JCTC Journal of Chemical Theory and Computation

Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields

Martin Korth*

Theory of Condensed Matter Group, Cavendish Laboratory, 19 J J Thomson Avenue, Cambridge CB3 0HE, United Kingdom

Received July 21, 2010

Abstract: Computational modeling of biological systems is a rapidly evolving field that calls for methods that are able to allow for extensive sampling with systems consisting of thousands of atoms. Semiempirical quantum chemical (SE) methods are a promising tool to aid with this, but the rather bad performance of standard SE methods for noncovalent interactions is clearly a limiting factor. Enhancing SE methods with empirical corrections for dispersion and hydrogen-bonding interactions was found to be a big improvement, but for the hydrogen-bonding corrections the drawback of breaking down in the case of substantial changes to the hydrogen bond, e.g., proton transfer, posed a serious limitation for its general applicability. This work presents a further improved hydrogen-bonding correction that can be generally included in parameter fitting procedures, as it does not suffer from the conceptual flaws of previous approaches: hydrogen bonds are now treated as an interaction term between electronegative acceptor and donor atoms, "weighted" by a function of the position of H atoms between them, and multiplied with a damping function to correct the short- and long-range behavior. The performance of the new approach is evaluated for PM6, AM1, OM3, and SCC-DFTB as well as several force-field (FF) methods for a number of standard benchmark sets with hydrogen-bonded systems. The new approach is found to reach the same accuracy as the second-generation hydrogen-bonding correction with less parameters, while it avoids among other issues the conceptual problem with electronic structure changes. SE methods augmented this way reach the accuracy of DFT-D approaches for a large number of cases investigated, while still being about 3 orders of magnitude faster. Moreover, the new correction scheme is transferable also to FF methods that were shown to have serious problems with hydrogen-bonding interactions.

1. Introduction

Many promising applications of computational methods like computer-aided drug design are related to large-scale simulations of biologically relevant molecular systems. While significant successes have already been achieved, e.g., in computer-aided drug lead generation and optimization,[1,2] the field is still confronted with serious challenges, especially considering the effects of protein flexibility and solvation.[3] As a possibly very valuable tool for tackling these problems, semiempirical quantum chemical (SE) methods have come into the focus of several groups in recent years.[4−12] SE methods offer a compromise between the accuracy of "full" ab initio treatments and the speed of force field (FF) approaches. This way, SE methods allow for extensive sampling of large systems, while keeping the ability to describe the effects of electronic structure changes. The latter point is of high importance, because customary used FF point charge models ignore effects such as charge transfer and polarization[13] that are likely to be quite important in biomolecular modeling applications.[14]

But biomacromolecules are dominantly influenced by noncovalent interactions like dispersion and hydrogen bonding that generally need very high-level quantum chemical methods to be modeled with sufficient accuracy.[15] In this sense, it comes to no surprise that standard SE methods perform rather poorly for these types of interaction. A big

* E-mail: mk642@cam.ac.uk and dgd@uni-muenster.de.

10.1021/ct100408b
Published on Web 11/01/2010

$cos(\theta)^n$ like **DH1**, FS1	H-bond	f_{geom} like **DH2**, DH+
-4.18	N...HN bond	-2.78
-0.17	cross-molecule	none
-0.54	intra-molecule	-0.02
-0.04	cross-molecule	none
-0.18	cross-molecule	none
-0.48	cross-bond	none
-0.24	cross-molecule	0.00
-0.84	cross-bond	0.00
-0.25	O...HN bond	-2.01
0.01	cross-molecule	-0.01
-0.62	cross-molecule sum	-0.01

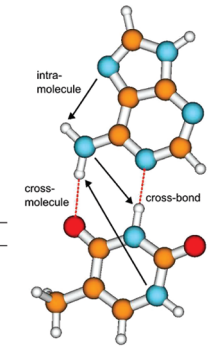
H-bond Correction Contributions for a 3225 Atom Protein[a,b]

type	No. of H-bonds[c]	overall correction
$cos(\theta)^n$ like DH1, FS1	≈ 1500	≈ 630 kcal/mol
$cos(\theta)^n$ w/o 'cross-molecule'	≈ 1200	≈ 530 kcal/mol
f_{geom} like DH2, **DH+**	≈ 230	≈ 120 kcal/mol

[a] all numbers already with long-range damping, i.e. no contributions beyond 10.5 Å

[b] in bold letters the method with which the actual numbers are produced

[c] i.e. number of contributions larger than 0.1 kcal/mol



Red dots indicate actual H-bonds, black arrows indicate unphysical contributions: Non-existing or very weak intra-molecule H-bonds, interactions with parallel H-bonds ('cross-bond'), or across the molecule core ('cross-molecule').

