 and S66 $\times$ 8 sets. Also in this case, similar performances of all the methods investigated were found. Both of the subsets mentioned above can play an important role in the testing and/or parametrization of lower-level methods, for which stretch as well as angular variations might not be so well reproduced.

An analysis of the S66 database shows that the smallest errors are provided by the SCS-MI-CCSD and MP2.5 methods, which yield highly accurate interaction energies for various types of noncovalent complexes. The latter method contains one parameter (0.50, see later), while the former one contains two parameters fitted on the basis of the S22 interaction energies. The SCS-MI-CCSD method is faster than the CCSD(T) since the evaluation of triple excitation is time-consuming. For larger systems, it is even a CPU-determining step. The SCS-MI-CCSD method thus allows the study of larger complexes, but this "enlargement" is only modest (by no more than 10−15 atoms). Significant enlargement can be expected from the application of the second method, the MP2.5, and especially of the MP2.X methods (see below).

MP2.5 and Generalized MP2.X Methods. It was shown above that the MP2.5 method provides very low errors for the S22 and S66 data sets. The interaction energy is constructed like in eq 1, but instead of ΔCCSD(T) correction term the half of ΔMP3 correction term (defined as $(\Delta E^{MP3} - \Delta E^{MP2})$) was used. When the CBS calculations were performed for the MP2 and MP3 calculations, the MP2.5 is exactly the arithmetic mean of both interaction energies. The computation of the MP3/CBS interaction energy is, however, relatively expensive and can become cheaper when a medium basis set (typically aug-cc-pVDZ) is used. This procedure yields accurate results owing to the fact that the ΔMP3 correction term is less basis-set dependent than the MP3 and MP2 interaction energies themselves. Unfortunately, even the MP3/aug-cc-pVDZ calculation is tedious and impractical for extended complexes ($\sim$100 atoms). To make these calculations feasible for extended complexes, it is necessary to pass to smaller basis sets. Recently,[20] the performance of the MP2.5 along with several basis sets including a small one like 6-31G* has been assessed on the basis of the S66 database. Figure 2 shows the S66 RMSE for various basis sets. For the sake of comparison, Figure 2 also shows the performance of the MP2, SCS-MI-MP2, and SCS-MI-CCSD methods. It should be kept in mind that the last two

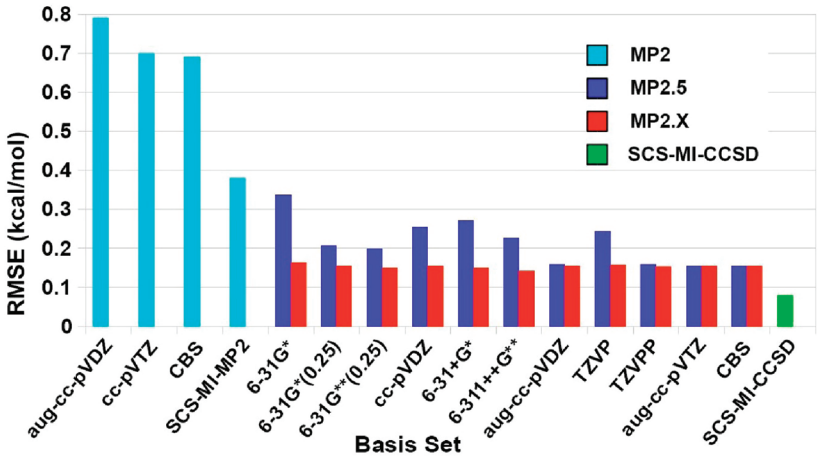


FIGURE 2. MP2, SCS-MI-MP2, SCS-MI-CCSD, MP2.5, and MP2.X percent RMSE (in kcal/mol) calculated with different basis sets for the S66 data set. 6-31G**(0.25) refers to 6-31G**(0.25, 0.15).

