JCTC Journal of Chemical Theory and Computation

A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods

Martin Korth,[†] Michal Pitoňák,[†,‡] Jan Řezáč,[†] and Pavel Hobza*[,†,§]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems, 16610 Prague 6, Czech Republic, Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, 84215 Bratislava 4, Slovak Republic, and Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

Received October 13, 2009

Abstract: Semiempirical methods could offer a feasible compromise between ab initio and empirical approaches for the calculation of large molecules with biological relevance. A key problem for attempts in this direction is the rather bad performance of current semiempirical methods for noncovalent interactions, especially hydrogen-bonding. On the basis of the recently introduced PM6-DH method, which includes empirical corrections for dispersion (D) and hydrogen-bond (H) interactions, we have developed an improved and transferable H-bonding correction for semiempirical quantum chemical methods. The performance of the improved correction is evaluated for PM6, AM1, OM3, and SCC-DFTB (enhanced by standard empirical dispersion corrections) with several test sets for noncovalent interactions and is shown to reach the quality of current DFT-D approaches for these types of problems.

1. Introduction

The ability to perform fast and accurate computer simulations of biomolecular systems has the potential to bring new insight and application opportunities in several scientific fields, for example, the development of selective receptors, catalysts, and enzyme inhibitors in computational drug design. Complementary computational methods for de novo drug design and virtual screening have already made striking successes possible, for example, through computer-aided drug lead generation and optimization.[1,2] Although these approaches can support and complement drug design, they can not be seen as fully mature, because both the modeling tools used and our understanding of protein−ligand recognition principles are still limited, especially regarding the effects of protein flexibility and solvation.[3]

Even though many advanced and accurate computational methods exist, their application to large-scale simulations of biomolecules is not possible, because these methods are computationally too demanding. As a result, the method of choice for these applications is molecular mechanics (MM). Although MM performs well in many cases, it has several drawbacks: By design, it cannot describe quantum effects like, for example, changes in electronic structure, such as chemical reactions or charge transfer, and most MM models also neglect polarization effects, which were shown to be important, for example, for the solvation of biomolecules.[4] Promising tools to overcome these limitations while maintaining efficiency (allowing extensive sampling of biologically relevant molecular systems) are semiempirical (SE) quantum mechanical methods.

The application of current SE methods to biochemical problems is unfortunately not straightforward, because the structure and function of biomacromolecules are dominantly influenced by noncovalent interactions like dispersion and hydrogen-bonding,[5] that generally need very high-level quantum chemical methods to be modeled with sufficient accuracy.[6] Despite this, the past few years have seen great success with the incorporation of dispersion effects via

* To whom correspondence should be addressed. E-mail: pavel.hobza@marge.uochb.cas.cz.

[†] Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Systems.

[‡] Comenius University.

[§] Palacky University.

10.1021/ct900541n
Published on Web 12/10/2009

empirical corrections for a wide range of DFT[7,8] and also SE (e.g., PM3-D, AM1-D[9]) methods. But because substrate recognition and binding is most often dominated by electrostatics, the accurate description of these effects and especially the hydrogen-bond interactions are also of fundamental importance for any biomolecular modeling approach. Examples for the importance of hydrogen-bonding for molecular recognition are, for example, DNA base pairing, protein folding, enzyme activity, crystal structures, properties of liquids, and pharmaceutical drug solubility and activity. While electrostatics in general are not a problem for SE methods, current SE methods are known to be deficient in the description of hydrogen-bonding (with hydrogen core−core terms, missing polarization functions on hydrogen, missing orthogonalization corrections, and in general parametrization as discussed reasons, see refs 10 and 11 and references therein). We see this to be the major obstacle limiting the accuracy of SE methods when applied to biomolecules.

As classical modeling approaches are further pushed to their limit, and more and more pitfalls are coming to light,[12] the interest in improving SE for biomolecular modeling purposes grew substantially over the past few years and has led to a number of related publications: As a result of the first biomolecular application attempts with OMn[13] and SCC-DFTB[14] and explorative approaches to describe protein ligand docking with PM3[15,16] and AM1,[17] it became clear that the (earlier known) deficiencies of SE methods for the description of hydrogen bonding[18,19] are of crucial importance in these applications.[16,20] On the other hand, first large-scale SE modeling of protein structures gave promising results[21−23] and showed that the capability of SE methods to detect native structures from collections of decoys is quite remarkable.[12] In order to surpass the accuracy of the description of noncovalent interactions by MM force fields, improving the description of hydrogen-bonding interactions in SE methods is clearly necessary.

A number of approaches offering improvement in this direction have been suggested in the literature so far, for example, on the basis of additional or modified core−core terms (like PM3-PIF[24,25] and PDDG/PM3[26]), third-order terms, and modified parameters for SCC-DFTB[27] and also reformulated QM/MM interaction terms (to improve hydrogen bonding at the QM/MM interface[28]). An overview of the problem and the proposed solutions can be found in refs 10 and 11. While a significantly better performance is observed when applying these techniques, the results still leave large space for further improvements. (It is nevertheless hard to understand why a recent comparison of the performance of semiempirical QM/MM approaches with force fields[29] ignores all developments except the PDDG approaches.) Concerning force field and ab initio results, the following has to be kept in mind: A recent study that evaluated the performance of a set of widely used force fields by calculating the geometries and stabilization energies for a large collection of intermolecular complexes showed that the magnitude of hydrogen-bonding interactions are severely underestimated by all of the force fields tested.[30] And albeit much better, also the performance of DFT methods for the calculation of (especially the relative) strength of hydrogen-bond interactions is not always of satisfactorily high accuracy (see ref 31 and references therein).

Recently, our group managed to successfully open up a new path to improve SE methods for hydrogen-bonding interactions: We augmented the new PM6 method[32] with empirical corrections for dispersion and hydrogen-bonding interactions (referred to as PM6-DH1 in the following)[33] and were able to achieve large improvements in accuracy for interaction energies of biologically relevant, noncovalently bound systems. PM6 was chosen, because this model is parametrized for 80 elements and was shown to be one of the most accurate SE approaches for a wide range of problems.[32] Furthermore, PM6 is implemented also as a linear-scaling, localized molecular orbital algorithm (termed MOZYME[34]) in Mopac2009[35] and VAMP 10.0, which allows the modeling of most of the proteins in the PDB (with less than about 5000 atoms) on standard desktop computers.[34] While our first-generation H-bonding correction was already a major step forward in accuracy, we have found further improvement possible, to be presented in the following.