Figure 1. Illustration of the importance to go beyond a simple $\cos(\theta)^n$ term for hydrogen-bonding correction terms. See the text for further explanation.

step forward for the description of biomolecular systems with SE methods was made with the inclusion of empirical dispersion corrections (e.g., PM3-D, AM1-D[16]), similar to the ones used for DFT (e.g., refs 17 and 18) methods. But in contrast to DFT methods that perform acceptably well for hydrogen-bonding interactions,[19] SE methods remain as deficient as commonly used FF approaches[20] for systems beyond pure dispersion interactions. While this disadvantage has been known for many years[21−24] and several attempts to cure this remedy were successful up to a certain point,[23−29] only the inclusion of force-field-type terms for empirical hydrogen-bonding corrections was able to improve the accuracy of SE methods to a level near that of DFT-D approaches. Because of conceptual problems with the initial PM6[30,31]-based approach,[32] the method was redesigned in a physically more sound way and is now publicly available in Mopac2009[33] as the "-DH2" add-on method. One remaining major drawback of DH2 is the breakdown of the correction in the case of an acceptor-atom change, a problem that is, among other issues, addressed in the following.

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

First-Generation and Second-Generation Corrections. The first-generation correction, termed "DH" and later on "DH1", (eq 1) made use of the charges (q) on the acceptor (A) and hydrogen (H) atoms, the 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:

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

This design led to a number of problems, with the possibility of a large number of unphysical contributions to the correction from nonexisting hydrogen bonds, e.g., through the back of acceptor atoms, because the orientation of the acceptor atom is not taken into account (see below for a detailed discussion). Other problems include large discontinuities (because only the attractive but not the repulsive term is multiplied with the angular dependency), a high number of unsystematic parameters, and unphysical cutoffs.

The second-generation correction (eq 2) is a complete redesign of this approach, with the most important change being the inclusion of the missing information about the sterical arrangement of the acceptor side of the hydrogen bonds (see ref 34 for a detailed explanation). This "H2" correction uses the same 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):

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

with ϕ and ψ as the deviations of the R_2−A···H angle and R_1R_2A···H torsion angle from the idealized optimal H-bond values. This redesign also allowed for keeping terms and parameters more physically sound (e.g., avoiding the above-mentioned large discontinuities), led to a much smaller number of now systematic parameters, and made the correction transferable to other SE methods. Large systematic gains in accuracy for hydrogen-bonded complexes were possible with only one overall parameter; the final accuracy using eight fitted parameters reached the DFT-D level for a large number of investigated cases.[34]

Figure 1 illustrates how important it is to go beyond a simple $\cos(\theta)^n$ ansatz for the geometrical definition of hydrogen-bond correction terms: even in a rather small system like the Watson−Crick bound adenine/thymine base pair, the number and impact of unphysical contributions (indicated in the picture with black arrows) is quite large when the simple $\cos(\theta)^n$ term is used. These contributions sum up to enormous interaction energies for larger systems, as shown for a medium-sized protein. Please note that our numbers should be considered rather conservative estimates, because we already used a long-

range damping function, so that no interactions beyond 10.5 Å contribute to the shown values.

The major remaining drawback of the second-generation correction is the kept direct dependence on the distance between the hydrogen and the acceptor atom (and the corresponding parametrization to acceptor atom types) that requires a constant bonding situation between acceptor, hydrogen, and donor atoms, making it (unlike the common empirical dispersion corrections) a bond-type term, with all the disadvantages attached: if the acceptor atom changes (e.g., in the case of proton transfer from the donor to the acceptor), then the correction is likely to break down (see below for example data). Other problems include the need for charge derivatives for analytical gradient calculations (ignored in the published DH2 version, but on the order of tenths of a kilocalorie per mole even for small systems in the case of strong H bonds), a problematic repulsive term including a distance cutoff to prevent problems with the optimization of strong H bonds, and a partially adjusted dispersion correction (with a modified C_6 for sp^3 carbon and a changed van der Waals radius for hydrogen) that could now profit from recent developments of dispersion corrections with system-dependent C6 coefficients (something not addressed in this paper, but under development). A final issue is the long-range behavior of previously published hydrogen-bonding corrections: As it is not generally clear what the exact long-range behavior should be (among other reasons because semiempirical methods already account to some extent for hydrogen-bonding interactions), we think the most reasonable thing to do is to design the correction to be of a rather short-ranged nature. This seems to be the safest way to go in the sense that no correction is safer than a correction from a huge sum of very tiny and most likely wrong contributions.

Shortly before we finished our manuscript, Foster and Sohlberg published another hydrogen-bonding correction scheme for the SE method AM1, termed FS1,[35] which they compare to PM6-DH but unfortunately not to the likewise earlier published DH2 scheme. FS1 has the same basic outline as PM6-DH1, with the repulsive term replaced by a damping function (as previously suggested as an alternative approach for DH2[34]) and the bond-type parametrization replaced with four general, fitted parameters, which somewhat spoils the accuracy but partly solves the problem with electronic structure changes. The authors nevertheless have to admit that a safe treatment of such changes would need a further modified (effectively doubled) version of their ansatz and, therefore, do not recommend using their published version in such cases. As we have tried a doubling of terms for DH2 before publishing it, we believe to have good reasons to question that such a modification will not further diminish the accuracy of the FS1 scheme. Independently of this issue, we consider the FS1 scheme to be a first-generation hydrogen-bonding correction, because it does not take the complete geometric information into account (and accordingly suffers from all of the related problems), as discussed above when comparing DH2 to the earlier PM6-DH1.