methods were parametrized toward NI using the S22 data set and the latter method provides the best results of all the methods tested. The MP2.5 is evidently very robust and depends only little on the basis set; the RMSE for the various basis sets lies in a rather narrow interval (0.34−0.15 kcal/mol). Surprisingly, all of the MP2.5 interaction energies possess smaller RMSE than all of the MP2 and even the SCS-MI-MP2 ones. Very favorable results have been obtained for the small 6-31G*(0.25) basis set used extensively in our laboratory in the 1990s[21] for treating H-bonded and stacked pairs of DNA bases. The RMSE at this basis set (0.21 kcal/mol) is only "slightly" larger than that for the CBS limit (0.15 kcal/mol). All of these results concern the MP2.5 method, that is, method where half of the ΔMP3 correction term is added to the CBS MP2 interaction energy. Further improvement can be achieved by performing a scaling of the ΔMP3 term. It means that instead of using the scaling factor of 0.5 (which gives the name to MP2.5 method and which controls the contribution of the MP3-based correction to CBS MP2 interaction energy) its value for each basis set was determined by parametrization. This is based on the fact that errors exhibited by the MP2 and MP3 methods are not exactly the same when the size of the basis set decreases. The scaling of the ΔMP3 term was based on the S66 data set and the respective values for basis sets presented in Figure 2 are the following: 0.86, 0.62, 0.62, 0.72, 0.73, 0.65, 0.52, 0.67, 0.53, 0.50, and 0.50. We intentionally used here the S66 data set for parametrization, since it is more general and larger than any other data set (including the S22 set). When passing to extended basis sets, the scaling factor approaches 0.5 (original value) and for the aug-cc-pVTZ basis set and the CBS limit it is exactly equal to 0.5. Figure 2 further shows that all of the MP2.X interaction energies provide very similar RMSE (red bars). The fact that MP2.X interaction energies can be made nearly insensitive toward the basis sets by optimization of just one "mixing" parameter X is very surprising and we do not have yet any theoretical explanation for it. What is only evident that a huge error cancellation should take place. Evidently, upon optimizing of only one "mixing" parameter (which is then valid for different complexes), the RMSE is significantly reduced. In the case of the smallest basis set, 6-31G*, the MP2.5 RMSE (0.34 kcal/mol) is reduced to less than one-half (0.16). The CPU time for MP2.5 and MP2.X is evidently the same, but passing from the nonoptimized scaling factor of 0.5 to optimized one of 0.68 yields considerable improvement. It is extremely significant that the RMSE produced by the MP2.X with smaller basis sets are no larger than those produced by the MP2.5 with larger basis sets and even at the CBS limit. The reason is clear: the CPU time differs dramatically. As an example, we present the relative timing of MP3 calculations for the benzene dimer with all basis sets shown in Figure 2 (except the CBS): 1.0: 1.1: 2.1: 2.0: 2.4: 4.7: 14.7: 23.3: 46.8: 285.4. This means that when performing the MP2.X with 6-31G* or 6-31G*(0.25) basis sets we obtain almost identical interaction energy as with the MP2.5/aug-cc-pVTZ method but for a fragment (1/285.4) of the CPU time. (In both cases, the CBS MP2 interaction energy should be calculated but then the ΔMP3 correction term will be calculated in the former case with 6-31G* or 6-31G*(0.25) basis sets and in the latter case with aug-cc-pVTZ basis set.) There is no doubt that the MP2.5 and MP2.X methods yield highly accurate interaction energies and can be used for extended complexes (with 100 or more atoms).

SQM Methods and Their Performance Evaluated on the Basis of the S22 Data Set

All of the SQM methods, including the recently introduced PM6[22] and OM3[23] ones and also the slightly different tight-binding density fitted (DFTB)[24] method, lack the ability to describe NI, specifically dispersion energy and, surprisingly, also hydrogen bonding. This is certainly a very demanding task, and only the most accurate ab initio QM procedures satisfactorily describe all these interactions (see above). The use of MM methods in the field of NI is limited, since they do not describe quantum effects such as proton and electron transfer, halogen bonding, and so forth. To improve the performance of the SQM methods, two modifications have recently been introduced: (i) an empirical dispersion energy term and (ii) an additional electrostatic term, improving the description of hydrogen-bonded complexes. This correction is directional and improves the description of H-bonds, which represents a weak point of all the standard SQM methods. The resulting corrections, called DH2,[25,26] can be added to any SQM method and considerably improve their performance toward NI.

