JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods

Jan Řezáč*[,†] and Pavel Hobza[†,‡]

[†]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague, Czech Republic

[‡]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacky University, 771 46 Olomouc, Czech Republic

Supporting Information

ABSTRACT: Semiempirical quantum mechanical methods with corrections for noncovalent interactions, namely dispersion and hydrogen bonds, reach an accuracy comparable to much more expensive methods while being applicable to very large systems (up to 10 000 atoms). These corrections have been successfully applied in computer-assisted drug design, where they significantly improve the correlation with the experimental data. Despite these successes, there are still several unresolved issues that limit the applicability of these methods. We introduce a new generation of both hydrogen-bonding and dispersion corrections that address these problems, make the method more robust, and improve its accuracy. The hydrogen-bonding correction has been completely redesigned and for the first time can be used for geometry optimization and molecular-dynamics simulations without any limitations, as it and its derivatives have a smooth potential energy surface. The form of this correction is simpler than its predecessors, while the accuracy has been improved. For the dispersion correction, we adopt the latest developments in DFT-D, using the D3 formalism by Grimme. The new corrections have been parametrized on a large set of benchmark data including nonequilibrium geometries, the S66x8 data set. As a result, the newly developed D3H4 correction can accurately describe a wider range of interactions. We have parametrized this correction for the PM6, RM1, OM3, PM3, AM1, and SCC-DFTB methods.

INTRODUCTION

Until very recently, semiempirical quantum mechanical (SQM) methods were developed mainly to reproduce the thermochemical properties of molecules. Because of the many approximations used and since no greater attention has been paid to them, noncovalent interactions have not been described well by the SQM methods. On the other hand, the SQM methods are the most efficient methods that still use a quantum mechanical description of the system, which provides them with several advantages over fully empirical molecular mechanics (MM). For one, the SQM methods are able to describe quantum effects that are not covered by molecular mechanics; another advantage is that SQM methods can be applied to any molecule without previous parametrization. The SQM methods can be derived from the ab initio Hartree−Fock (HF) method by introducing further approximations, such as reduction of the basis set to the absolute minimum and neglect or simplification of a large part of the integrals. To compensate for these approximations, additional empirical terms are added. The parameters, both in the integrals and in the additional corrections, are either derived directly from the experimental data or optimized to reproduce them. Although not in the mainstream, the SQM methods are still evolving, and the recently published PM6,[1] RM1,[2] OM-x,[3] and other methods have brought substantial improvements over their predecessors. When they are combined with linear scaling algorithms, such as the localized orbital method MOZYME,[4] it is possible to calculate routinely whole proteins at the SQM level.

Two types of interactions that are difficult to describe using the SQM methods are London dispersion and hydrogen bonds (H-bonds), both of which are common in the studied systems and crucial for obtaining accurate results. The London dispersion can be described explicitly only by methods that account for electron correlation. The SQM methods can include part of this interaction by other means, through the parameters and core−core potentials, but a major part of the dispersion is still missing. In methods where the dispersion is missing completely, such as in the HF or density functional theory (DFT), it can be easily added as an a posteriori empirical correction (the resulting corrected DFT is referred to as DFT-D). A pairwise potential based on the physically sound c_6/r^6 formula (where c_6 is a coefficient determining the strength of the interaction and r is the interatomic distance) scaled at short distances by a damping function has proven to be a very effective solution. It is widely used in DFT and has also been applied to the semiempirical methods,[3,5] in some cases in conjunction with a partial reparameterization of the SQM method itself.

However, none of the resulting dispersion-corrected SQM methods had been accurate enough to provide a quantitative description of all of the types of noncovalent interactions until corrections for both dispersion and hydrogen bonding were applied. We were the first to develop a H-bond correction for the SQM method,[6] namely, PM6. In this approach, the correction energy is a function of the hydrogen-bond distance and angle, and of partial charges of the atoms involved in the H-bond. In combination with a parametrization of a dispersion correction used previously in DFT, the resulting method (PM6-DH)

Received: October 25, 2011
Published: December 22, 2011

achieved an accuracy of less than 1 kcal/mol in small model complexes.

The second generation of the correction[7] (DH2) was developed to solve the problems of the first version. The dispersion correction was modified in order to avoid double counting of the dispersion energy already described by the underlying method. This was achieved by scaling the whole dispersion term and a specific scaling of the c6 coefficient for sp3-hybridized carbon. In the H-bonding correction, discontinuities of the potential were fixed, and additional geometrical parameters of the hydrogen bond were added to avoid false contributions from atoms not involved in a real H-bond and to improve the geometry of the H-bond. The DH2 correction was parametrized for use with multiple semiempirical methods, namely, PM6,[1] AM1,[8] OM3,[3] and SCC-DFTB;[9] the best results have been achieved with PM6.

The third generation of the correction[10] (DH+) addresses two problems of the DH2 correction. The first is the use of partial atomic charges from the underlying semiempirical calculation in the H-bond energy correction. The derivative of the charge with respect to the coordinates, which is expensive to calculate, enters the expression for the gradient of the correction. For practical purposes, the derivative of the charges was assumed to be zero, but this approximation cannot be used in some cases, such as in accurate optimizations or in molecular dynamics. In the DH+ correction, the charges are no longer used, and the exact gradient can be obtained easily. The second issue addressed is the fixed definition of the hydrogen donor and acceptor atoms. In the DH2 formalism, a proton transfer along a hydrogen bond exhibits a discontinuous PES. In the DH+ correction, the two potentials for both the reactant and product are switched smoothly. The dispersion correction in the DH+ is identical to the one in DH2.

It is clear from this brief review that many problems have been successfully solved during the development of the corrections for the SQM methods. On the other hand, the complexity of the H-bonding correction has grown substantially. The energy of the DH2 and DH+ corrections depends not only on the atom distances and angle of the H-bond but also on multiple other internal coordinates. Not only does this make the actual calculation complicated, the main problem is in defining these coordinates for different groups involved in the H-bond. In practical implementation, this information on all of the possible H-bonds in the system is stored in memory, making the calculation inefficient for large molecules.

Additionally, two more important points were neglected in the construction of the DH+ correction. First, the earlier versions of the H-bonding correction used the atomic charges and thus naturally described strong hydrogen bonds involving charged groups. In DH+, the same parameters are used for neutral and charged H-bonds, which leads to an underestimation of the interaction in charged systems. Second, the angular terms in both DH2 and DH+ do not have smooth first derivatives, which makes it impossible to optimize the geometry of some systems.

Despite the limitations described above, the PM6-DH2 method has been successfully applied to practical problems.[11−14] It was used for calculations of protein−ligand interactions in computational drug design, yielding much better correlation with the experimental data than an equivalent protocol based on molecular mechanics. The great potential of the corrected SQM methods in this area and other possible applications has led us to develop the corrections further.

Here, we propose the next generation of the H-bonding and dispersion corrections for semiempirical methods. In the H-bond correction, we have wanted to preserve the improvements brought by the DH+ approach, solve its poor performance in charged systems, and, importantly, simplify the form of the correction. Unlike its predecessors, the new correction has not only a smooth potential energy surface but also its first and second derivatives. Another important feature is that the correction potential is strictly local and does not have to be evaluated for more distant potential H-bonds. This makes the computational expense grow only linearly with the size of the calculated system. Finally, our goal is to improve the accuracy, or at least keep it at the level of the previous, more complex approaches. These developments are summarized in Table 1, which lists the most important features of the correction in the DH2, DH+, and D3H4 versions.

