

# Corrected Small Basis Set Hartree-Fock Method for Large Systems

Rebecca Sure and Stefan Grimme\*

A quantum chemical method based on a Hartree-Fock calculation with a small Gaussian AO basis set is presented. Its main area of application is the computation of structures, vibrational frequencies, and noncovalent interaction energies in huge molecular systems. The method is suggested as a partial replacement of semiempirical approaches or density functional theory (DFT) in particular when self-interaction errors are acute. In order to get accurate results three physically plausible atom pair-wise correction terms are applied for London dispersion interactions (D3 scheme), basis set superposition error (gCP scheme), and short-ranged basis set incompleteness effects. In

total nine global empirical parameters are used. This so-called Hartree-Fock-3c (HF-3c) method is tested for geometries of small organic molecules, interaction energies and geometries of noncovalently bound complexes, for supramolecular systems, and protein structures. In the majority of realistic test cases good results approaching large basis set DFT quality are obtained at a tiny fraction of computational cost. © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23317

## Introduction

Noncovalent interactions such as van der Waals interactions or H-bonding play a crucial role in the chemistry of supramolecular and biomolecular systems as well as for nanostructured materials.<sup>[1,2]</sup> They control host-guest and enzyme-substrate binding, structures of proteins and DNA, antigen-antibody recognition or the orientation of molecules on a surface. Theoretical methods based on first principles to complement experimental studies which often can provide only limited information about these complex soft-matter systems seem indispensable.

Many of these systems or at least reasonable models thereof can nowadays be computed routinely with quite good accuracy by (dispersion corrected) density functional theory (DFT) together with relatively large basis sets (triple-zeta quality or better). For recent reviews how to treat the important long-range London dispersion interactions in DFT, see Refs. [3, 4]. One perspective of such treatments is to provide accurate input data to parameterize simpler force-field or even coarse-grained theoretical models although full protein structures can be treated.<sup>[5]</sup> But despite of the good cost-accuracy ratio of DFT for large systems, these calculations are often prohibitive in terms of the necessary computational efforts. Furthermore, the quadrature of the exchange-correlation energy in DFT causes numerical noise in geometry optimizations or frequency calculations which is a particular problem in these often flexible systems. Accurate harmonic frequencies are an important ingredient for the computation of thermodynamic properties as for example free enthalpies of association of supramolecules.<sup>[6]</sup> Another issue in DFT are charged systems (e.g., proteins with charged residues) where the self interaction error (SIE<sup>[7,8]</sup>) can lead to artificial charge-transfer and convergence problems of the self consistent field (SCF)<sup>[5,9,10]</sup> at least when “cheap” semilocal functionals of general gradient

approximation (GGA) type are used. Modern semiempirical methods like DFTB3,<sup>[11]</sup> OM2,<sup>[12]</sup> or PM6<sup>[13]</sup> (for an overview see Ref. [14]) represent an alternative in principle but suffer from missing parametrization for important elements or robustness in certain situations (e.g., charged complexes<sup>[15]</sup>).

As will be shown in this work, most of the above mentioned problems can be alleviated by applying Hartree-Fock (HF) theory together with small AO basis sets. The basic idea is to fill the gap between existing semiempirical methods and DFT in terms of the cost-accuracy ratio with a physically sound approach. Using HF has the following advantages: First, in contrast to DFT, HF does not suffer from SIE and extended charged systems even when treated unscreened (*in vacuo*) are unproblematic. Second, a HF calculation is performed completely analytical, including the computation of gradients and Hessians so that no problems with numerical noise in geometry optimizations or frequency calculations occur. Third, contrary to standard semiempirical approaches HF is inherently able to treat the important hydrogen bonding so that there is no need for atom-type dependent H-bond corrections which are normally applied for neglect of diatomic differential overlap (NDDO)-type methods.<sup>[16]</sup> Furthermore, the proposed HF method can be applied without any parametrization to almost any element of the periodic table and includes important physical effects like Pauli-exchange repulsion correctly. The accurate description of these steric interactions was always a problem in semiempirical methods<sup>[14]</sup> and even current density functionals are not free of inaccuracies for short interatomic

R. Sure, S. Grimme

Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie der Universität Bonn, Beringstr. 4, D-53115 Bonn, Germany  
E-mail: grimme@thch.uni-bonn.de

Contract grant sponsor: Fonds der Chemischen Industrie

© 2013 Wiley Periodicals, Inc.

distances.<sup>[17,18]</sup> For density functionals which try to mimic the HF short-range repulsive behavior see, for example, Ref. [19].