The RMSE (in kcal/mol) for the S22 data set for the AM1, OM3, PM6, and DFTB methods is very large (5.50, 2.27, 2.51, and 2.15, respectively) and clearly prevents the use of these methods in the realm of NI. The inclusion of corrections to dispersion energy and H-bonding led to a significant improvement, and the RMSE (0.87, 0.84, 0.53, and 1.11, respectively) are now comparable[2] to those obtained by much more expensive methods, for example, the MP2/CBS method. Of the methods investigated,[2] the PM6-DH2 represents the most robust method, and because it uses a linear scaling variant of the SCF procedure, it can be applied to complexes with several thousands of atoms. In our laboratory, we are using this technique in in silico drug design.[27−29]

Conclusions

The introduction of efficient databases of benchmark interaction energies and geometries of noncovalent complexes triggered the development of novel computation methods allowing the treatment of large and even extended noncovalent complexes with up to recently unprecedentedly high accuracy. The S22 and recently introduced S66 data sets provide benchmark CCSD(T)/CBS interaction energies and geometries. The latter data set contains not only 66 interaction energies evaluated at their equilibrium geometries but also 1056 interaction energies determined at nonequilibrium geometries. The full S66 data set thus represents an ideal tool for the parametrization and/or verification of novel computation techniques tailored for NI, and a detailed analysis was made for various WFT methods. A negligible error lower than 0.2 kcal/mol was exhibited by the SCS-MI-CCSD and MP2.5 methods. The relative timing for a medium-sized noncovalent complex for CCSD(T)/CBS, SCS-MI-CCSD, and MP2.5 methods is roughly 2000:20:1. Passing to larger complexes, the ratio changes even further in favor of the last method. The MP2.5 method (and its considerably faster variant MP2.X method) thus represents the method of choice providing highly accurate interaction energies for complexes with up to 200 atoms. A lower accuracy is yielded by the SCS-MI-MP2 method applicable to considerably larger complexes.

SQM methods corrected for dispersion energy and H-bonding provide accurate interaction energies with average errors comparable to those obtained by much more expensive methods, such as the MP2/CBS method. The SQM methods allow the treatment of extended noncovalent complexes with up to several thousand atoms, and for general use the PM6-DH2 method can be recommended. The use of MM methods is limited by problems with the description of quantum effects.

There are two main massages of this Account. First, the high accuracy needed for treating NI playing a key role in bio- and nanostructures could only be reached by QM methods carefully parametrized toward suitable databases. Second, benchmark interaction energies and geometries of the complexes in S22 and S66 date sets are available through the BEGDB Web site (www.begdb.com) for download and interactive browsing.[30]

This work was a part of research Project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, ASCR and was supported by the Operational Program Research and Development for Innovations – European Regional Development Fund (Project CZ.1.05/2.1.00/03.0058 of the MEYS of the CR). The support of Praemium Academiae, ASCR, awarded to P.H. in 2007 is acknowledged. This work was supported by the Czech Science Foundation (P208/12/G016) and Korea Science and Engineering Foundation (World Class University Program: R32-2008-000-10180-0). The author thanks all four referees for their comments which improved the quality of the manuscript considerably.

BIOGRAPHICAL INFORMATION

Pavel Hobza was born in Prerov, Czech Republic in 1946. He is currently Distinguished Chair at the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic,

Prague. His research focuses on noncovalent interactions and their applications to biodisciplines.

REFERENCES

1 Hobza, P.; Müller-Dethlefs, K. *Noncovalent Interactions. Theory and Experiment*; The Royal Society of Chemistry: Cambridge, 2010.

2 Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. *Chem. Rev.* **2010**, *110*, 5023.