Table 1. A Comparison of the Hydrogen-Bonding Correction in the DH2, DH+, and Newly Introduced D3H4 Approaches

	H2	H+	H4
exact gradient	NO	YES	YES
proton transfer	NO	YES	YES
accurate for charged systems	YES	NO	YES
smooth energy derivatives	NO	NO	YES
coordinates per H-bond (torsions)	4 (2)	7 (4)	3 (0)

Table 2. A Comparison of the Dispersion Correction in the DH2, DH+, and Newly Introduced D3H4 Approaches

	D2, D+	D3
parameters for elements	18	94
valence-dependent parameters	NO	YES
parameters	element-wise	pairwise

We have also updated the dispersion correction, adopting the latest advances in the DFT-D methods.[15−17] We have based our dispersion correction on the DFT-D3 method.[17] The improvement of the accuracy is not large, but the new approach has other important benefits (see Table 2). First, it uses a large set of atomic parameters consistently constructed for all of the elements up to plutonium. This makes it a useful complement to the PM6 method, which can treat 70 elements; for many of these, the parameters for the earlier dispersion correction were missing. The DFT-D3 correction uses different parameters for the possible valence states of the atoms and switches between them smoothly.

In this work, we have focused mainly on PM6, which we found to be the most accurate SQM method for the description of biomolecular systems. Additionally, we report the parametrization of the correction for many other semiempirical methods we have used in our previous work, namely, AM1[8] and OM3[3] and the self-consistent charge density-functional tight-binding[9] (SCC-DFTB) method. In this study, we parametrize the corrections for two more methods, PM3[18] and the more recent RM1.[2]

The dispersion and hydrogen bonding corrections described here are close to the accuracy limit that could be achieved with *a posteriori* corrected semiempirical methods. The D3H4 approach also solves all of the issues we encountered in the previous generations of the corrections. Therefore, we consider the D3H4 corrections to be a final version that can be recommended for general use.

It should also be noted that empirical corrections are not the only way to achieve a better description of hydrogen bonds in semiempirical methods. There is no fundamental reason that would

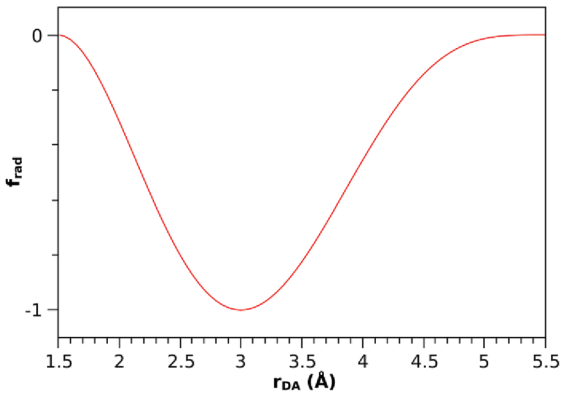

Figure 1. Radial potential of the hydrogen bond correction. This polynomial function is scaled by a coefficient specific for each combination of donor and acceptor elements.

prevent SQM methods from describing hydrogen bonds quantum mechanically. This had been discussed in the literature,[19,20] and significant improvements have been achieved when polarization functions were added to hydrogen atoms. SINDO1[21] was the first SQM method using this approach. The recently introduced PMO method[22] yields very promising results, but its parametrization is limited to hydrogen and oxygen only.

H-BOND CORRECTION

The correction potential for each potential hydrogen bond is constructed from the radial part f_{rad}, which determines the strength of the correction from the donor−acceptor distance (r_{DA}) scaled by the angular term (f_{ang}), and the proton transfer term (f_{PT}), which depends on the position of the hydrogen between the donor and acceptor. In the case of charged groups, additional scaling by the f_{charge} term is applied to make the correction stronger, and when water acts as a hydrogen donor, further scaling f_{wat} is applied. The complete expression is

$$E_{HB} = c \times f_{rad}(r_{DA}) \times f_{ang}(\alpha_{DHA}) \times f_{PT}(r_{DH}, r_{AH}) \times f_{charge} \times f_{wat} \quad (1)$$

where c is the parameter determining the strength of the correction, α_{DHA} is the donor-hydrogen-acceptor angle (defined as zero in the linear arrangement), and r_{DH} and r_{AH} are the distances between the hydrogen and the donor and acceptor.

Radial Potential. The shape of this potential mimics the difference between the dissociation curve of the H-bond calculated with the corrected method and a reference. This difference does not vanish even at larger distances (6−10 Å). In the previous versions of the H-bond correction, a $1/r^x$ form damped at short distances was used. The unlimited range of such a correction led to problems in condensed systems with many potential H-bonds in this range. Small contributions that are not a real H-bond added up to an erroneous stabilization. This had been addressed by the addition of further criteria for the identification of true H-bonds based on additional internal coordinates. This approach works well but makes the calculation rather complex. Another solution, developed in this work, is to make the correction potential more short-ranged. We found that H-bonds with a donor−acceptor distance larger than 5.5 Å can be left uncorrected without a loss of accuracy, and we use this cutoff radius in our correction. The correction potential (Figure 1) is a polynomial determined by the following points: It has a minimum at the average H-bond distance, $r_{DA,0}$ = 3.0 Å. At a cutoff radius of 5.5 Å, it smoothly approaches zero. Here, the first and second derivatives are also required to be smooth. The third point defines the curvature of the potential at shorter than equilibrium distances by setting the distance where the correction approaches zero. The distance $r_{DA,min}$ = 1.5 Å was determined by fitting the dissociation curves of the training set. This region of the potential is unlikely to be visited. Therefore, we make sure that the energy surface is smooth, but we do not apply any conditions to the energy derivatives. The eight conditions described here are used to construct a seventh-order polynomial. Outside this interval, the correction function is set to zero (the correction is not calculated in a practical implementation). The obtained coefficients are listed in eq 2 in a rounded form; in a real implementation, it is necessary to use more precise values in order to avoid large errors in the result. The coefficients are provided at high precision in the Supporting Information (Table S4).

$$f_{rad}(r_{DA}) = -0.003r_{DA}^7 + 0.074r_{DA}^6 - 0.701r_{DA}^5 + 3.253r_{DA}^4 - 7.207r_{DA}^3 + 5.318r_{DA}^2 + 3.407r_{DA} - 4.685 \quad (2)$$

The depth of the minimum of this potential is determined by the only free parameter in the correction, coefficient c. The parameters obtained by fitting to reference values (as described below) are tabulated for all of the combinations of donor and acceptor elements.

Angular Term. Goniometric functions, $\cos(\alpha)$ in DH2 and $\cos(\alpha)^2$ in DH+, were used previously, as they seem to be a natural expression of the angular dependence of the potential. However, the derivative of this term has a cusp at $\alpha = 0$; it is in the linear arrangement of the H-bond (Figure 2). This makes it practically impossible to optimize a system with a linear hydrogen bond. For this reason, and in order to gain more control over the shape of the potential, we replaced the cosine with a polynomial constructed to have smooth derivatives. First, we define a polynomial switching function $f_{sw}(x)$ that smoothly changes from 0 at $x = 0$ to 1 at $x = 1$, having first and second derivatives of zero at the boundaries of this interval.

$$f_{sw}(x) = -20.0x^7 + 70.0x^6 - 84.0x^5 + 35.0x^4 \quad (3)$$

The angular term then uses this function to construct a potential with the desired properties in the interval of 0 to $\pi/2$:

$$f_{ang}(\alpha_{DHA}) = 1 - (f_{sw}(2\alpha_{DHA}/\pi))^2 \quad (4)$$

We have tested multiple functions with different shapes, namely, the width of the minimum, and one with a rather wide, flat maximum (eq 4) worked best (in terms of interaction energies in the model complexes).