2. Empirical H-Bonding Corrections for Semiempirical Quantum Chemical Methods

The First-Generation Correction. To incorporate the major characteristics of hydrogen-bond interactions, the first-generation correction made use of the charges q on the acceptor (A) and hydrogen (H) atoms, the H-bond distance r between these atoms, and a cosine term that promotes a 180° bonding situation for the A···H−D (with the donor atom D) angle:

$$E_{\text{H-bond}} = a\left[\frac{q_{\text{A}} \times q_{\text{H}}}{r^2} \times \cos(\theta) + b \times c^r\right] \qquad (1)$$

The parameters a, b, and c were optimized for eight different bond types, leading to overall 24 parameters for the description of common H bonds involving nitrogen and oxygen acceptor and donor atoms. As the discussion of the results in section 4 will illustrate, this approach leads to a significantly improved performance of PM6 for the description of H-bond interactions.

An in-depth analysis of our correction revealed the following improvement opportunities: While the H-bond distance and the 180° condition for the A···H−D angle are the most important geometrical features of hydrogen-bonding, two additional internal coordinates are needed to complete the sterical description by taking care of the "orientation of the lone pair" at the acceptor atom. We will show later that the full description of all important geometrical features of hydrogen-bonding in the second-generation correction is the major reason for its improved accuracy and reliability. It turns out that the change to a physically more sound description of hydrogen bonding allows us to fix two other problems of the first-generation correction: First, the second term in eq 1 is only dependent on the H-bond distance coordinate. This leads to discontinuous potentials around values of 90° for the A···H−D angle. Second, for some H-bond types, the second—meant to be

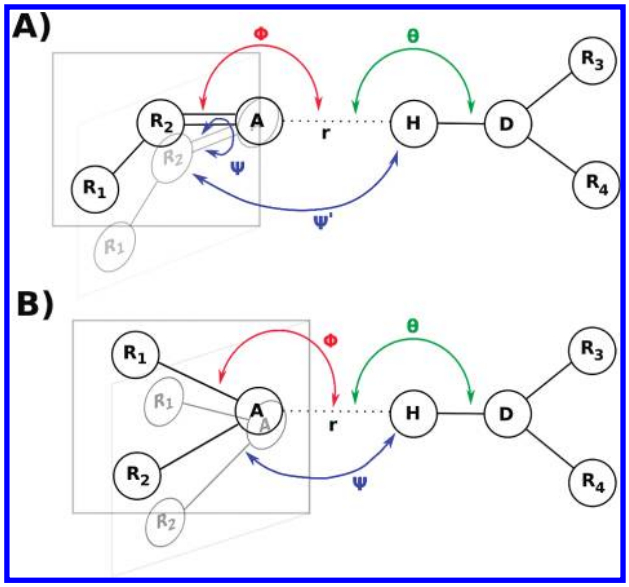

Figure 1. Illustration of the geometric features of hydrogen-bonding; see text for explanation.

repulsive—term is of an attractive nature (the optimization of the *b* parameter with constraints on the sign was tested but discarded because of a diminished accuracy of the results). By fixing these two problems in the new version, we are able to significantly increase the robustness of our empirical correction approach.

To prevent problems with the optimization of very strong H bonds (like in the case of the formic acid dimer), we had to keep the distance cutoff of the first-generation correction (see ref 33 for a more detailed discussion of this approximation). The cutoff is applied in such a way that H-bond distances below 1.8 Å are set to 1.8 Å for the calculation of the hydrogen-bond energy. It acts like a damping function that forces the correction to a constant value in the repulsive region, where the description seems to be quite deficient otherwise. For molecular dynamics simulations, this cutoff will be implemented with an interpolating polynomial to prevent kinks in the potential surface. We also tried multiplication by a damping function for the second generation correction but found no advantages over the cutoff distance solution.

The application of the above listed changes in the second-generation correction furthermore allows us to change from bond-type to atom-type parameters and to apply our ansatz also to other semiempirical methods.

The Second-Generation Correction. The sterical arrangement of the two system parts involved in a H bond can be defined with six internal coordinates (see Figure 1 for an illustration of the following explanations): the H-bond distance *r*, the two angles A···H—D (termed Θ) and R_2—A···H (termed Φ, with R_2 being a donor "base atom"), and the corresponding three torsional angles of which only one directly influences the H-bond interaction energy, R_1R_2A···H (termed Ψ). The first two mentioned coordinates, the H-bond distance and the angle Θ between acceptor, hydrogen, and donor atoms were incorporated into the first-generation correction. The second two mentioned coordinates define the relative position of the acceptor atom system part (or so to say the spatial arrangement of the acceptor lone pair), which is important to prevent nonphysical contributions to the H-bond interaction energies (e.g., through other atoms or in the case of purely dispersion-bound complexes). Figure 1 shows *r*, Θ, Φ, and Ψ for two different cases, an sp^2 oxygen-type acceptor atom (a) and sp^2 nitrogen or general sp^3-type acceptor atoms (b), which require a different choice of atoms for the definition of the torsion angle coordinate. Note that, for our choice of coordinates, the out-of-plane "movement" in case a (described by Ψ′) is actually realized by a combined change of the two internal coordinates Φ and Ψ. As a result of these considerations, the new version of our empirical H-bonding correction takes the following form:

$$E_{\text{H-bond}} = \left[a \times \frac{q_A \times q_H}{r^b} + c \times d^r \right] \times \cos(\theta) \times \cos(\phi) \times \cos(\psi) \quad (2)$$

with ϕ as the deviation of the R_2—A···H angle from the idealized optimal H-bond angle (taken as 109.48° for sp^3 and 120° for sp^2 structures) and ψ as the deviation of the R_1R_2A···H torsion angle from the idealized optimal H-bond torsion angle (taken to be 109.48° for N sp^3, 109.48/2° for other sp^3 structures, and 0° for sp^2 structures). Two special cases arise: For sp^2 oxygen acceptor atoms, not only 120° but also 180° has to be considered as the idealized optimal H-bond angle; the one with the smaller deviation to the actual binding situation is chosen for the calculation. For NR_3 nitrogen acceptor atoms, the possible planarization of this group has to be taken into account; we chose to calculate the idealized optimal ϕ and ψ values by linear extrapolation between the tetrahedral and planar values (109.48 and 90 for ϕ, 109.48/2 and 90 for ψ) subject to the actual value of the torsion angle between R_1R_2A and the remaining third of the NR_3 group. An alternative (but considering gradient derivations unnecessarily complicated) implementation of our general approach would be to use a cosine term for the angle between the H-bond and an "idealized lone-pair orientation" vector.