It is clear, however, that the Coulomb correlation energy is entirely missing in HF and a small basis set can introduce further severe errors. The suggested approach is hence not meant to be generally applicable or as a replacement of DFT. Rather, it should yield reasonable results for simple molecular properties like equilibrium structures or vibrational frequencies or for noncovalent interactions, that is, when changes in the basic electronic structure during a chemical process is small. The accurate computation of chemical reaction energies requires the account of various short-ranged polarization and correlation effects and is not of concern here (and likely not computable with a minimal or small AO basis set).

Several years ago Pople noted that HF/STO-3G optimized geometries for small molecules are excellent, better than HF is inherently capable of yielding.<sup>[20,21]</sup> Similar observations were made by Kolos already in 1979, who obtained good interaction energies for a HF/minimal-basis method together with a counterpoise-correction as well as a correction to account for the London dispersion energy.<sup>[22]</sup> It seems that part of this valuable knowledge has been forgotten during the recent "triumphal procession" of DFT in chemistry. The true consequences of these intriguing observations could not be explored fully at that time due to missing computational resources but are the main topic of this work.

We recently noted the good performance of HF/large-basis in combination with our latest dispersion correction scheme D3<sup>[23,24]</sup> for noncovalent interactions, and we will use this well-established dispersion correction (see Refs. [25–28] for recent D3 applications) also in this work. Recently, work along similar lines (i.e., using HF-D3/STO-3G) has been done by the group of T. Martinez.<sup>[29]</sup> The basis set superposition error (BSSE) is significant for a small or minimal basis set and will be treated with our recently developed geometrical counterpoise correction (gCP).<sup>[30]</sup> Importantly, this approach also accounts for intramolecular BSSE which is difficult to correct efficiently otherwise. Both schemes are used essentially in unmodified form here. Additionally, a new short-ranged basis (SRB) incompleteness correction term is applied. This corrects for systematically overestimated bond lengths for electronegative elements (e.g., N, O, F) when employing small basis sets. According to common practice, basis set effects are separated into BSSE and basis set incompleteness error (BSIE). In this sense, the SRB term corresponds to the BSIE and the gCP scheme accounts for the atom pair-wise part of the BSSE (for related BSSE correction schemes see Refs. [31, 32]).

The basis set used here is of minimal quality for the often occurring ("organic") elements H, C, N, O and mostly of split-valence (SV) or polarized SV (SVP) quality for the other elements. It is dubbed "MINIX" from now on and an inherent (fixed) ingredient of the method. For simplicity, this HF-D3-gCP-SRB/MINIX method will be abbreviated HF-3c in the following where the term "3c" stands for the three applied corrections, and the mentioned compound basis set is always implied. It should also indicate that the method accounts for the important dispersion contributions by the relatively accurate D3 scheme.<sup>[23,24]</sup>

We present HF-3c results in comparison to those obtained with the semiempirical PM6<sup>[13]</sup> method and to standard DFT. The PM6 method is used because it is parametrized for very many elements so that the same systems can be calculated for comparison. We investigate geometries of small organic molecules as well as interaction energies and geometries of small noncovalent complexes. As more realistic tests, geometries and association free enthalpies of supramolecular complexes will be considered. This also includes a test of the quality of the harmonic vibrational frequencies. Finally, HF-3c results for protein structures will be presented and compared to experimental X-ray and solution NMR data.

## Theoretical and Computational Methods

### The HF-3c method

The starting point for calculating the electronic energy is a standard HF treatment with a small Gaussian AO basis set. The herein used so-called MINIX basis set consists of different sets of basis functions for different groups of atoms (Table 1). The valence scaled minimal basis set MINIS<sup>[33]</sup> and the split valence double-zeta basis sets SV, SVP,<sup>[34]</sup> and def2-SV(P)<sup>[35]</sup> (the latter together with effective core potentials (ECP)<sup>[36]</sup> for heavier elements) are employed. Many other possibilities have been considered but the chosen one not only represents a very good compromise between accuracy and speed, but furthermore, this basis seems to be balanced and easily to correct for deficiencies (see below).

Table 1. Composition of the MINIX basis set.

Element	Basis
H-He, B-Ne	MINIS
Li-Be	MINIS+1(p)
Na-Mg	MINIS+1(p)
Al-Ar	MINIS+1(d)
K-Zn	SV
Ga-Kr	SVP
Rb-Xe	def2-SV(P) with ECP

The HF calculations are conducted in conventional mode, that is, the two-electron integrals are computed once and stored on disk or in memory if possible. This option is a further advantage of the small basis set approach and leads to large computational savings. Only huge systems are treated in direct mode by recalculating integrals in every SCF iteration. The so-called resolution of the identity (RI) approximations are not applied because the savings are negligible for small basis sets and this approach can even slow-down the computations due to overhead from the necessary linear algebra parts.