Proton-Transfer Term. The correction described so far would work on equilibrium geometries but breaks down when the hydrogen is moved along the H-bond to the acceptor. We address this analogously to the H+ correction by scaling the correction using a switching function dependent on the donor−hydrogen distance. When the hydrogen is at a covalent distance, this function should be equal to 1. In contrast to the H+ correction, the coefficient c changes when the donor and acceptor are exchanged during the proton transfer (the donor is defined as the atom closer to the hydrogen). A smooth transition is ensured by the proton transfer switching function, which scales the whole correction to zero when this exchange occurs (at $r_{DH} = r_{AH}$). We use the polynomial switching function f_{sw} defined in eq 3. Here, it

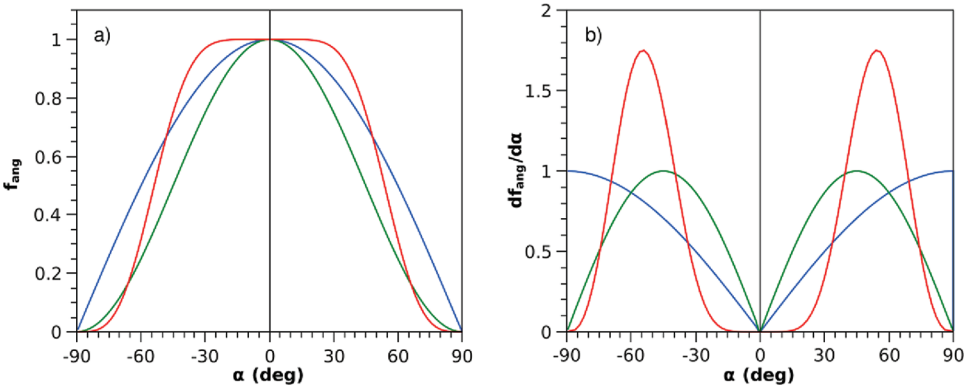

Figure 2. Angular term (a) and its derivative (b) in the DH2 (blue), DH+ (green), and D3H4 (red) hydrogen-bond corrections. The angle of zero degrees corresponds to a linear arrangement of the atoms.

switches from 1 to 0 in the interval from r_0 = 1.15 Å (∼ max. covalent bond length) to r_1 = half of the sum of the donor−hydrogen and acceptor−hydrogen distances.

$$f_{PT}(r_{DH}) = \begin{array}{ll} 1 & \text{when } r_{DH} < r_0 \\ 1 - f_{sw}((r_{DH} - r_0)/(r_1 - r_0)) & \text{in } (r_0 \ldots r_1) \end{array} \quad (5)$$

Scaling of Charged H-Bonds. When the donor or/and acceptor are charged, the hydrogen bond becomes stronger than in a neutral system. The first two generations of the correction accounted for this directly, because the atomic charges determined the strength of the correction. In the third generation, these charges were replaced by fixed parameters, which leads to substantial errors for charged H-bonds. Here, we correct this problem by additional scaling of the correction for charged groups. The scaling factors are fitted to the reference data as described below. The identification of the charged groups can be done automatically in simple cases. We implemented it for the COO− and NHR3+ groups most common in biomolecules. For calculations on the minimum geometry, this scaling can be applied easily on the basis of the atom types determined by a connectivity search. This approach is not valid in reactions, e.g., proton transfer, where atom types defined this way would change abruptly. To keep the potential continuous, we have introduced a fractional measure of atom valence as the description of the similarity to the desired atom type. For example, to identify the NHR3+ donor group, we evaluate the distance from the nitrogen (atom *i* in general) to all of the atoms in the vicinity. Each of these atoms contributes to the atom valence v_i by a factor determined by a function v_{ij} of the interatomic distance r_{ij}, which smoothly switches from 1 for covalent distance r_{cov} to 0 at 1.6 r_{cov}. The factor 1.6 has been chosen here so that two atoms bound to the same center in a tetrahedral arrangement do not affect one other. The polynomial switching function described above (eq 3) is used:

$$v_{ij}(r_{ij}) = \begin{array}{ll} 1 & \text{if } r_{ij} < r_{cov} \\ 0 & \text{if } r_{ij} > 1.6 r_{cov} \\ f_{sw}((r_{ij} - r_{cov})/(0.6 r_{cov})) & \text{otherwise} \end{array} \quad (6)$$

$$v_i = \sum_{j \neq i} v_{ij}(r_{ij}) \quad (7)$$

The scaling for NHR3+ is applied if the valence of the nitrogen v_N is between 3 and 5; the scaling factor f_{charge} is linearly dependent on the valence and peaks at the value of c_s determined for the minimum geometry when the valence reaches v_{max} = 4:

$$f_{charge} = \begin{array}{ll} 1 + (c_s - 1) \times (1 - |v_N - v_{max}|) & \text{for } v_N \text{ in } (v_{max} - 1 \ldots v_{max} + 1) \\ 1 & \text{otherwise} \end{array} \quad (8)$$

In the COO− group, the scaling factor f_{charge} is the product of the contributions from both oxygens, calculated using the same functions as above but peaking at a valence of 1.0. Such a continuous description of the atom type can be applied to any other group.

Water-Donor Scaling. In PM6, interaction energies of hydrogen bonds with a water molecule as a donor (including the water dimer) are not as underestimated as in other hydrogen bonds featuring a −OH group donor. This is probably a consequence of the special focus on water in the parametrization of PM6. Therefore, a large improvement can be achieved when a water molecule as a donor is treated separately from other oxygen donors. We define the atom type for water oxygen as a continuous property analogously to the identification of the charged groups. The scaling factor f_{wat} is equal to an optimized coefficient c_{wat}, where the oxygen has exactly two hydrogens within covalent distance and decays smoothly when the fractional valence diverges from the ideal value, as determined using eq 8.

DISPERSION CORRECTION

We adopt the formalism and atomic parameters proposed by Grimme for the DFT, the D3 dispersion correction.[17] We do not use the higher-order contributions (effectively covered by the $1/r^8$ term in D3), as they do not improve the accuracy when used with the SQM methods. The dispersion is pairwise interatomic potential:

$$E_{disp} = s_6 \sum \sum \frac{c_{6,ij}}{r_{ij}^6} f_{damp}(r_{ij}) \quad (9)$$

where $f_{damp}(r_{ij})$ is a function damping the dispersion at short distances:

$$f_{damp}(r_{ij}) = \frac{1}{1 + 6\left(\frac{r_{ij}}{s_r r_{0,ij}}\right)^{-\alpha}} \quad (10)$$

In DFT-D, it is used to prevent overbinding in the region where the exchange-correlation functional describes the interaction

dx.doi.org/10.1021/ct200751e | J. Chem. Theory Comput. 2012, 8, 141–151

well. In the SQM methods, the interaction at short range is also described by the method itself by means of parametrized core−core potentials.

There are three adjustable parameters that should be optimized for each corrected method—the global scaling of the correction s_6, the scaling of the cutoff radii in the damping function s_r, and the exponent α that determines the steepness of the damping function.

The main difference from the previously used dispersion correction lies in the parameters employed in the formulas above, the c_6 coefficients that determine the dispersion energy for each pair of atoms and the cutoff radii r_0 determining the onset of the damping function that switches off the correction at short distances. Previously, the c_6 coefficients were tabulated for each element, and pairwise $c_{6,ij}$ coefficients were constructed using empirical combination rules. In D3, all of the pairwise coefficients have been calculated using time-dependent DFT not only for each pair of elements but also considering the possible different valence states of the atoms. To allow a change of the parameters during a reaction, the valence is defined as a continuous function of the coordinates of the surrounding atoms, and the $c_{6,ij}$ coefficient is interpolated from the values tabulated for the reference valences. Also the cutoff radii were determined by calculation as dispersion-specific pairwise radii, replacing the previously used sum of atomic van der Waals radii. For further details on the dispersion correction, we refer the reader to the original paper.