The resulting equation was analyzed for the importance of the different terms and parametrized for acceptor atom types in a stepwise optimization process that led to some remarkable observations illustrated by data for the H-bonded complexes of the S26 set (the S22 benchmark set[6] for noncovalent interactions, extended with four singly hydrogen-bonded complexes,[36] see also Figure 2) in Table 1:

• Even when using only the attractive term, $(q_A \times q_H)/(r^b) \times \cos(\theta) \times \cos(\phi) \times \cos(\psi)$, with overall only a single parameter (termed P1 in Table 1: $b = 3.2$) to correct PM6-D for all different types of nitrogen and oxygen hydrogen-bond interactions, our new correction leads to very good (PM6-DH1-quality) values for almost all S26 H-bond interaction energies (please note that the intermediate parametrization results are given to show the robustness of our approach and that the Pn parametrizations are therefore not meant to be some kind of "intermediate" methods).

• The additional distinction between nitrogen and oxygen acceptor atom types (resulting in overall two parameters

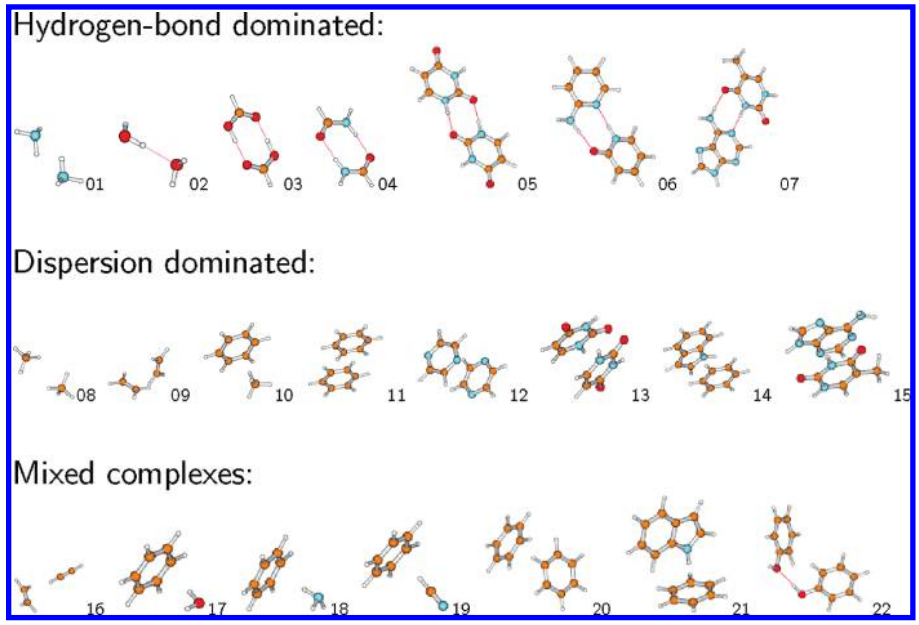

Figure 2. The S22 benchmark set for noncovalent interactions.[6]

Table 1. Intermediate Results from the Optimization Process[a]

S26 entry	reference	PM6	PM6-D	PM6-DH1	P1[b]	P2[b]	P3[b]	PM6-DH2	AM1-D*	AM1-D*/P1[b]	DFTB-D	DFTB-D/P1[b]	OM3-D	OM3-D/P1[b]
ammonia dimer	−3.17	0.86	0.33	−0.55	−0.14	−0.32	−0.33	−0.04	1.45	0.55	2.57	2.23	−0.28	−0.89
water dimer	−5.02	1.08	0.70	0.35	−1.18	−0.95	−0.05	0.12	1.16	−1.78	1.80	0.09	−5.44	4.15
formic acid dimer	−18.61	7.47	6.47	1.23	−0.08	0.63	−0.08	−0.03	18.26	6.62	2.49	−3.57	−2.53	0.49
formamide dimer	−15.96	3.40	2.18	0.56	−2.21	−1.70	0.34	0.09	7.74	−1.44	2.67	−0.33	−2.12	−0.66
uracil dimer C_{2h}	−20.65	7.32	5.55	1.82	0.37	0.96	0.37	−0.56	11.76	0.43	3.74	−0.17	−0.60	−0.07
2-pyridoxine/2-aminopyridine	−16.71	6.72	4.50	−0.64	0.69	0.41	0.18	0.37	8.09	0.83	4.84	2.49	−2.47	1.08
adenine/thymine Watson/Crick	−16.37	7.30	4.89	−1.46	0.67	0.20	−0.02	−0.09	7.65	−0.68	5.80	3.29	−0.49	−0.07
phenol dimer	−7.05	3.67	1.32	0.31	0.06	0.21	0.06	−0.01	0.96	−0.93	3.02	1.96	−1.84	0.00
methanol dimer	−5.70	2.20	1.35	0.26	−0.25	−0.06	−0.25	−0.55	2.45	−0.15	2.45	1.04	−1.42	0.88
methanol/formaldehyde	−5.31	1.89	1.23	0.47	0.12	0.26	0.12	0.10	1.78	−0.09	2.46	1.62	−2.27	2.26
methyl amide dimer (α)	−6.69	1.76	0.38	−0.34	−0.93	−0.84	−0.27	0.31	1.59	−1.48	0.92	0.07	−0.85	−0.01
methyl amide dimer (β)	−7.65	1.78	0.72	−0.05	−1.00	−0.80	−0.01	−0.02	3.98	0.49	1.59	0.45	−0.68	−0.15
MSE		3.79	2.47	0.16	−0.32	−0.17	0.01	0.04	5.57	0.20	2.86	0.76	−1.75	0.60
MUE		3.79	2.47	0.67	0.64	0.61	0.19	0.19	5.57	1.29	2.86	1.44	1.75	0.89
RMSE		4.56	3.26	0.85	0.88	0.76	0.28	0.24	7.57	2.12	3.15	1.89	2.22	1.47
$\Delta_{Max-Min}$		6.61	6.13	3.28	2.90	2.66	0.93	0.84	17.30	8.40	4.88	6.86	5.16	5.05

[a] Errors, mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol. AM1-D* refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D, see text for details. [b] For explanation see text, PM6-D and PM6-D/Pn with PM6-DH1 dispersion parameters.