Third-Generation Correction. The third-generation correction (eq 3) does not make the assumption of a specific acceptor/hydrogen/donor binding situation but instead takes the hydrogen bond as a charge-independent atom−atom term between two atoms capable of serving as an acceptor or donor part (e.g., O, N), weighted by a function that accounts for the steric arrangement of the two fragments to each other and the preferably favorable positioning of a H atom somewhere between them (a definition of coordinates analogous to the above, with A and B being the two possible acceptor/donor atoms and C_A and C_B the corresponding hydrogen-bonding correction parameters from Table 1 for semiempirical and Table 2 for force field methods), multiplied with a damping function to correct the short- and long-range behaviors:

Table 1. Hydrogen-Bonding Correction Parameters $C_{element}$ for Semiempirical QM Methods

element	OM3	PM6	AM1	DFTB
N	−0.05	−0.16	−0.29	−0.21
O	−0.07	−0.12	−0.29	−0.08

Table 2. Hydrogen-Bonding Correction Parameters $C_{element}$ for Force Field Methods

element	MM2*	MM3*	AMBER*	OPLS*	OPLSAA	MMFF94
N	−0.64	−0.63	−0.21	−0.24	−0.25	−0.21
O	−0.08	−0.17	−0.03	−0.00	−0.00	−0.05

$$E_{\mathrm{H-bond}} = \frac{C_{\mathrm{AB}}}{r_{\mathrm{AB}}^2} \cdot f_{\mathrm{geom}} \times f_{\mathrm{damp}} \tag{3}$$

$$f_{\mathrm{geom}} = \cos(\theta_{\mathrm{A}})^2 \times \cos(\phi_{\mathrm{A}})^2 \times \cos(\psi_{\mathrm{A}})^2 \times \cos(\phi_{\mathrm{B}})^2 \times \cos(\psi_{\mathrm{B}})^2 \times f_{\mathrm{bond}} \tag{4}$$

$$f_{\mathrm{bond}} = 1 - \frac{1}{1 + \exp[-60(r_{\mathrm{XH}}/1.2 - 1)]} \tag{5}$$

$$f_{\mathrm{damp}} = \left(\frac{1}{1 + \exp[-100(r_{\mathrm{AB}}/2.4 - 1)]}\right) \times \left(1 - \frac{1}{1 + \exp[-10(r_{\mathrm{AB}}/7.0 - 1)]}\right) \tag{6}$$

$$C_{\mathrm{AB}} = \frac{C_{\mathrm{A}} + C_{\mathrm{B}}}{2} \tag{7}$$

The damping functions can be chosen as a "safe bet", so that no fitting is necessary for them (albeit the long-range cutoff could in principle be taken as a fit parameter, e.g., if it turns out that the structures of very large molecules are found to be too dense): the f_{damp} function is switched on between a donor−acceptor distance of 2.3 and 2.5 Å (safe choice for the assumption of no H bonds below 2.5 Å) and slowly switched off between 3.5 and 10.5 Å (safe choice for the assumption of full H-bond strength up to 3.5 Å and no strength anymore at three times this distance). Figure 2 shows this damping function and a resulting example energy profile for the overall correction. The f_{bond} function brings the correction to zero if the hydrogen wanders away too far from both electronegative atoms (with r_{XH} being the smaller one of the two distances r_{AH} and r_{BH}): It is switched off between 1.15 and 1.25 Å (safe choice for the assumption of a maximum distance of 1.15 for a covalent hydrogen bond).

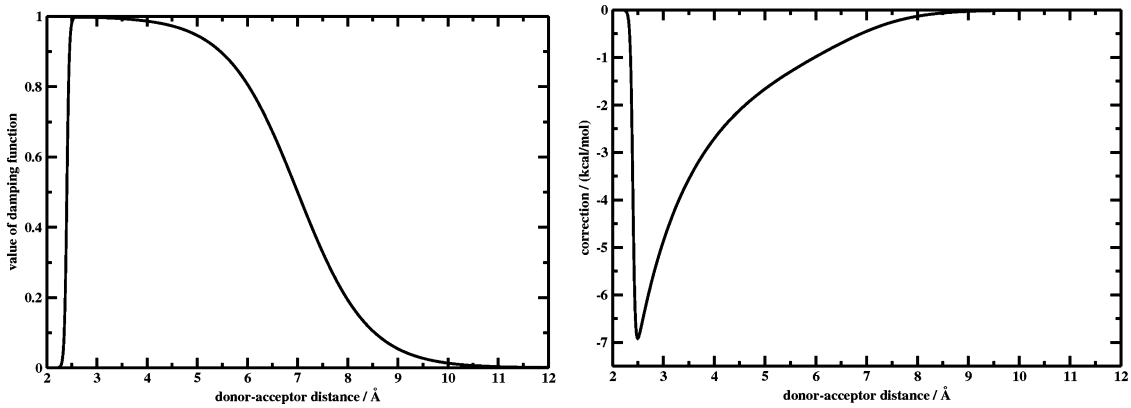


Figure 2. The f_{damp} function (left) and an example for the overall correction energy (right).