However, the dispersion correction itself does not yield satisfactory results when used with PM6 and other semiempirical methods. We found that there is a specific error in the description of hydrocarbons where the intermolecular distance is strongly underestimated owing to weak Pauli repulsion between hydrogens. This cannot be corrected by the dispersion, which is only attractive. Therefore, we had to add a repulsive term to all pairs of hydrogen atoms. We have chosen the form of a smooth sigmoidal function:

$$E_{\mathrm{rep}}(r_{ij}) = s_{\mathrm{HH}} \times \left(1.0 - \frac{1.0}{1.0 + \exp\left(-e_{\mathrm{HH}} \left(\frac{r_{ij}}{r_{0,hh}} - 1.0 \right) \right)} \right) \quad (11)$$

where s_{HH} sets the strength of the correction, $r_{0,\mathrm{HH}}$ defines the distance where the function acts, and the exponent e_{HH} determines how steep it is. This function is flat at short distances and therefore does not affect the covalent-bond region in any other way than by adding a constant. At the close noncovalent region, it approximates the exponential repulsion well. While this repulsive correction is independent of the dispersion, in practical implementation it is calculated along with the dispersion correction and is included in the D3 term in our notation.

REFERENCE DATA

Training Set. Both corrections have been parametrized on the recently introduced S66x8 data set.[23] It features CCSD(T)/CBS dissociation curves for 66 noncovalent complexes, covering hydrogen bonds, dispersion, and mixed dispersion/electrostatic interactions. The hydrogen bonds in the set cover all of the combinations of the most common donor/acceptor groups and include also cyclic double hydrogen bonds. Unlike the previously used S22 set,[24] it also provides a more balanced coverage of dispersion, including both $\pi-\pi$ stacking and dispersion in aliphatic hydrocarbons. These improvements over the previously used benchmark data and availability of the dissociation curves instead of equilibrium geometries are a great advantage in this application and lead to a more robust resulting method.

Charged H-Bonds. To develop the H-bonding correction for systems with charged groups, we built a new data set consistent with the S66 set. It covers the charged moieties found in proteins, using model molecules acetate, methylammonium, guanidinium, and imidazolium (heterocycle in the protonated histidine), interacting with small donor/acceptor molecules (water, methanol, methylamine, and formaldehyde). Details on this set are provided in the Supporting Information. The geometries of the complexes and the benchmark CCSD(T)/CBS results are also available online in the BEGDB database[25] (www.begdb.com).

Validation Sets. We test the methods on multiple data sets covering noncovalent interactions in organic molecules and biomolecules. In addition to the S66x8 set used in parametrization, we report separately the results for the equilibrium geometries, the S66 set.[23] The same complexes are used in the S66a8 set[26] to cover those geometries distorted in the intermolecular angular coordinates. We also employ the S22 data set[24] (used to parametrize the previous generations of the corrections), as recalculated in the large basis set by Szalewicz.[27] Two separate sets are utilized for the validation of the H-bonding and dispersion separately: a set of 104 H-bonds (abbreviated as HB104 here) optimized and calculated at the MP2 level, developed for the parametrization of the original -DH correction,[6] and a set of hydrocarbon dimers[28] (abbreviated as HC12). The latter covers at the CCSD(T)/CBS level the series propane to hexane, cyclopropane to cyclohexane, butadiene, and hexatriene and their cyclic analogs. Another test set, labeled here as AA24, is a set of neutral and charged amino acid side chains in the geometries most commonly found in proteins.[29]

COMPUTATIONAL SETUP

The PM6, RM1, AM1, and PM3 calculations, including the DH+ correction, have been carried out in MOPAC 2009.[30] The OM3 method was used as implemented in the MNDO 2005 program.[31] For the SCC-DFTB calculations, the DFTB+ program has been used.[32]

The SCC-DFTB calculations use the third-order expansion and modified N−H parameters[33] that improve the interaction energies. In our comparison, we have also included the SCC-DFTB with the original dispersion correction[34] (denoted as -D) and with modified electrostatics of hydrogen (denoted as γ), which improves the hydrogen bonding.

The D3H4 correction energy is solely a function of the molecular geometry. All of the corrections here are independent of each other, and the corrected total energy is a sum of the energy from the semiempirical calculation (E_{SQM}) and the correction terms:

$$E_{\mathrm{SQM\text{-}D3H4}} = E_{\mathrm{SQM}} + E_{\mathrm{HB}} + E_{\mathrm{disp}} + E_{\mathrm{rep}} \quad (12)$$

where E_{HB} is given by eq 1, E_{disp} by eq 9, and E_{rep} by eq 11.

The D3H4 correction developed here was implemented separately, so that it can be applied to results from any of the

dx.doi.org/10.1021/ct200751e | J. Chem. Theory Comput. 2012, 8, 141–151

programs above. This experimental code is based on the Cuby framework developed in our laboratory. We plan to implement these corrections in MOPAC.[30] A standalone program for calculation of the hydrogen bond correction is available at the author's Web site (www.molecular.cz/~rezac). A program implementing the D3 correction (without parameters for SQM methods) is available at the Web site of Grimme's group (toc.uni-muenster.de/DFTD3/).

PARAMETERIZATION

The dispersion correction is parametrized first on complexes without hydrogen bonds. Subsequently, the hydrogen-bonding correction is optimized, taking into account the contribution of dispersion to the H-bonded complexes.

All of the parametrizations described here are least-squares optimizations, minimizing the root-mean-square error of the interaction energy when compared to the CCSD(T)/CBS reference.

Dispersion. The parametrization of the dispersion correction is not trivial and cannot be fully automated. These problems are caused by the unbalanced description of different types of molecules given by the SQM methods and, in the case of PM6, also the partial coverage of the dispersion. Therefore, multiple separate steps are needed to get a robust set of parameters. The final balancing of the different types of interaction was done by hand. To obtain the parameters presented here, the following protocol was employed:

(1) As we have already shown, in the case of PM6, it is necessary to introduce scaling of the dispersion correction energy. We derive the scaling coefficient s_6 from the long-distance interactions where the effect of the damping function is negligible. We have used the most distant points in the S66x8 data set (displaced to 2× the equilibrium distance) of the dispersion group. The optimization of the coefficient yields a value of 0.88, which is almost the same value as that used in the previous generation of the correction. This scaling is not needed in the other methods investigated here (s_6 is 1.0).

(2) In the next step, the damping function is optimized. For this, we use dispersion and mixed-type complexes from S66x8, excluding the aliphatic hydrocarbons that exhibit anomalous behavior, which is corrected later.

(3) If no special measures are taken, the interaction between the aliphatic hydrocarbons is overestimated by all of the methods considered here. This effect is strongest in the case of PM6, the intermolecular distance becomes extremely short when optimized with not only PM6-D or PM6-D3 but also PM6 alone (1.5 Å hydrogen−hydrogen contact in the worst case). This problem cannot be fixed by tweaking the dispersion, but a separate repulsive correction has to be added. This repulsion has been optimized on the most problematic system in the S66 set, the neopentane dimer. The optimization of the function on other hydrocarbons yields too weak a repulsion to correct the geometry of the neopentane dimer. In order to introduce the least perturbation, we do not optimize all of the variables freely, but we seek the smallest s_{HH} for which one can obtain the correct shape of the dissociation curve while optimizing the remaining two parameters. The effect of the repulsive H−H correction on the dissociation curve of the hydrogen molecule dimer is illustrated in plot S1 in the Supporting Information. Even in this simple system, this correction is necessary for reproducing both the interaction energy and intermolecular distance of the equilibrium structure.