(termed P2)), then leads to comparably small but still significant improvements.

• In the case of PM6-D, the accurate description of H-bonding interactions involving water and peptide systems requires the inclusion of at least one additional atom-type parameter (termed P3).

• Also for AM1-D* (not AM1-D, see next paragraph for explanation), SCC-DFTB-D and OM3-D, a significant improvement is found with only a single parameter (b = 2.2 for AM1, 3.2 for DFTB, and 3.1 for OM3). For these methods, no additional parameters for water or peptide systems are necessary, but one additional parameter is required for the proper description of acid acceptor atoms.

We decided not to use the AM1-D method published by McNamara and Hillier[9] with our correction scheme, because their method is not simply AM1 with a standard empirical dispersion correction but is additionally based on a refit of 18 AM1 parameters that also account to some extent for hydrogen bonding. The variant we use here, termed AM1-D* in the following, refers to standard AM1 with a standard empirical dispersion correction of the Jurečka type.[8] The optimization of the D parameters for the dispersion-bound complexes of the S22 benchmark set led to the values s_r = 0.91, α = 56.0, and s_6 = 1.18. This way, the mean unsigned error for the dispersion-bound S22 complexes was decreased from 6.7 kcal/mol for AM1 to 0.4 kcal/mol for AM1-D*, slightly lower than the 0.6 kcal/mol found for AM1-D.[9]

We also adjusted the dispersion correction for PM6, because we have found that PM6-DH1 overestimates dispersion effects in saturated systems. The new parameters are s_r = 1.04, α = 20.0, and s_6 = 0.89 (instead of s_r = 1.07, α = 11.0, and s_6 = 1.00 for the old version), in combination with c_6 = 0.95 for sp^3 carbon and c_6 = 1.65 for other carbon atoms (instead of c_6 = 1.65 for all carbon atoms) and a hydrogen van der Waals radius of 156 pm (instead of 120 pm). The effects of these changes are very small for the

systems investigated here but become significant, for example, for large, saturated hydrocarbon chains. We decided to nevertheless include these changes here to avoid spreading the details of PM6-DH2 over multiple publications.

As a training set for the stepwise H-bonding correction parameter optimization procedure, the equilibrium structures of the original H-correction training set[33] (105 hydrogen-bond interaction energies) were extended with the (noncharged) 37 hydrogen-bonded DNA base pairs and 13 peptide interaction energies from the JSCH2005 set.[6] Another optimization run on the much smaller S26 test led to essentially the same parameter values, showing on one hand the stability of our approach with respect to the chosen parameters and on the other hand the usefulness of this test set for the parametrization of new methodological developments for H-bonding interactions. (Due to the implications of the limited size of the S26 complexes we do not assume this to be equally true for empirical dispersion corrections).

Because the fitting of correction terms only to equilibrium structure data is prone to result in problems in real-life applications, we extended in the next step the S26 set with the S22 × 4 set,[37] which contains four nonequilibrium structures with high-level reference data for each of the S22 complexes. While not necessary at all for the description of the equilibrium structures, the second (constrained to be repulsive) term in eq 2 was included to further improve the accuracy at short distances, especially in the case of very strong hydrogen bonds like those found in the formic acid dimer.

The inclusion of more parameters for a (reasonably) larger number of acceptor atom types or additional parameters for donor atom types led to no significant improvements. In addition, it was found that method-independent values can be chosen for the parameters *b*, *c*, and *d* in eq 2, because slight differences are absorbed by the method-dependent *a* parameters. For the final parametrization of the new correction for the PM6-D, AM1-D*, OM3-D, and SCC-DFTB-D methods on the combined S26/S22 × 4 set, fixed values of *b*, *c*, *d*, and five acceptor atom types with different *a* parameters were chosen. Albeit a significant improvement can already be found with only one *a* parameter, the additional increase of accuracy (especially for water and peptides in the case of PM6-D) outweighed our concerns of using five different acceptor-atom-type-based *a* parameters. This view was further supported by the rather well-behaved nature of the parameters (that nicely reflect the capabilities of the underlying SE methods to describe hydrogen bonding) and the overall number of parameters in SE methods.

As a result, the outlined procedure led to three global and five method-dependent parameters for the description of common H bonds involving nitrogen and oxygen acceptor and donor atoms. One additional method-dependent parameter for the description of H bonds involving sulfur acceptor atoms was generated accordingly for every method (except OM3, where no sulfur SE parameters were available to us), using the sulfur hydrogen-bonded DNA base pairs from the JSCH2005 set.[6] The final parameters are shown in Table 2. We believe that the differences of the individual parameter values are more likely to reflect advantages and deficiencies of the parametrization of the underlying SE methods, rather than physical issues of different H-bonding interactions. As noted before, the qualitative parameter differences between methods nicely reflect the initial capability of the underlying SE methods to describe hydrogen bonding (with a rather bad performance of AM1 and a quite good performance of OM3 at the two ends of the scale).

Table 2. Final Parameters

parameter	element	PM6	AM1	OM3	DFTB
		Global			
b	all	3.0			
c	all	0.65			
d	all	5.0			
		Method-Dependent			
a	N	1.48	4.54	0.86	4.41
	O	1.56	3.75	0.75	1.84
	O_{acid}	1.55	5.55	1.51	1.15
	$O_{peptide}$	0.96	3.46	0.78	1.56
	O_{water}	0.76	3.52	0.49	1.57
	S	0.85	1.05	−[a]	0.53

[a] OM3 sulfur parameters unavailable.