Table 3. Comparison of First-, Second-, and Third-Generation Hydrogen-Bonding Correction Schemes

generation	scheme	transferability	generality[a]	full geometric information[b]	safe long-range behavior[c]	no charges used[d]	number of fitted parameters
1	DH1	PM6	no	no	no	no	24
	FS1	AM1	yes/no[e]	no	no[f]	no	4
2	DH2	SE methods	no	yes	no	no	8
3	DH+	SE and FF methods	yes	yes	yes	yes	2

[a] Robust scheme, does not break down for electronic structure changes, can be generally included in parameter fits for new semiempirical QM (SE) and force field (FF) methods. [b] Uses full geometric information, not just a cosine term, which is likely to lead to problems with larger systems (see Figure 1). [c] Shows safe long-range behavior by avoiding huge sums of very small (and likely wrong) contributions. [d] Allows for affordable analytical gradients, as no gradients with respect to charges are required. [e] Usage for proton transfer "effectively possible" in a suggested, modified (doubled) version, but not recommended by the authors of AM1-FS1. [f] Damping function with bad long-range behavior (significantly above zero at long distances).

The (torsion) angles of the f_{geom} function are defined similarly to those of the DH2 correction,[34] with ϕ and ψ now symmetrically used for both the donor and acceptor atoms. This ansatz is not a doubling or "double-potential" version of the DH2 correction (which was tested by the author before the publication of DH2 but found to be quite problematic): the important difference is the change from the use of the hydrogen−acceptor distance (with its requirement of a hydrogen−donor bond definition) to the core−core interaction picture that results from using the donor−acceptor distance instead of the hydrogen−acceptor distance. Through this change, the implicit hydrogen−donor bond definition (also still present in the recently published AM1-FS1 method, see above) can be avoided. The target angles can nevertheless be kept as for the second-generation correction,[34] which of course has to be the case for "text-book" ideal values.

This way, the new scheme accounts for the major drawback of the (first- and) second-generation correction, i.e., the problem of a substantial change to the hydrogen bond, but several additional benefits are gained as side effects: while keeping the high accuracy of the "DH2" scheme, the number of fitted parameters can be reduced from eight to two. As charges are no longer used, no charge-derivative terms are needed for the analytical gradient. The repulsive term can be replaced by a damping function, and cutoff distances are no longer needed for an accurate description of nonequilibrium structures. (In the development of DH2, an unphysical short-distance cutoff was introduced to avoid problems with very strong hydrogen bonds where the correction was much too high, because strong partial charges and short H-bond distances both increase the value of the DH2 correction.) At the same time, the damping function greatly improves the long-range behavior in the sense that we think it to be preferable to have no (rather than very likely wrong) long-range contributions from hydrogen-bonding corrections. Finally, we found that the new scheme is also well suited for the application to force field methods. This is of great importance as it was recently shown how strongly common force fields underestimate hydrogen-bonding interactions (while they actually perform very well for dispersion interactions)[20] and a possible improvement, e.g., of water models, could have major impact for biomolecular modeling in general. We should mention though that while our straightforward implementation of the third-generation correction is about 2 orders of magnitude faster than the underlying semiempirical methods for midsized proteins (as is "DH2"), the application to force fields will require a more sophisticated approach to avoid slowing down the force field calculations by about 1 order of magnitude.

The new correction was parametrized on the hydrogen-bonded complexes of the S26 + S22x4 set for the AM1, PM6, OM3, and SCC-DFTB methods (all enhanced with standard dispersion corrections, see Table 1) and of the S22 set only for several FF methods (see Table 2). Optimization of the parameters with respect to the mean unsigned error (MUE) and the root-mean-square error (RMSE) over all reactions led to nearly identical parameter values; the final parameters are taken from the MUE optimizations. We call our approach "H+", to indicate the conceptual difference of "DH+" from the first- and second-generation "DHn" approach. Table 3 summarizes the developments from the first to the second and third generation H-bonding corrections explained in this section.

Table 4. Results for the H-Bonded Complexes of the S26 Set[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−1.75	−0.66	−0.36	−2.82	0.03	0.13	−5.58	0.25	0.81	−2.86	0.14	−0.11
MUE	1.75	0.66	0.62	2.82	0.19	0.66	5.58	0.73	2.28	2.86	0.88	1.01
RMSE	2.22	0.96	0.84	3.56	0.27	0.88	7.57	0.95	2.60	3.15	1.01	1.20
Δ	5.16	2.35	3.10	6.20	0.92	3.18	17.30	3.33	8.58	4.88	2.93	3.69

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] AM1-D refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 5. Results for the H-Bonded Complexes of the S26 Set, Optimized with Each Method[a]

	OM3- D[b]	-DH 2[b]	-DH+[b]	PM6-D	-DH2	-DH+	AM1-D[c]	-DH2	DH+
MSE	4.74	5.28	4.48	−2.63	0.68	0.45	−3.68	1.16	2.31
MUE	5.18	6.18	5.03	2.63	1.10	0.75	3.68	1.21	2.42
RMSE	11.52	12.17	7.70	3.20	1.56	0.91	5.02	1.56	2.84
Δ	38.83	40.56	20.20	4.94	5.36	2.85	10.45	3.43	5.80