(4) This repulsive correction now leads to an underestimation of the interaction in all of the hydrocarbons except for neopentane. Here, no universal solution that works for all systems can be found. Therefore, we have manually adjusted the strength of the repulsive term (by changing the s_{HH} parameter) to get the best interaction energies overall with the condition of conserving a reasonable (within 5% from the benchmark) intermolecular distance in the neopentane dimer.

Table 3. The Dispersion Correction Parameters for the Methods Considered in This Study[a]

parameter	PM6	SCC-DFTB	RM1	OM3	AM1	PM3
s_6	0.88	1.0	1.0	1.0	1.0	1.0
s_r	1.18	1.215	1.0	1.14	0.90	0.90
α	22	30	16	23	15	22
s_{HH} (kcal/mol)	0.4	0.3	0.3	0.3	0.9	0.9
e_{HH}	12.7	14.31	4.46	9.60	4.46	6.86
$r_{0,HH}$ (Å)	2.30	2.35	2.11	2.10	2.11	2.23

[a] The parameters are dimensionless unless indicated otherwise.

The final set of parameters is listed in Table 3. The final performance of this correction is negligibly worse (by 0.01 kcal/mol in the S66 dispersion complexes in PM6) than the solution obtained by a blind optimization of the dispersion correction without the repulsive term, but the description of the hydrocarbons is improved significantly, bringing more balanced errors in the different types of interactions.

Hydrogen Bonding. The first step of the development of the H-bonding correction was the design of the functional form, mainly the shape of the radial and angular terms. Here, we have attempted to build functions with the desired properties described above, matching the distance and angular dependence of the error between PM6 and the reference data. For the radial term, we have used the dissociation curves of the hydrogen-bonded complexes in the S66x8 data set; the angular term was optimized on angular scans in methanol and methylamine dimers.

Once the form of the potential is set, it is straightforward to optimize the free parameters, the coefficients determining the strength of the correction. The coefficients for all donor−acceptor combinations (c_{OO}, c_{ON}, c_{NO}, c_{NN}) have been optimized along with the coefficient for scaling the H-bonds with the water donor, c_{wat}, on the hydrogen-bonded complexes in the S66x8 data set.

Finally, the scaling coefficients for the charged groups present in the training set (carboxylic acids, ammonium, guanidinium and imidazolium), $c_{s,COO-}$, $c_{s,NHR3+}$, $c_{s,gua}$, and $c_{s,imz}$, have been optimized on a set of charged hydrogen bonds. The resulting parameters are listed in Table 4. Note that the parameters for different cations differ significantly; it is not possible to use a single parameter and achieve the desired accuracy here.

We have attempted to reoptimize some of the parameters in the functional form of the correction, e.g., the radii defining the radial potential, on the S66x8 set. Although the overall error can be slightly decreased this way, the geometric parameters are worse (the minima become shifted away from the reference geometry). Therefore, we keep the original radial function

designed to have the optimal shape to ensure a robust description of the complex geometries.

RESULTS AND DISCUSSION

Tests on Benchmark Data. The corrected SQM methods developed here have been tested on multiple benchmark data sets. Table 5 lists the RMSE for all of the tested methods and sets. Other error measures, the mean and maximum unsigned errors, are listed in Tables S2 and S3 in the Supporting Information. This overview includes sets used for the parametrization of the corrections in some of the methods (S66x8 and the charged H-bonds for D3H4 and S22 for the DH2 and DH+ corrections).

Table 4. The Hydrogen-Bonding Parameters for the Methods Considered in This Study[a]

parameter	PM6	SCC-DFTB	RM1	OM3	AM1	PM3
c_{OO} (kcal/mol)	2.32	1.11	3.76	1.95	4.89	2.71
c_{ON} (kcal/mol)	3.10	2.58	3.90	1.64	6.23	4.37
c_{NO} (kcal/mol)	1.07	0.80	3.14	0.93	2.54	2.29
c_{NN} (kcal/mol)	2.01	2.01	2.95	1.35	4.56	3.86
c_{wat}	0.42	1.32	0.94	0.50	0.49	0.91
$c_{s,COO-}$	1.41	1.22	1.10	1.63	1.08	0.89
$c_{s,NHR3+}$	3.61	2.33	1.21	0.9	2.78	2.54
$c_{s,gua}$	1.26	2.42	1.18	1.37	0.86	1.54
$c_{s,imz}$	2.29	3.44	1.10	1.18	2.11	1.84

[a] The parameters are dimensionless unless indicated otherwise.

The average of the errors over all of the data sets (E_{all}) is used as a simple measure of the overall performance of each method. Additionally, the average over the sets that were not used in the parametrization of any method (HB104, large set of H-bonds; hydrocarbon dimers, HC12 and AA24, amino acid side chains), E_{val}, is provided as an independent validation. The optimization of the new correction for the hydrogen bonds of the charged groups leads to an important decrease of error in these cases. Here, a separate parametrization is needed for each functional group. The data set listed here is the one used for parametrization and covers all of the charged amino acids in proteins.

When different corrections are compared for each SQM method, the D3H4 approach developed here yields the best results both on the validation sets and overall. In the following text, we have ordered the methods by their overall best score, discussing the results in detail and comparing the possible correction schemes for each given method. Here, we have also discussed the advantages and disadvantages of each method for practical applications.

Among the methods tested, OM3-D3H4 yields the lowest errors (E_{all} 0.69 kcal/mol, E_{val} 0.63 kcal/mol), which is an improvement of about 35% over OM3-DH2. We cannot compare OM3-DH+ here, because we do not have software that implements this combination, but on the basis of the original paper, we expect it to be somewhere between -DH2 and -D3H4. Although OM3-D3H4 scores best for interaction energies, there are two issues that make it impractical for many applications. First, the OM3 method is parametrized for only a few elements

Table 5. The Root Mean Square Errors (in kcal/mol) of the Studied Methods in Multiple Benchmark Data Sets[a]