We also tested our correction with third-order SCC-DFTB with and without a modified γ parameter for an improved description of hydrogen bonding[27] but ended up with the same accuracy as with SCC-DFTB-DH2 (which is significantly higher than third-order SCC-DFTB with a modified γ parameter, giving an MAD of 1.0 and an error span of 6.0 kcal/mol for the S26 test set).

As the last step, an analytical gradient for the proposed correction was implemented. This was done analogously to the first-generation correction, that is, without derivatives of the atom charges. This approximation was found to have only a minor impact for the cases investigated here and allows us to keep our approach simple and fast but surely needs deeper investigation in the future.

3. Computational Details

Semiempirical PM6 and AM1 calculations applying the MOZYME algorithm were done with MOPAC2009,[35] OM3 calculations with MNDO2005, and SCC-DFTB calculations with DFTB+.[38] TPSS[39] and B3-LYP[40,41] DFT calculations with empirical dispersion corrections of the Jurecka type[8] were done with Turbomole 5.10[42] using TZVPP[43] Gaussian AO basis sets and the RI approximation[44,45] for two-electron integrals. The second-generation H-bonding correction is implemented as an add-on correction to MOPAC2009, mndo99 and DFTB+ in our own development code (the latest version of this software can be obtained from the authors upon request), and will be included in a future release of MOPAC2009.

4. Results and Discussion

Tables 3−8 show results of PM6, AM1, OM3, and SCC-DFTB calculations with dispersion correction and first- and second-generation H-bonding corrections for the S26 (Table 3), S22 (Table 4, in additional comparison to literature data), and S26+S22 × 4 benchmark sets (Table 5), the PM6-DH1 training set of 105 small hydrogen-bonded complexes (Table

Table 3. Results for the S26 Set[a]

S26 entry	CCSD(T)/CBS	PM6	PM6-D	PM6-DH1	PM6-DH2	TPSS-D[a]	B3LYP-D[a]	AM1	AM1-D*	AM1-DH2	DFTB	DFTB-D	DFTB-DH2	OM3	OM3-D	OM3-DH2
ammonia dimer	−3.17	0.86	0.57	−0.57	−0.04	0.57	0.71	2.38	1.45	0.67	2.82	2.57	1.12	1.24	0.49	0.06
water dimer	−5.02	1.08	0.91	0.35	0.12	1.31	1.49	2.13	1.16	−1.43	1.82	1.80	−0.63	0.93	0.28	0.09
formic acid dimer	−18.61	7.47	6.76	1.23	−0.03	0.85	0.70	20.14	18.26	1.09	2.74	2.49	−1.11	7.06	5.44	0.26
formamide dimer	−15.96	3.40	2.49	0.56	0.09	0.17	0.47	10.23	7.74	−0.97	3.40	2.67	−0.32	4.28	2.53	1.07
uracil dimer C_{2h}	−20.65	7.32	5.88	1.82	−0.56	−0.14	0.37	14.85	11.76	−0.30	4.98	3.74	−1.64	4.47	2.12	0.28
2−pyridoxine/2−aminopyridine	−16.71	6.72	4.97	−0.64	0.37	1.16	1.06	12.25	8.09	0.15	6.45	4.84	−0.91	5.36	2.47	1.51
adenine/thymine Watson/Crick	−16.37	7.30	5.37	−1.46	−0.09	0.71	0.78	12.08	7.65	−2.27	7.58	5.80	−1.54	5.03	1.84	0.39
methane dimer	−0.53	0.56	0.18	−0.10	0.18	0.17	−0.03	0.85	−0.18	−0.18	0.54	0.06	0.06	0.67	−0.08	−0.08
ethene dimer	−1.51	1.11	0.45	−0.01	0.45	0.17	0.11	1.38	−1.25	−1.25	1.32	0.70	0.70	1.65	−0.17	−0.17
benzene/methane	−1.50	1.02	0.11	−0.25	0.11	−0.39	−0.41	1.90	−0.70	−0.70	1.32	0.29	0.29	1.63	−0.01	−0.01
benzene dimer stacked	−2.73	2.84	−0.85	−0.90	−0.85	−0.31	−0.88	6.24	−0.38	−0.38	3.10	−0.37	−0.37	3.86	−1.14	−1.14
pyrazine dimer	−4.42	2.60	−0.93	−1.00	−0.93	−0.75	−0.98	6.91	−0.16	−0.16	4.11	0.79	0.79	3.74	−1.47	−1.47
uracil dimer C_2	−10.12	5.66	0.70	0.42	0.67	−1.04	−0.52	10.23	−0.02	−0.06	6.17	1.94	1.92	6.16	−1.29	−1.30
indole/benzene stacked	−5.22	5.28	0.16	0.01	0.16	−1.00	−1.64	10.60	0.77	0.77	5.46	0.63	0.63	6.60	−0.76	−0.76
adenine/thymine stacked	−12.23	7.29	0.57	−0.55	0.54	−1.07	−0.84	15.14	0.08	0.04	8.34	2.10	2.08	8.93	−1.87	−1.89
ethene/ethine	−1.53	0.98	0.58	0.42	0.58	0.02	−0.09	1.18	0.13	0.13	0.99	0.54	0.54	0.85	0.10	0.10
benzene/water	−3.28	1.00	0.10	−0.13	0.10	0.49	0.43	2.59	0.54	0.54	2.00	1.62	1.62	1.53	−0.01	−0.01
benzene/ammonia	−2.35	0.82	−0.20	−0.42	−0.20	−0.10	−0.13	2.02	−0.41	−0.41	1.84	0.77	0.77	1.58	−0.04	−0.04
benzene/HCN	−4.46	2.48	1.48	1.27	1.48	−0.64	−0.77	3.65	1.53	1.53	2.73	1.64	1.64	2.78	0.77	0.77
benzene dimer T−shaped	−2.74	1.98	0.15	−0.10	0.15	−0.81	−0.95	3.10	−0.53	−0.53	2.42	0.69	0.69	2.85	0.09	0.09
indole/benzene T−shaped	−5.73	3.32	0.79	0.42	0.79	−0.84	−1.25	4.67	0.44	0.44	4.05	1.71	1.71	4.56	0.79	0.79
phenol dimer	−7.05	3.67	1.84	0.31	−0.01	0.11	0.52	5.69	0.96	−0.85	4.25	3.02	1.19	3.89	0.60	0.24
methanol dimer	−5.70	2.20	1.72	0.26	−0.55	1.28	1.51	4.00	2.45	−0.10	2.66	2.45	0.07	2.58	1.42	1.11
methanol/formaldehyde	−5.31	1.89	1.51	0.47	0.10	0.12	0.61	3.39	1.78	0.16	2.77	2.46	1.29	3.28	2.27	2.38
methylamide dimer (α)	−6.69	1.76	0.74	−0.34	0.31	0.08	0.34	3.81	1.59	0.49	1.98	0.92	0.39	2.44	0.85	0.53
methylamide dimer (β)	−7.65	1.78	0.99	−0.05	−0.02	−0.15	0.36	5.85	3.98	0.36	2.38	1.59	0.40	1.93	0.68	0.03
Complete S26 Set																
MSE		3.17	1.42	0.04	0.11	−0.00	0.04	6.43	2.57	−0.12	3.39	1.83	0.44	3.46	0.61	0.11
MUE		3.17	1.58	0.54	0.36	0.56	0.69	6.43	2.85	0.61	3.39	1.85	0.94	3.46	1.14	0.64
RMSE		3.94	2.46	0.71	0.51	0.69	0.82	8.18	5.17	0.81	3.93	2.31	1.10	4.04	1.63	0.91
$\Delta_{Max-Min}$		6.91	7.69	3.28	2.41	2.38	3.15	19.29	19.51	3.80	7.80	6.17	3.72	8.26	7.31	4.27
Hydrogen-Bonded Systems																
MSE		3.79	2.81	0.16	−0.03	0.51	0.74	8.07	5.57	−0.25	3.65	2.86	−0.14	3.54	1.75	0.66
MUE		3.79	2.81	0.67	0.19	0.55	0.74	8.07	5.57	0.74	3.65	2.86	0.88	3.54	1.75	0.66
RMSE		4.56	3.55	0.85	0.27	0.73	0.84	9.77	7.57	0.96	4.05	3.15	1.01	3.95	2.22	0.96
$\Delta_{Max-Min}$		6.61	6.19	3.28	0.93	1.46	1.17	18.01	17.30	3.36	5.76	4.88	2.93	6.13	5.16	2.35