3 Riley, K. E.; Hobza, P. Noncovalent interactions in biochemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 3–17.

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

5 Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. Basis set consistent revision of the S22 test set of noncovalent interaction energies. *J. Chem. Phys.* **2010**, *132*, 144104.

6 Podeszwa, R.; Patkowski, K.; Szalewicz, K. Improved interaction energy benchmarks for dimers of biological relevance. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974–5979.

7 Molnar, L. F.; He, X.; Wang, B.; Merz, K. M., Jr. Further analysis and comparative study of intermolecular interactions using dimers from the S22 database. *J. Chem. Phys.* **2009**, *131*, 065102.

8 Grafova, L.; Pitonak, M.; Rezac, J.; Hobza, P. Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

9 Rezac, J.; Riley, K. E.; Hobza, P. S66: A well balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

10 Rezac, J.; Riley, K. E.; Hobza, P. Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.

11 Pittner, J.; Hobza, P. CCSDT and CCSD(T) calculations on model H-bonded and stacked complexes. *Chem. Phys. Lett.* **2004**, *390*, 496–499.

12 Pitonak, M.; Neogrady, P.; Rezac, J.; Jurecka, P.; Urban, M.; Hobza, P. Benzene dimer: high-level wave function and density functional theory calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1829–1834.

13 Dabkowska, I.; Jurecka, P.; Hobza, P. On geometries of stacked and H-bonded nucleic acid base pairs determined at various DFT, MP2, and CCSD(T) levels up to the CCSD(T)/complete basis set limit level. *J. Chem. Phys.* **2005**, *122*, 204322.

14 Hesselmann, A. Improved supermolecular second order Moller-Plesset intermolecular interaction energies using time-dependent density functional response theory. *J. Chem. Phys.* **2008**, *128*, 144112.

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

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

17 Pitonak, M.; Neogrady, P.; Cerny, J.; Grimme, S.; Hobza, P. Scaled MP3 non-covalent interaction energies agree closely with accurate CCSD(T) benchmark data. *ChemPhysChem* **2009**, *10*, 282–289.

18 Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. Improvement of the coupled-cluster singles and doubles method via scaling same- and opposite-spin components of the double excitation correlation energy. *J. Chem. Phys.* **2008**, *128*, 124111.

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

20 Riley, K. E.; Rezac, J.; Hobza, P. MP2.X: A generalized MP2.5 method that produces improved binding energies with smaller basis sets. *Phys. Chem. Chem. Phys.*, in press.

21 Hobza, P.; Sponer, J. Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations. *Chem. Rev.* **1999**, *99*, 3247–3276.

22 Stewart, J. J. P. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

23 Möhle, K.; Hofmann, H.-J.; Thiel, W. Description of peptide and protein secondary structures employing semiempirical methods. *J. Comput. Chem.* **2001**, *22*, 509–520.

24 Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaziras, E. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. *J. Chem. Phys.* **2001**, *114*, 5149.

25 Rezac, J.; Fanfrlik, J.; Salahub, D. R.; Hobza, P. Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

26 Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. A transferable H-bonding correction for semiempirical quantum-chemical methods. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.

27 Dobes, P.; Rezac, J.; Fanfrlik, J.; Hobza, P. Semiempirical quantum mechanical method PM6-DH2X describes the geometry and energetics of CK2-Inhibitor complexes involving halogen bonds well, while the empirical potential fails. *J. Phys. Chem. B* **2011**, *115*, 8581–8589.

28 Dobes, P.; Rezac, J.; Fanfrlik, J.; Otyepka, M.; Hobza, P. Transferable scoring function based on semiempirical quantum mechanical PM6-DH2 method: CDK2 with 15 structurally diverse inhibitors. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 223–235.

29 Fanfrlik, J.; Bronowska, A. K.; Rezac, J.; Prenosil, O.; Konvalinka, J.; Hobza, P. A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.

30 Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Rezac, T.; Pitonak, M.; Vondrasek, J.; Hobza, P. Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (www.begdb.com): A Users Manual and Examples. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.