	S66	S66x8	S66a8	S22	H bonds	charged HB	hydrocarbons	aaside chains	avg	$avg_{testing}$
PM6	3.02	2.49	2.12	4.16	3.18	3.92	2.64	4.08	3.20	3.30
PM6-DH2	0.91	0.79	0.73	0.54	1.52	2.21	**0.67**	1.32	1.09	1.17
PM6-DH+	0.82	0.76	**0.67**	0.80	1.43	1.94	**0.67**	1.89	1.12	1.33
PM6-D3H4	**0.65**	**0.66**	0.68	**0.78**	**1.05**	**1.11**	0.71	**1.17**	**0.85**	**0.98**
PM6-D3H4*	0.70	0.71	0.74	0.84	1.12	2.26	0.71	1.86	**1.12**	**1.23**
DFTB	2.88	2.40	2.24	3.45	2.82	4.78	2.90	3.44	3.11	3.05
DFTB-D	1.50	1.43	1.28	1.63	1.96	4.28	0.59	2.27	1.87	1.60
DFTB-D,γ	1.17	1.17	1.04	1.21	1.61	3.67	**0.56**	1.82	1.53	1.33
DFTB-DH2	1.44	1.15	0.98	1.86	1.54	2.13	0.59	1.62	1.41	1.25
DFTB-D3H4	**0.67**	**0.62**	**0.61**	**0.97**	**0.71**	**1.43**	0.59	**0.88**	**0.81**	**0.73**
RM1	5.39	4.38	4.13	7.15	5.40	5.60	3.65	5.34	5.13	4.80
RM1-D3H4	0.92	**0.90**	**0.78**	1.03	0.90	**2.05**	**0.24**	**0.73**	**0.94**	**0.62**
RM1-D3H4*	**0.91**	0.90	0.79	**1.03**	**0.89**	2.09	**0.24**	0.93	0.97	0.69
OM3[b]	3.33	2.70	2.49	4.17	2.88	3.00	3.93	4.99	3.44	3.93
OM3-DH2[b]	0.80	0.96	0.62	0.96	0.84	1.83	1.11	**1.53**	1.08	**1.16**
OM3-D3H4[b]	**0.48**	**0.60**	**0.42**	**0.58**	**0.56**	**1.50**	**0.70**	2.34	**0.90**	1.20
AM1	6.24	5.27	4.03	8.66	6.10	7.64	3.73	6.38	6.01	5.40
AM1-DH2	1.93	1.96	1.47	**0.85**	**2.08**	3.58	3.94	3.71	2.44	3.25
AM1-D3H4	**1.35**	**1.76**	**1.45**	1.76	2.11	**3.04**	**0.82**	**2.02**	**1.79**	**1.65**
PM3	5.08	4.51	3.77	7.64	4.98	7.03	2.25	4.60	4.98	3.94
PM3-D3H4	**1.40**	**1.26**	**0.97**	**2.51**	**0.83**	**2.23**	**0.40**	**1.05**	**1.33**	**0.76**
B3LYP/6-31G*	2.68	2.40	1.87	3.63	1.31	3.17	4.20	2.97	2.78	2.82
TPSS/TZVP-D	0.69	0.53	0.57	0.58	1.04	1.89	0.72	0.89	0.86	0.88
BLYP/def2QZVP-D3	0.25	0.17	0.21	0.33	0.41	0.59	0.21	0.46	0.33	0.36
MP2/cc-pVTZ	0.70	0.59	0.57	1.85	1.40	1.81	0.88	1.62	1.18	1.30

[a] The last two columns list the average of these errors over all of the sets and over the validation sets only. [b] Due to the limited parameter set, the methionine complexes in AA side chains set are not considered.

dx.doi.org/10.1021/ct200751e | J. Chem. Theory Comput. 2012, 8, 141–151

(H, C, N, and O). More importantly, this method critically fails in geometry optimizations of complexes containing acetic acid, as described in the following section.

The second most accurate method (E_{all} 0.81 kcal/mol, E_{val} 0.73 kcal/mol) is SCC-DFTB when used with the D3H4 correction and modified parameters for hydrogen−nitrogen interaction.[33] This is a clear improvement over DH2 but also over the original dispersion correction[34] and modified electrostatics (SCC-DFTB-D,γ) developed by the authors of DFTB to improve the description of the hydrogen bonds.[33]

PM6-D3H4 scores third overall (E_{all} 0.85 kcal/mol, E_{val} 0.98 kcal/mol). The parametrization of PM6-D3H4 on the S66x8 set leads to very good results for all of the sets in the S66 family. Achieving such low errors with a substantially simplified H-bonding correction is a very encouraging result indicating a good choice of the form of the correction. The performance on the S22 set is better than that of PM6-DH+ but not as good as that of PM6-DH2 (both of these methods used S22 for parametrization). What is more important is the independent tests on systems outside the training set. In the set of 104 H-bonds, PM6-D3H4 outperforms all of their predecessors with a RMSE of 1.1 kcal/mol. This is the most important result, which clearly shows the accuracy and robustness of the new H-bonding correction. In the set of hydrocarbon dimers, the results are slightly worse than in the previous version, but the D3 correction corrects the problems with short intermolecular distances in aliphatic hydrocarbon dimers. Although PM6 has some limitations, the accuracy that can be achieved for noncovalent interactions, in combination with the coverage of a major part of the periodic table (both by PM6 and the D3 dispersion), makes PM6-D3H4 very useful for applications.

To demonstrate the effects of the introduction of the water atom type and the scaling of charged hydrogen bonds, we list results obtained without these modifications (PM6-D3H4* in Table 5). In the case of water as a hydrogen bond donor (included in S66, S22, and H-bonds sets), the increase of the overall errors is rather small, but the error for the water dimer is rather large (the binding is overestimated by 1.2 kcal/mol, which is 25% of the interaction energy). We are convinced that hydrogen bonds in and with water are important in many applications, and correcting this error is worth the specific scaling in the H-bond correction. Regarding the hydrogen bonds in charged systems, the error is about twice as large as when the scaling is applied. A RMSE of 2.26 kca/mol translates to rather small relative errors, as the interaction energies in these systems are larger than in neutral H-bonds. The improvement brought by the system-specific scaling allowed us to achieve high accuracy consistently in a wide range of systems. When the ultimate accuracy is not needed, it is possible to use the H-bond correction without this scaling; in such a case, the errors are comparable to the previous generations of the correction.

In this paper, the RM1 method has been coupled with empirical corrections for the first time, and the results are very promising. The error in the validation sets is very low (0.62 kcal/mol), but the overall results are slightly worse because of the large error in the charged H-bonds. Unlike all of the other methods, RM1 works comparably well without separate scaling of the H-bonds with the water donor and in charged H-bonds (the method is denoted as RM1-D3H4*). This system-independent nature of the errors indicates that the method is robust and not overparameterized. We plan to test this method in applications where the limited set of parameters (H, C, N, O, P, S, and halogens) makes it possible.

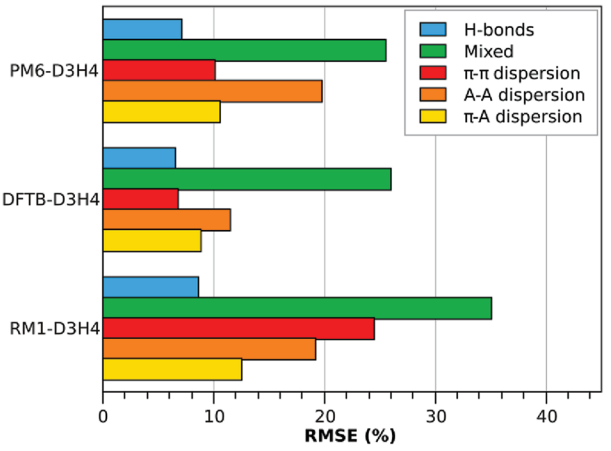

Figure 3. The relative errors of the selected methods for interactions of different types in the S66 set: H-bonds; mixed type; and $\pi-\pi$, aliphatic−aliphatic, and π−aliphatic dispersion. Errors are plotted as a percent of the average interaction energy in the group.

The AM 1 and PM3 methods were included only for comparison; it is obvious from the results that the more modern SQM methods can describe noncovalent interactions better. Here, PM3 performs better than AM1 but still yields rather large errors for some data sets.

Table 5 also lists results of selected DFT and wave function methods for comparison. The widely used B3LYP/6-31G* setup was chosen as a representative DFT method without any treatment of dispersion. DFT with empirical dispersion is represented by the TPSS/TZVP-D method,[16] which yields very good results with a reasonably small basis set, and BLYP/def2-QZVP with the D3 dispersion correction[17] using the Becke-Johnson damping function[35] as a demonstration of the highest accuracy that can be achieved with DFT-D when large basis set and advanced dispersion corrections are used. Finally, MP2 with the cc-pVTZ basis set (with counterpoise correction of basis set superposition error) was chosen as an example of a relatively inexpensive correlated QM method. The best corrected semiempirical methods (PM6-D3H4, DFTB-D3H4, and OM3-D3H4) outperform DFT and MP2 and are about as accurate as DFT-D with a medium-sized basis set.