[a] TZVP basis set. [a] Errors, mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol.

Table 4. Results for the S22 Set[a]

method	MUE
MP2/CBS	0.8[c]
B3LYP-D/TZVP	0.7
TPSS-D/TZVP	0.6
M08-HX/6-311+G(3df,2p)/CP	0.5[c]
M06−2X/6-311+G(3df,2p)/CP	0.4[c]
TPSS-D/6-311++G(3df,3pdf)	0.3[d]
B2-PLYP-D/TZVPP/0.5CP	0.3[e]
$PM3_{BP}$	5.2[f]
SCC-DFTB-D	1.9
OM3-D	1.1
PM3-D[b]	0.9[f]
AM1-D[b]	0.9[f]
PM6-DH1	0.6
SCC-DFTB-DH2	1.0
AM1-DH2	0.7
OM3-DH2	0.6
PM6-DH2	0.4

[a] Comparison of mean unsigned errors (MUE) with respect to the benchmark CCSD(T)/CBS interaction energies for various wave function theory, density-function theory and enhanced semiempirical quantum chemical methods are presented. All values in kcal/mol. (0.5)CP stands for (half) counter-poise corrected values. [b] With 18 adjusted AM1/PM3 parameters, see ref 9. [c] From ref 46. [d] From ref 8. [e] From ref 47. [f] From ref 9.

6), the 37 noncharged, H-bonded DNA base pair complexes from the JSCH2005 set (table 7), and the 13 noncharged, H-bonded peptide-structures from the JSCH2005 test set (Table 8) with corresponding TPSS-D/TZVP and B3-LYP/TZVP data for comparison. Each table shows the reference interaction energy at the CCSD(T)/CBS level, errors relative to these values for the investigated methods, followed by statistical measures overs these errors: the mean signed error (MSE), mean unsigned error (MUE), root-mean-square error (RMSE), and the error span ($\Delta_{Max-Min}$). The errors are calculated so that a positive error means that the investigated method underestimates the binding energy and vice versa.

The general trends for the different benchmark sets are very similar, so that the observations can be summarized altogether in the following way: The standard SE semiempirical methods perform quite badly for both dispersion and hydrogen-bonding interactions, but PM6, OM3, and SCC-DFTB (S26 MUEs around 3 kcal/mol) are significantly more accurate than AM1 (S26 MUE around 6 kcal/mol). The inclusion of empirical dispersion corrections is a great improvement for all tested semiempirical methods. With these corrections, the semiempirical methods are able to model dispersion bound complexes with comparably high accuracy (MUEs between 1 and 3 kcal/mol), so that the largest remaining errors are found for hydrogen-bond interactions.

Table 5. Results for the Combined S26+S22 × 4 Sets (114 Entries)[a]

	PM6	PM6-D	PM6-DH2	TPSS-D[b]	B3LYP-D[b]	AM1	AM1-D*	AM1-DH2	DFTB	DFTB-D	DFTB-DH2	OM3	OM3-D	OM3-DH2
MSE	2.23	1.04	0.14	−0.56	−0.63	4.43	1.86	0.06	2.42	1.34	0.38	2.20	0.51	−0.10
MUE	2.23	1.17	0.36	0.58	0.63	4.43	2.23	0.89	2.42	1.49	0.68	2.20	0.88	0.72
RMSE	3.21	2.07	0.61	0.69	0.72	7.00	4.87	1.72	3.20	2.18	0.88	2.98	1.49	1.29
$\Delta_{Max-Min}$	9.62	9.69	4.96	1.68	1.46	28.52	29.15	14.40	10.21	10.64	4.35	8.97	9.76	7.71

[a] Mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol. [b] TZVP basis set.