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Already, the OM3 method itself (without D or H corrections) has a serious problem with the very strongly bound formic acid dimer, which is the main reason why the errors are much larger than for the other examples. [c] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 6. Results for the H-Bonded Complexes of the S26 and S22x4 Sets[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	1.25	0.21	−0.02	2.35	0.01	−0.39	4.91	0.27	−0.97	2.83	0.33	0.22
MUE	1.45	0.91	1.03	2.35	0.24	0.81	4.91	1.42	2.88	2.83	0.74	0.80
RMSE	2.05	1.36	1.46	3.20	0.34	1.00	7.76	2.55	3.54	3.33	0.89	1.01
Δ	7.92	6.30	7.45	7.88	1.65	4.14	26.18	14.38	15.46	8.12	3.19	4.51

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 7. Results for the 105 Small, H-Bonded Complexes of the PM6-DH1 Fit Set[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−0.88	−0.51	0.03	−1.66	−0.43	0.46	−2.55	−0.12	1.85	−2.33	−0.40	−0.14
MUE	0.91	0.66	0.46	1.77	1.15	1.21	2.71	1.59	2.40	2.36	0.85	1.07
RMSE	1.14	0.86	0.59	2.35	1.54	1.44	4.04	2.12	2.87	2.79	1.06	1.44
Δ	6.52	4.48	3.71	9.61	7.37	6.18	22.64	12.14	13.08	10.47	5.15	8.64

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

3. Computational Details

Semiempirical PM6 and AM1 calculations applying the MOZYME algorithm were done with MOPAC2009,[33] OM3 calculations with MNDO2005, and SCC-DFTB calculations with DFTB+.[36] AM1-D* refers to standard AM1[16] with a standard Jurecka-type[18] empirical dispersion correction (see ref 34 for details), not AM1-D, which is additionally based on a refit of 18 AM1 parameters. B3-LYP[37,38] DFT calculations with empirical dispersion corrections of the Jurecka type[18] were done with Turbomole 5.9[39] using TZVP[40] and QZVP[41] Gaussian AO basis sets and the RI approximation[42,43] for two-electron integrals.

Energies and analytical gradients for our new "H+" hydrogen-bonding correction are implemented as a standalone program that is freely available from the author upon request. Preparations to make the correction available within the open source FF code GROMACS are underway.

4. Results and Discussion

Tables 4−9 show results of OM3, PM6, AM1, and SCC-DFTB (shortened to "DFTB" in the tables) calculations with dispersion and second- and third-generation hydrogen-bonding corrections for the hydrogen-bonded complexes of the S26[44] (Table 4, again in Table 5 with structures optimized at each level of theory) and S26 + S22x4[34] (Table 6) benchmark sets; the PM6-DH1 training set of 105 small hydrogen-bonded complexes[32] (Table 7); the 37 noncharged, H-bonded DNA base pair complexes from the JSCH2005 set[15] (Table 8); and the 13 noncharged, H-bonded peptide structures from the JSCH2005 test set (Table 9). (We do not supply DFTB data in Table 5 because our interface does not allow for DFTB geometry optimizations with DH+ yet.) The geometries of these benchmarks are optimized at the MP2/cc-pVTZ level or higher (S22, S26, S22x4, PM6-DH1 training set, JSCH2005 partly) or represent experimental data (JSCH2005 partly); see the references given above for details.

Table 8. Results for the JSCH2005 H-Bonded DNA Base Pairs[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	−2.41	−0.81	−0.49	−6.10	−0.54	−0.87	−10.16	0.11	−0.05	−5.49	2.64	0.90
MUE	2.50	1.22	1.02	6.10	1.76	1.35	10.16	2.29	1.68	5.49	3.06	1.64
RMSE	2.79	1.44	1.29	6.30	2.23	1.59	10.91	2.87	2.40	5.80	3.48	1.97
Δ	6.81	4.52	5.74	7.67	7.94	5.94	16.31	12.46	11.78	6.85	8.53	7.20

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 9. Results for the JSCH2005 H-Bonded Peptides[a]

	OM3-D	-DH2	-DH+	PM6-D	-DH2	-DH+	AM1-D[b]	-DH2	-DH+	DFTB-D	-DH2	-DH+
MSE	0.33	0.36	0.36	−0.07	−0.00	−0.00	1.37	1.45	1.50	−0.84	−0.75	−0.76
MUE	0.60	0.62	0.62	0.65	0.69	0.68	1.49	1.56	1.60	0.92	0.83	0.85
RMSE	0.80	0.81	0.82	0.85	0.88	0.86	1.94	2.03	2.11	1.07	0.98	0.99
Δ	2.81	2.79	2.80	3.45	3.42	3.40	4.30	4.27	4.60	2.91	2.84	2.85

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole. [b] Refers to standard AM1 with a standard empirical dispersion correction, unlike AM1-D.