For the selected method that performed best, we have analyzed the errors in the S66 data set in more detail. The set is divided into five groups—H-bonds; $\pi-\pi$, π−aliphatic, and aliphatic−aliphatic dispersion; and mixed-type interactions. For each group, the relative error is calculated as RMSE expressed in the percentage of the average interaction energy in the group to make the errors comparable between interactions of different strengths. The results are plotted in Figure 3. The hydrogen bonds are described very well by all of these methods. In DFTB-D3H4 and PM6-D3H4, the largest error in the dispersion complexes is found in the complexes of aliphatic hydrocarbons, because these complexes exhibit very large errors already in the uncorrected methods. RM1-D3H4 has the largest error among the dispersion groups for $\pi-\pi$ interactions, where the stacking interactions of uracil are underestimated by as much as 1.7 kcal/mol. It is not surprising that the largest errors are observed in the mixed-type groups, where the interactions are not specifically corrected (e.g., X−π interactions and C−H hydrogen bonds).

Geometries. Another important test of the corrected method is its application to the optimization geometry of noncovalent

Table 6. The Results of the Geometry Optimization of the S66 Complexes by the Studied Method[a]

	avg. RMSD	max. RMSD	RMSE S66	RMSE $S66_{opt}$
PM6	0.38	2.34	3.1	2.7
PM6-D/PM6-DH2	0.26	0.85	0.9	1.0
PM6-DH2 (numer.)	0.22	0.73	0.9	1.3
PM6-DH+	0.25	1.30	0.8	1.0
PM6-D3/PM6-D3H4	0.25	1.30	0.7	0.7
PM6-D3H4	0.20	0.51	0.7	0.7
DFTB	0.37	1.79	2.9	2.3
DFTB-D	0.24	0.83	1.6	1.1
DFTB-D,γ	0.23	0.77	1.2	0.8
DFTB-D3H4	0.21	0.77	0.7	0.9
RM1	0.55	5.08	5.4	4.4
RM1-D3H4	0.23	1.01	1.0	0.9
RM1-D3H4*	0.23	0.93	0.9	0.9
OM3	0.65	6.36	3.4	5.9
OM3-D/OM3-DH2	0.24	1.02	0.9	6.3
OM3-D3H4	0.20	0.73	0.5	5.0
AM1	0.80	5.95	6.3	4.1
AM1-D3H4	0.39	2.23	1.4	2.0
PM3	0.58	2.23	5.1	4.0
PM3-D3H4	0.37	2.26	1.4	1.2
TPSS/TZVP-D	0.11	1.35	0.7	1.3

[a] We describe the changes in geometry by the average and largest root-mean-square deviation (RMSD, in Å) and the changes in the interaction energy by listing the RMSE (in kcal/mol) in the S66 set before and after geometry optimization.

complexes. We evaluate two criteria: First, the optimized geometry should be as close as possible to a benchmark one optimized with a high-level QM method. Second, the interaction energies calculated on the optimized geometries should be close to the reference values calculated at reference energy. When these two criteria are satisfied, the method is applicable to practical calculations that involve geometry optimization and an evaluation of the properties of the resulting structure.

We performed these tests on the S66 data set. We optimized each of the complexes with the studied method with high accuracy (convergence limits of 0.03 kcal/mol/Å for the RMS gradient, 0.06 kcal/mol/Å for the max. gradient component, and a 3.0e−4 kcal/mol energy difference between the subsequent steps). The interaction energies are recalculated on the new geometries and compared to the benchmark. The results are summarized in Table 6. We list the average and largest root-mean-square deviations (RMSD, in Ångstrom) compared to the reference MP2/cc-pVTZ (counterpoise corrected) geometry along with the RMSE of the interaction error on the benchmark geometries and after optimization with the tested methods.

In some cases, the optimization is problematic. The DH2 correction uses only an approximate gradient, and it is not possible to converge the optimizations. Instead, we list the results of the optimizations with dispersion only, but we calculate the interaction energies with both dispersion and H-bond correction; this is the protocol we have used and recommend for applications of this method. For PM6-DH2, we also performed full optimization using the much more expensive numerical evaluation of the gradient. The DH+ correction fixes this issue, but the gradient is not smooth. The cusp corresponds to the linear arrangement; therefore, it is impossible to converge the optimizations of symmetric systems (namely, complexes 1, 8, 17, and 23). Since the resulting structure is very close to the minimum although the gradient is nonzero, we use the unconverged geometries for further analysis.

Table 7. The Interaction Energy (in kcal/mol) Per Molecule in a Cubic Box with 216 Water Molecules[a]

	ΔE/molecule	ΔE_{dimer}
PM6	−5.2	−3.9
PM6-DH2 angle only	−9.5	−4.9
PM6-DH2	−8.4	−4.9
PM6-DH+	−9.6	−6.5
PM6-D3H4	−8.3	−4.9
TIP3P	−7.4	−6.0
CCSD(T)		−5.0

[a] The interaction energy in the water dimer is listed for comparison.

A serious issue is observed in the OM3 method, regardless of whether the corrections are applied or not. The complexes containing acetic acid optimize into covalently fused structures where the hydrogen has an equal distance of 1.2 Å from the donor and acceptor atoms. The interaction energy in these cases is exaggerated by 100 and 200%. If this issue were removed, the method would perform rather well, as indicated by the low average RMSD.

The D3H4 consistently yields the lowest average and maximum RMSD when compared to the other possible correction schemes. The improvement in PM6, where PM6-D3H4 yields very low maximum RMSE compared to its predecessors, is an important achievement. This is partly due to the special treatment of the dispersion in aliphatic hydrocarbons, where the newly introduced repulsive correction is needed to obtain good geometries. Also, the new H-bonding correction works slightly better, although it is simpler than the previous versions and uses less information from the local geometry of the hydrogen bond (see Table 1). The average RMSD in the 23 H-bonds in the S66 set is 0.24 Å for PM6-DH+ and improves only slightly to 0.23 Å in PM6-D3H4, but the largest RMSD decreases from 0.94 to 0.51 Å. What is also very important is that the interaction energies do not change significantly when the structures are optimized.

We applied the same optimization protocol to DFT-D[16] in a medium-sized basis set (TZVP). This method is comparable to the corrected semiempirical methods in terms of interaction energies. The results of geometry optimizations are better overall (the average RMSD is 0.11 Å while the best SQM-D3H4 methods yield 0.2 Å). A slightly larger maximum RMSD and an increase of the error in interaction energies in the optimized complexes is caused by a single system, where the methylamine−methanol complex with amine hydrogen bond donor optimizes to the global minimum with the alcohol as a hydrogen donor.

Tests on Large Systems. As the corrections are developed on small model complexes, it is necessary to evaluate their transferability to large systems. The dispersion correction should have no problems here, as the dispersion interaction is almost additive and the pairwise potential used is a good approximation to it. The scaling of the whole correction is obtained from calculations of the complexes displaced to twice the equilibrium geometries in order to ensure that the dispersion is not overestimated at longer distances.

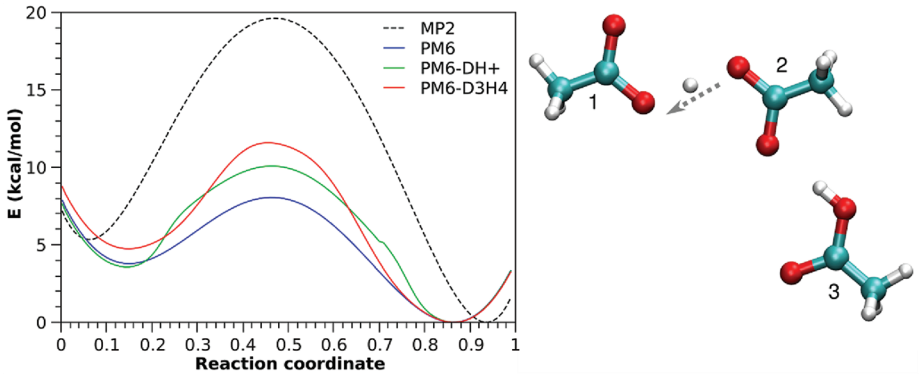

Figure 4. The proton transfer in a network of hydrogen bonds. The energy profile along the reaction coordinate obtained with PM6 (blue), PM6-DH+ (green), and PM6-D3H4 (red) is compared with the MP2/aug-cc-pVDZ reference (black).