Table 6. Results for 105 Small, Hydrogen-Bonded Complexes of the PM6-DH1 Training Set[a]

	PM6	PM6-D	PM6-DH2	TPSS-D[b]	B3LYP-D[b]	AM1	AM1-D*	AM1-DH2	DFTB	DFTB-D	DFTB-DH2	OM3	OM3-D	OM3-DH2
MSE	−2.63	−1.66	−0.43	0.90	1.01	−5.06	−2.55	−0.12	−3.19	−2.33	−0.40	−2.67	−0.88	−0.51
MUE	2.64	1.76	1.15	0.91	1.01	5.06	2.70	1.59	3.21	2.36	0.85	2.67	0.91	0.66
RMSE	3.16	2.35	1.54	1.08	1.15	6.07	4.03	2.12	3.56	2.79	1.06	2.87	1.14	0.86
$\Delta_{Max-Min}$	9.22	9.61	7.37	3.66	3.52	22.75	22.64	12.13	10.32	10.47	5.15	6.73	6.52	4.48

[a] Mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol. [b] TZVP basis set.

Table 7. Results for the Noncharged Hydrogen-Bonded JSCH2005 DNA Base Pairs (37 Entries)[a]

	PM6	PM6-D	PM6-DH2	TPSS-D[b]	B3LYP-D[b]	AM1	AM1-D*	AM1-DH2	DFTB[c]	DFTB-D[c]	DFTB-DH2[c]	OM3[d]	OM3-D[d]	OM3-DH2[d]
MSE	−8.07	−6.06	−0.83	0.42	0.55	−14.08	−9.56	−0.01	−7.21	−5.34	2.09	−5.67	−2.41	−0.80
MUE	8.07	6.06	1.85	0.72	0.70	14.08	9.56	2.20	7.21	5.34	2.78	5.67	2.49	1.21
RMSE	8.23	6.23	2.36	0.97	0.97	14.71	10.33	2.84	7.47	5.64	3.25	5.89	2.79	1.44
$\Delta_{Max-Min}$	8.73	7.67	8.84	3.97	3.86	18.85	16.43	12.43	7.88	6.85	9.58	7.01	6.81	4.53

[a] Mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol. [b] TZVP basis set. [c] Without adenine/fluorotoluene Watson/Crick complex because of missing fluorine parameters. [d] Without seven thio base pairs because of missing sulfur parameters.

Table 8. Results for the Hydrogen-Bonded JSCH2005 Peptides (13 Entries)[a]

	PM6	PM6-D	PM6-DH2	TPSS-D[b]	B3LYP-D[b]	AM1	AM1-D*	AM1-DH2	DFTB	DFTB-D	DFTB-DH2	OM3[c]	OM3-D[c]	OM3-DH2[c]
MSE	−2.97	−0.19	−0.13	−0.35	−0.45	−4.49	1.40	1.47	−3.77	−0.80	−0.72	−3.96	0.33	0.36
MUE	2.97	0.68	0.71	0.67	0.60	4.49	1.49	1.56	3.77	0.87	0.79	3.96	0.60	0.62
RMSE	3.24	0.87	0.89	0.80	0.82	4.95	1.91	2.00	4.01	1.01	0.94	4.21	0.80	0.81
$\Delta_{Max-Min}$	4.56	3.44	3.41	2.53	2.59	7.31	4.30	4.27	5.34	2.91	2.84	5.20	2.81	2.79

[a] Mean signed error (MSE), mean unsigned error (MUE), root mean square error (RMSE), and the error span $\Delta_{Max-Min}$ with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values in kcal/mol. [b] TZVP basis set. [c] Without seven thio base pairs because of missing sulfur parameters.

As mentioned before, the inclusion of our first-generation H-bonding correction in PM6-DH1 is already a major step toward a higher accuracy for these interactions (with a MUE of 0.7 kcal/mol for the hydrogen-bonding interactions in the S26 set). The largest errors are found for double hydrogen bonds, because of the parametrization to single hydrogen bonds and the higher likeliness of nonphysical contributions to the H-bonding correction in these cases.

The new correction manages to reach even higher accuracy (with a corresponding MUE of 0.2 kcal/mol) and greatly reduced error span (from 3.3 to 0.8 kcal/mol). Furthermore, it can be seen that the new H-bonding correction does not lead to nonphysical interaction energy contributions for purely dispersion-bound complexes (the values for PM6-DH being essentially the same as for PM6-D). Albeit a less accurate final performance is found for AM1, OM3, and DFTB when compared to PM6, the large decrease of errors (especially for AM1) is still impressive for these SE methods. As we focused on PM6 during the initial development phase, we do not want to exclude the possibility that further improvements are possible, especially for AM1, for which the chosen repulsive term seems to fit least well.

While the results for the hydrogen-bonded complexes of the PM6-DH1 training set and the hydrogen-bonded JSCH2005 peptides (both with smaller interaction energies of −6.2 and −4.4 kcal/mol and quite good values already for the "pure" SE methods) are less impressive, the hydrogen-bonded JSCH2005 DNA base pairs set (with an average interaction energy of −19.5 kcal/mol) shows how large the gain of applying the second generation H-bonding correction can be, if H-bonding interaction energies become larger. We believe that the rather poor performance for the peptide test set stems at least partly from an unbalanced description of dispersion and hydrogen-bond interactions through the combination of the two empirical corrections, which will be addressed in our future work.

It can nevertheless be stated that the obtained quality of the PM6-DH2, AM1-DH2, OM3-DH2, and DFTB-DH2 calculations reaches the accuracy of DFT-D methods (with TPSS-D/TZVP being one of the most accurate for the noncovalent interactions) for a large part of the investigated cases, while being several orders of magnitude faster. For the S22 set (included in our fit set, but also used as fit set for DFT corrections), PM6-DH2 (MUE 0.4 kcal/mol) nearly

catches up even with very sophisticated DFT approaches, like B2-PLYP-D/TZVPP/0.5CP[47] (MUE 0.3 kcal/mol) and M06−2X/6-311+G(3df,2p)/CP[46] (MUE 0.4 kcal/mol).