Table 10. Results for the H-Bonded Complexes of the S22 Set[a]

	MM2*	-H+	MM3*	-H+	AMBER*	-H+	OPLS*	-H+	OPLSAA	-H+	MMFF94	-H+	B3LYP-D/TZVP
MSE	−8.77	−0.63	−10.90	−1.37	−3.47	−0.74	−2.76	−0.20	−3.27	−0.61	−3.01	0.05	0.74
MUE	8.77	2.12	10.90	3.61	3.93	2.63	3.30	2.03	3.55	1.73	3.15	0.84	0.74
RMSE	10.68	2.70	13.13	5.00	5.29	3.60	4.42	2.57	4.49	2.53	3.88	1.19	0.84
Δ	−17.18	−8.75	−19.07	−16.17	−11.65	−9.83	−9.10	−8.09	−8.46	−7.70	−7.57	−4.30	1.17

[a] Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are presented. All values are in kilocalories per mole.

Structures and reference energies for these test sets can be obtained online from the Benchmark Energy and Geometry DataBase BEGDB, see http://www.begdb.com. Mean signed (MSE), mean unsigned (MUEs), and root-mean-square errors (RMSE) as well as the maximum error span (Δ) with respect to the benchmark CCSD(T)/CBS interaction energies are given in kilocalories per mole. Table 10 gives the same statistical error measures for the hydrogen-bonded complexes of the S22 set, this time for a number of force field methods without and with augmentation by our third-generation hydrogen-bonding correction. Force field interaction energies for the "frozen" geometries of the S22 and S22x4 sets were kindly provided by the authors of ref 20, from their extensive study on the performance of force field methods for noncovalent interactions. The force fields are the MacroModel implementations of MM2*,[45] MM3*,[46] AMBER*,[47−49] and OPLS*[50] and native versions of OPLSAA[51] and MMFF94.[52] Further details can be found in the original publication.[20]

Perusing Tables 4−9, the following conclusions can be drawn: All six tables illustrate that even dispersion-corrected semiempirical QM methods perform quite badly for hydrogen-bonding interactions (a known issue, see Introduction). While OM3 is doing rather well, AM1 especially gives large errors for H-bond interaction energies. Tables 4 and 6−9 also show that the inclusion of the second-generation "H2" correction (in combination with standard dispersion corrections such as "DH2") consistently improves the accuracy of all methods, but unfortunately DH2 suffers from several conceptual problems (as explained in the theory section above). Our new "H+" correction (in combination with standard dispersion corrections such as "DH+") is able to reach the same overall accuracy as the DH2 correction, while it avoids all of the conceptual problems connected with the DH2 ansatz: Tables 4 and 7−9 show that, e.g., MUEs are improved by a factor of 1.5 to 3 for all sets (with significantly strong H-bonding interactions, unlike the peptide set in Table 9) and methods.

That this is also the case for nonequilibrium structures can be seen in Table 6, and that this conclusion still holds for structures optimized at the corresponding level of theory is shown in Table 5. [The used benchmark sets are designed to be used with the given benchmark geometries, because the goal is a correct energetic description within the correct geometrical arrangement. For a comparison of interaction energies of structures optimized at different levels, it is necessary to carefully check and compare all final geometries and resulting energetic effects. To allow for comparison with earlier work, we present this data here for the S22 set—where for both DH2 and DH+ no substantial change of binding motifs occurs—but stick with the intended use of the benchmark sets in the other cases.]

While the overall accuracies of DH2 and DH+ are very similar, DH2 seems to do better for AM1, while DH+ seems to be better suited for OM3 (see Tables 7 and 8). The performance for DFTB is better with DH2 for the DH1 training set (Table 7) but better with DH+ for the JSCH2005 set, presumably because the latter one includes multiple hydrogen bonds. In addition, while DH2 works exceptionally well for the S26 and S26+S22x4 fit sets (as three parameters for each method were added to achieve exactly this goal), this additional gain in accuracy is not transferred to systems beyond the fit sets, e.g., the diverse hydrogen-bonded structures of the PM6-DH1 set (Table 7), where DH+ is at the same level, and especially not the DNA base pairs (Table

Table 11. Results for Several Methods for the Hydrogen-Bonded Complexes of the S26 Set (S22 Set for Force Field Methods)[a]

methods	MUE	average error per H bond
MM2*	8.77	5.0
MM2*-H+	2.12	1.2
MMFF94	3.15	1.8
MMFF94-H+	0.84	0.5
SCC-DFTB-D	1.75	1.2
SCC-DFTB-DH+	1.01	0.7
PM6-D	2.82	1.9
PM6-DH+	0.66	0.4
OM3-D	1.75	1.2
OM3-DH+	0.62	0.4
B3LYP-D	0.74	0.5

[a] Mean unsigned errors (MUEs) and average errors per H bond with respect to the benchmark CCSD(T)/CBS interaction energies are presented. DFT methods with TZVP basis sets. All values in kiloclaories per mole.

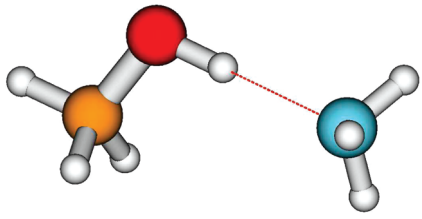


Figure 3. Simple Model System for Proton Transfer.