More attention has to be paid to the hydrogen bonds. In condensed systems, the number of potential H-bonds grows rapidly, and even if they contribute only negligibly to the total energy, the overall stabilization arising from the H-bond correction might be overestimated. This problem, observed in the first generation of correction, was addressed by a more complex definition of the hydrogen bonds in DH2 and DH+, using additional internal coordinates. Here, we have attempted to solve this problem by making the correction rather short-ranged, which eliminates all of the false contributions from potential H-bonds other than those actually interacting.

We test this on a system with very high H-bond density—water. We calculate the total interaction energy in a cubic box of 216 water molecules and list it as interaction energy per single water molecule. In Table 7, we compare the results of PM6 with the DH2, DH+, and D3H4 corrections with PM6-DH2 without the additional angular and torsional coordinates (using the H-bond angle α only). The interaction energies in the optimized structure of a water dimer are included for comparison. For reference, we list the values obtained with the TIP3P force field and accurate CCSD(T) interaction energy.

These results show that the new approach is as efficient as the use of additional coordinates in the DH2 correction, while the new correction is much simpler and can be calculated more efficiently. In this case, PM6-DH+ yields larger average interaction in the cluster, although it uses a formalism very similar to DH2. This is caused by the overestimated interaction of the H-bond of this type already in the water dimer.

Proton Transfer. The new hydrogen-bonding correction can seamlessly describe proton transfer, because it smoothly switches the donor and acceptor when the hydrogen atom is in the middle. The description is more complicated when the proton transfer studied involves a charged group, because the scaling of the H-bonding correction applied to that group changes as well. Using the continuous scaling introduced above, a smooth potential is obtained even in the most complex cases. This is illustrated in Figure 4 on the potential energy curve along a proton transfer in a network of carboxylic groups with a total charge of −1, with the MP2/aug-cc-pVDZ calculation serving as a reference. In this simple model, all of the coordinates have been fixed, while the proton is transferred along the H-bond axis. In this system, central molecule 2 becomes charged as it loses the proton, which makes the second hydrogen bond between molecules 2 and 3 stronger. The correction stabilizes the minima but does not affect the energy of the transition state, which effectively increases the barrier. This is an improvement toward the reference curve when compared to uncorrected PM6, although the barrier height is still underestimated. We also include the PM6-DH+ results for comparison. Unlike its predecessors, this method should yield a smooth potential energy curve. In the practical implementation in MOPAC2009, there is a minor discontinuity close to a reaction coordinate value of 0.7. This most probably arises from the use of a distance cutoff in the H-bond correction. Overall, the shape of the curve, the energy difference between the minima, and the barrier height are not as good as in PM6-D3H4.

CONCLUSIONS

Empirical corrections for noncovalent interactions can substantially improve the performance of semiempirical quantum mechanical methods, reaching chemical accuracy (error of 1 kcal/mol) in most of the benchmark data sets studied. These results are very close to much more expensive methods, such as DFT-D or MP2, while the efficiency of the SQM method makes it possible to study very large systems on a routine basis.

The accuracy of the corrected SQM method approached its limits already with the DH2 correction, and the later advancements including the one presented here aim mainly to improve the robustness of the method. This was achieved by adopting the latest developments in the dispersion corrections for the DFT methods and redesigning the H-bonding correction from scratch. Although the H-bonding correction has been substantially simplified, the accuracy was improved.

We have addressed multiple weaknesses of the previous generations of the corrections. Most importantly, the D3H4 correction is the first one that can be used for geometry optimizations and molecular dynamics, as it and its derivatives have a continuous and smooth potential energy surface. For the first time, we have used scaling of the correction in charged hydrogen bonds in order to improve the accuracy in these systems.

The new H-bond correction does naturally describe proton transfer along a hydrogen bond, yielding a smooth potential energy surface even in the most complex cases. Stabilization of the minima effectively increases the barrier height, improving the SQM results toward a more accurate reference.

Among the tested methods, PM6-D3H4, DFTB-D3H4, and RM1-D3H4 yield errors lower than 1 kcal/mol in multiple benchmark data sets. We have also shown that these methods reproduce geometries of noncovalent complexes with good accuracy, which makes them useful for many applications.

Semiempirical methods with corrections for noncovalent interaction can yield very accurate results on small model systems and have been successfully applied to real-world systems. However, we would like to end this paper with a warning: The accuracy of both the SQM methods and of the corrections is achieved by empirical parametrization, and they can yield large errors when applied to systems that are outside of this parametrization. Therefore, it is advised to examine the results critically and possibly check them against more reliable calculations.

■ ASSOCIATED CONTENT

Supporting Information. Description and geometries of the charged hydrogen bonds data set and additional tables listing mean and maximum errors for all the studied methods in multiple benchmark data sets are provided as Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Fax: +420 220 410 320. E-mail: rezac@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was a part of Research Project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and was supported by Grant No. MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. It was also supported by the Research and Development for Innovations Operational Program of the European Social Fund (CZ.1.05/2.1.00/03.0058). The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. We are grateful to James Stewart for sharing his knowledge of semiempirical methods, to Martin Korth for discussion on testing of the H-bond correction in condensed systems, and to Filip Lankaš for his help with the construction of the polynomial functions.

■ REFERENCES

(1) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.
(2) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101–1111.
(3) Tuttle, T.; Thiel, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2159–2166.
(4) Stewart, J. J. P. *J. Mol. Model.* **2008**, *15*, 765–805.
(5) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
(6) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.
(7) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.
(8) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(9) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
(10) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808–3816.
(11) Fanfrlík, J.; Bronowska, A.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.
(12) Pecina, A.; Přenosil, O.; Fanfrlík, J.; Řezáč, J.; Granatier, J.; Hobza, P.; Lepšík, M. *Collect. Czech. Chem. Commun.* **2011**, *76*, 457–479.
(13) Dobeš, P.; Fanfrlík, J.; Řezáč, J.; Otyepka, M.; Hobza, P. *J. Comput.-Aided Mol. Des.* **2011**.
(14) Dobeš, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P. *J. Phys. Chem. B* **2011**, *115*, 8581–8589.
(15) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473.
(16) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. *J. Comput. Chem.* **2007**, *28*, 555–569.
(17) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.
(18) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
(19) Clark, T. *THEOCHEM* **2000**, *530*, 1–10.
(20) Winget, P.; Selcuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254–266.
(21) Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, *14*, 639–646.
(22) Zhang, P.; Fiedler, L.; Leverentz, H. R.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**, *7*, 857–867.
(23) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.
(24) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
(25) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrášek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261–1270.
(26) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.
(27) Podeszwa, R.; Patkowski, K.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5974.
(28) Granatier, J.; Pitoňák, M.; Hobza, P. Unpublished data.
(29) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982–992.
(30) Stewart, J. J. P. *MOPAC 2009*; Stewart Computational Chemistry: Colorado Springs, CO, 2009.
(31) Thiel, W. *MNDO 2005*; Max Planck Institute for Coal Research: Mülheim, Germany, 2005.
(32) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.
(33) Yang, Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861–10873.
(34) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
(35) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.