5. Concluding Remarks

With the presented work, we have made an attempt to alleviate the problem of the deficient description of H bonding in semiempirical quantum-chemical methods, a problem that is of vital importance for any application of these methods in computer simulations of biomolecular systems. We have proposed a transferable H-bonding correction for semiempirical methods that was shown to improve the performance of PM6, AM1, OM3, and SCC-DFTB for several test sets of H-bonding interactions by up to an order of magnitude.

The most appealing features of the new H-bonding correction are as follows:

1. Compared to the existing core−core term modifications for the improvement of the description of hydrogen bonding, our solution is able to achieve higher accuracy through the incorporation of the geometric features of H bonding.

2. Compared to the first version of our correction scheme, the new version avoids physically unsound equation terms and parameters and is now more robust and furthermore transferable between different semiempirical methods.

3. Even with only a single parameter, the new correction scheme gives significantly better results than most published correction schemes.

4. With the final parametrization (based on three global and five method-dependent parameters), the new scheme outperforms all other published H-bonding corrections by a significant margin and yields results comparable with current DFT-D approaches for a large part of the investigated cases.

5. Several details (e.g., well-behaved parameter values within and between methods, etc.) indicate the robustness of our general idea (it should nevertheless be kept in mind that our purely additive scheme cannot correct possible artifacts of the underlying methods, for example, in the case of dissociation curves).

A remaining major drawback, and subject to our future work, is the inability of our scheme to model processes where the acceptor atom-type changes, for example, proton transfer reactions. A second important step would be the combination of the second-generation H-bonding correction with an improved dispersion correction beyond a simple parameter refit of the current approach (as now the dispersion correction has become the overall less accurate part of PM6 with empirical dispersion and H-bonding corrections). From a longer perspective, we would expect it to be fruitful to incorporate empirical DH corrections directly into the parametrization process of semiempirical methods to avoid double counting and allow for a better balance of the different interactions types.

Acknowledgment. The authors would like to thank Marcus Elstner and Michael Gauss for making tests with third-order SCC-DFTB possible. This work was a part of Research Project Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and it was supported by Grant LC512 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. M.P. was supported by the Slovak Research and Development Agency under Contract No. APVV-20-018405, and the Slovak Grant Agency VEGA under Contract No. 1/0428/09. This work was also supported by Korea Science and Engineering Foundation (World Class Univ. program: R32-2008-000-10180-0).

Supporting Information Available: Calculated interaction energies for all mentioned methods and test sets. This material is available free of charge via the Internet at http://pubs.acs.org

References

(1) Jorgensen, W. L. Acc. Chem. Res. 2009, DOI: 10.1021/ar800236t.

(2) Jorgensen, W. L. *Science* **2004**, *303*, 1813.

(3) Klebe, G. *Drug Discovery Today* **2006**, *11*, 580.

(4) van der Vaar, A.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1999**, *121*, 9182.

(5) Hobza, P.; Müller-Dethlefs, K. *Non-Covalent Interactions. Theory and Experiment*: RSC Publishing: London, 2010; pp 1−220.

(6) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(7) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1436.

(8) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.

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

(10) Clark, T. J. *THEOCHEM* **2000**, *530*, 1.

(11) Winget, P.; Selcuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254.

(12) Wollacott, A. M.; Merz Jr, K. M. *J. Chem. Theory Comput.* **2007**, *3*, 1609.

(13) Möhle, K.; Hofmann, H.-J.; Thiel, W. *J. Comput. Chem.* **2001**, *22*, 509.

(14) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.

(15) Nikitina, E.; Sulimov, V.; Zayets, V.; Zaitseva, N. *Int. J. Quantum Chem.* **2004**, *97*, 747.

(16) Nikitina, E.; Sulimov, V.; Grigoriev, F.; Kondakova, O.; Luschekina, S. *Int. J. Quantum Chem.* **2006**, *106*, 1943.

(17) Vasilyev, V.; Bliznyuk, A. *Theor. Chem. Acc.* **2004**, *112*, 313.

(18) Dannenberg, J. J. *THEOCHEM* **1997**, *401*, 279.

(19) Csonka, G. I.; Angyan, J. G. *THEOCHEM* **1997**, *393*, 31.

(20) Villar, R.; Gil, M. J.; Garcia, J. I.; Martinez-Merino, V. *J. Comput. Chem.* **2005**, *26*, 1347.

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

(22) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. *Drug Discovery Today* **2007**, *12*, 725.

(23) Thiriot, E.; Monard, G. *THEOCHEM* **2009**, *898*, 31.

(24) Bernal-Uruchurtu, M. I.; Ruiz-Lopez, M. F. *Chem. Phys. Lett.* **2000**, *330*, 118.

(25) Monard, G.; Bernal-Uruchurtu, M. I.; Van Der Vaart, A.; Merz, K. M., Jr.; Ruiz-Lopez, M. F. *J. Phys. Chem. A* **2005**, *109*, 3425.

(26) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.

(27) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.

(28) Wang, Q.; Bryce, R. A. *J. Chem. Theory Comput.* **2009**, DOI: 10.1021/ct9002674.

(29) Seabra, G. M.; Walker, R. C.; Roitberg, A. E. *J. Phys. Chem. A.* **2009**, DOI: 10.1021/jp903474v.

(30) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944.

(31) Rao, L.; Ke, H.; Fu, G.; Xu, X.; Yan, Y. *J. Chem. Theory Comput.* **2009**, *5*, 86.

(32) Stewart, J. J. P. *J. Mol. Mod.* **2007**, *13*, 1173.

(33) Řezáč, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.

(34) Stewart, J. P. *J. Mol. Mod.* **2009**, *15*, 765.

(35) OPENMOPAC. www.openmopac.net (accessed Aug 31, 2009).

(36) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257.

(37) Grafova, L.; Pitonak, M.; Hobza, P. To be published.

(38) DFTBplus. http://www.dftb-plus.info (accessed Aug 31, 2009).

(39) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(40) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(41) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623.

(42) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

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

(44) Eichhorn, K.; Treutler, O.; öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.

(45) Eichhorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119.

(46) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.

(47) Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.

CT900541N