8), where DH+ (two fit parameters) even outperforms DH2 (eight fit parameters). The peptide systems in Table 9 were added to show that the DH+ correction (as the DH2 correction) does not worsen the energies of complexes with very weak hydrogen bonds by introducing unphysical contributions.

Table 10 shows that the overall good performance of the "H+" correction is also transferable to FF methods, with MUEs for the hydrogen-bonded systems of the S22 set nearly comparable to much more sophisticated computational approaches. Tests additionally including the nonequilibrium structures from the S22x4 set showed a reduction of the mean unsigned error (MUE) from 6.5 to 2.5 kcal/mol for MM2* and from 2.1 to 0.9 kcal/mol for MMFF94, quite comparable to the gain shown in Table 10 for equilibrium structures only. More thorough studies are in preparation but are beyond the scope of this work.

For a large number of investigated cases, the new "DH+" correction reaches the accuracy of DFT-D approaches, while being several orders of magnitude faster and now free of the conceptual problems of the older DH2 correction scheme. This is again summarized in Table 11 where different force field and semiempirical QM methods as well as a "standard" DFT-D approach are compared for the MUE and average error per H bond over the hydrogen-bonded systems of the S26 set (s22 set for force field methods).

To illustrate the practical applicability of DH+ to model proton transfer reactions, we have looked at the simple model reaction of methanol and ammonia visualized in Figure 3. Starting from a MP2/TZVP optimized structure, all coordinates are kept frozen except the O−H distance, which is varied between 1.0 and 1.9 Å in steps of 0.025 Å, corresponding to a proton transfer from methanol to ammonia.

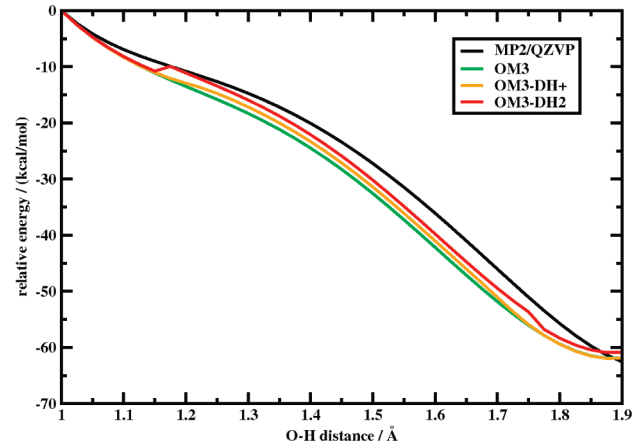


Figure 4. OM3(-DH2/-DH+) proton transfer energetics for the model system from Figure 3, with MP2/QZVP reference data (energies at 1 Å taken as a reference point).

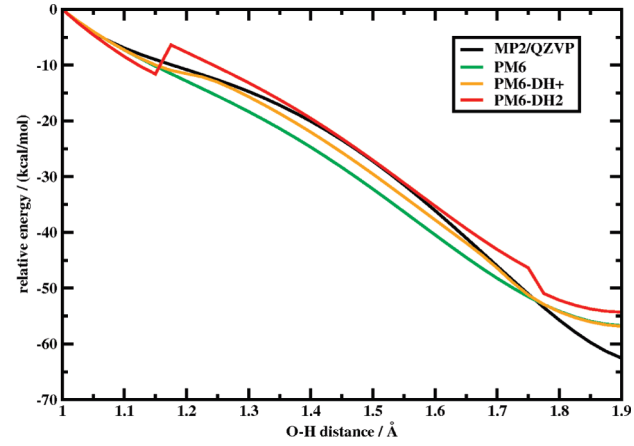


Figure 5. PM6(-DH2/-DH+) proton transfer energetics for the model system from Figure 3, with MP2/QZVP reference data (energies at 1 Å taken as a reference point).

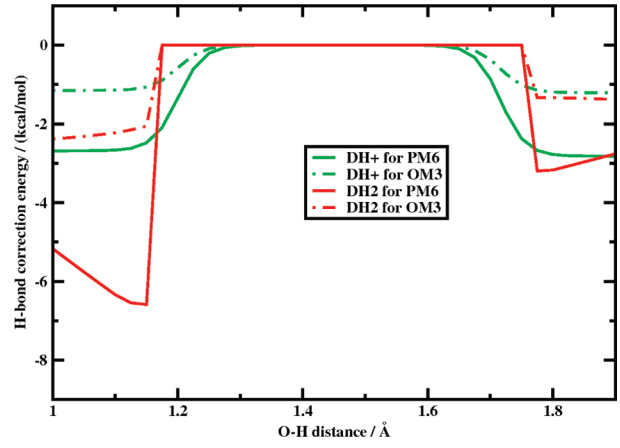


Figure 6. DH2 and DH+ hydrogen-bonding correction energies for the model system from Figure 3 (with parametrization corresponding to the OM3 and PM6 methods).

Figures 4 and 5 show the resulting proton transfer energetics for OM3 and PM6 without and with the DH2 and DH+ corrections, illustrating that in opposition to DH2, DH+ does not break down in the case of proton transfer. Apart from that, the impact of both corrections is small in comparison with the overall reaction energy, with a cor-

Table 12. Selected Hydrogen-Bond Interaction Energies from Linear, Hydrogen-Bonded Formamide Chains, As Well As B3LYP/D95**, MP2/TZVP, and MP2/QZVP Reference Data[a]

	OM3	OM3-DH2	OM3-DH+	PM6	PM6-DH2	PM6-DH+	B3LYP/D95**[b]	MP2/TZVP	MP2/QZVP
dimer	−5.13	−6.31	−6.57	−5.36	−6.71	−7.81	−7.31	−6.65	−6.47
hexamer terminal	−6.84	−8.27	−8.27	−7.17	−8.82	−9.56	−10.04	−8.66	
hexamer central	−9.05	−10.82	−10.39	−9.27	−11.33	−11.45	−13.20	−11.26	

[a] All values are in kilocalories per mole. [b] From ref 53.

respondingly small, indirect effect on the "barrier" height through the energetic lowering of reactants and products. Figure 6 shows a direct comparison of the DH2 and DH+ correction energies for our simple model reaction (using the corresponding the OM3 and PM6 parameters), illustrating the problems of DH2 and the conceptual improvement of DH+ in more detail.

Also of some importance is the performance of DH+ for hydrogen-bonding cooperativity, because such a type of cooperativity has been shown to exist in β sheets, which makes it important for the accurate modeling of large proteins, where the stability of secondary structures might be influenced.[53] Table 12 shows interaction energies for selected hydrogen bonds in linear formamide chains of lengths two and six, a model system taken from the work of Dannenberg and co-workers.[53,54] Besides OM3(-DH2/DH+) and PM6(-DH2/DH+) data, DFT and MP2 reference values are given. (Following ref 53, interaction energies are calculated by simple subtraction; e.g., the energy of the terminal H-bond in the hexamer is taken to be the energy of the hexamer less the combined energies of the pentamer and the monomer.)

First of all, Table 12 illustrates the strong cooperative nature of the H-bond interactions (emphasized already in the above-mentioned work by Dannenberg); i.e., the central interaction in the hexamer is predicted to be nearly two times as much as the dimer interaction. Comparing OM3 and PM6 with MP2, it looks as if semiempirical methods seem to be rather well capable of modeling such hydrogen-bond cooperativity effects, especially also concerning the ratio of interaction strengths (with a factor of 1.6 and 1.8 between the dimer and the central hexamer H-bond strength for all approaches). DH2 and DH+ show again a very similar performance, systematically improving the underlying SE methods, with DH2 being slightly more advantageous at least for PM6, presumably because DH2 has one parameter specially dedicated to amide interactions and fitted to the formamide dimer. Overall, empirical correction schemes seem to also work surprisingly well with semiempirical QM methods for hydrogen-bonding cooperativity effects.

5. Conclusions

This work presents a further improved, "third-generation" hydrogen-bonding correction scheme that can now be generally included in parameter fits of semiempirical QM and force field methods, as it does not suffer any longer from several conceptual limitations of previous approaches in this direction: hydrogen bonds are now treated as an interaction term between electronegative acceptor and donor atoms, "weighted" by a function of the positioning of H atoms between them. This way, the new correction scheme improves over existing ones with regard to the following issues: Electronic structure change (e.g., proton transfer) becomes generally possible; a safe long-range behavior is enforced; exact analytical gradients are affordable;, transferability to force field methods is achieved, and straightforward extendability for other hydrogen- (and halogen-)bonding types is given; and the same (high) overall accuracy can be achieved with significantly less parametrization. Our new correction scheme consistently improves the accuracy of the semiempirical QM methods PM6, AM1, OM3, and SCC-DFTB as well as the MM2*, MM3*, AMBER*, OPLS*, OPLSAA, and MMFF94 force field methods for several benchmark sets of hydrogen-bonding interactions by up to 1 order of magnitude at the cost of a force-field-type calculation.

Acknowledgment. The author would like to thank Pavel Hobza for introducing the author to hydrogen-bonding correction schemes and Jonathan Goodman for kindly supplying force field data and advice. This work was supported by Grant LPDS-2009-19 from the German National Academy of Science Leopoldina.

Supporting Information Available: Geometries for the proton transfer and hydrogen-bond cooperativity model systems. This material is available free of charge via the Internet at http://pubs.acs.org

References

(1) Jorgensen, W. L. *Acc. Chem. Res.* **2009**, *42*, 724.

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

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

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

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

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

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

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

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

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

(11) 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(26) 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.

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

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

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

(30) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.

(31) Stewart, J. P. *J. Mol. Model.* **2009**, *15*, 765.

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

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

(34) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.

(35) Foster, M. E.; Sohlberg, K. *J. Chem. Theory Comput.* **2010**, *6*, 2153.

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

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

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

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

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

(41) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119*, 12753.

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

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

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

(45) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(46) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(47) McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7743.

(48) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(49) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

(50) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(51) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(52) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.

(53) Kobko, N.; Paraskevas, L.; del Rio, E.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2001**, *123*, 4348.

(54) Kobko, N.; Dannenberg, J. J. *J. Phys. Chem. A* **2003**, *107*, 10389.

CT100408B