Three terms are added to correct the HF energy  $E_{\text{tot}}^{\text{HF/MINIX}}$  in order to include London dispersion interactions, to account for the BSSE and to correct for overestimated bond lengths. The corrected total energy is calculated as

$$E_{\text{tot}}^{\text{HF-3c}} = E_{\text{tot}}^{\text{HF/MINIX}} + E_{\text{disp}}^{\text{D3(BJ)}} + E_{\text{BSSE}}^{\text{gCP}} + E_{\text{SRB}} \quad (1)$$

The first correction term  $E_{\text{disp}}^{\text{D3(BJ)}}$  is the atom pair-wise London dispersion energy from the D3 correction scheme<sup>[23]</sup> and applying Becke-Johnson (BJ) damping<sup>[24,37,38]</sup>

$$E_{\text{disp}}^{\text{D3(BJ)}} = -\frac{1}{2} \sum_{A \neq B}^{\text{atoms}} \left( s_6 \frac{C_6^{\text{AB}}}{R_{\text{AB}}^6 + (a_1 R_{\text{AB}}^0 + a_2)^6} + s_8 \frac{C_8^{\text{AB}}}{R_{\text{AB}}^8 + (a_1 R_{\text{AB}}^0 + a_2)^8} \right) \quad (2)$$

Here,  $C_n^{\text{AB}}$  denotes the  $n$ th-order dispersion coefficient (orders = 6, 8) for each atom pair AB,  $R_{\text{AB}}$  is their internuclear distances and  $s_n$  are the order-dependent scaling factors. The cutoff radii  $R_{\text{AB}}^0 = \sqrt{C_8^{\text{AB}}/C_6^{\text{AB}}}$  and the fitting parameters  $a_1$  and  $a_2$  are used as introduced in the original works.<sup>[37,38]</sup> For this method, the three usual parameters  $s_8$ ,  $a_1$ , and  $a_2$  were refitted using reference interaction energies of the the S66 test set complexes.<sup>[17]</sup> This results in  $s_8=0.8777$ ,  $a_1=0.4171$ , and  $a_2=2.9149$ . The parameter  $s_6$  was set to unity as usual to enforce the correct asymptotic limit and the gCP correction (see below) was already applied in this fitting step.

The second term  $E_{\text{BSSE}}^{\text{gCP}}$  denotes our recently published geometrical counterpoise (gCP) correction<sup>[30]</sup> for BSSE, which depends only on the atomic coordinates of a given molecule. The difference in atomic energy  $E_A^{\text{miss}}$  between a large (nearly complete) basis set and the target basis set (MINIX in our case) for each free atom A is calculated for the HF Hamiltonian. The  $E_A^{\text{miss}}$  term is multiplied with a decay function depending on the interatomic distances  $R_{\text{AB}}$ . The sum over all atom pairs reads

$$E_{\text{BSSE}}^{\text{gCP}} = \sigma \sum_A^{\text{atoms}} \sum_{A \neq B}^{\text{atoms}} E_A^{\text{miss}} \frac{\exp(-\alpha(R_{\text{AB}})^\beta)}{\sqrt{S_{\text{AB}} N_{\text{B}}^{\text{virt}}}}, \quad (3)$$

where  $\alpha$ ,  $\beta$ , and  $\sigma$  are fitting parameters,  $S_{\text{AB}}$  is a Slater-type overlap integral and  $N_{\text{B}}^{\text{virt}}$  is the number of virtual orbitals on atom B in the target basis. The  $S_{\text{AB}}$  is evaluated over a single s-type orbital centered on each atom and using optimized Slater exponents weighted by the fourth fitting parameter  $\eta$ . The gCP parameters were fitted in a least-squares sense against counterpoise correction data obtained by the scheme of Boys and Bernardi<sup>[39]</sup> as described in the original publication.<sup>[30]</sup> This way, for each combination of a Hamiltonian (HF or DFT) and a basis set, a specific set of parameters  $\alpha$ ,  $\beta$ ,  $\sigma$ , and  $\eta$  was created. We found that this gCP correction performs particularly well for HF in combination with a small basis set. For further details and recent applications see Refs. [30, 40].

The last term  $E_{\text{SRB}}$  is a short-ranged correction to deal with basis set deficiencies which occur when using small or minimal basis sets. It corrects for systematically overestimated covalent bond lengths for electronegative elements and is again calculated as a sum over all atom pairs:

$$E_{\text{SRB}} = -s \sum_A^{\text{atoms}} \sum_{A \neq B}^{\text{atoms}} (Z_A Z_B)^{3/2} \exp\left(-\gamma \left(R_{\text{AB}}^{\text{D3}}\right)^{3/4} R_{\text{AB}}\right) \quad (4)$$

Here,  $R_{\text{AB}}^{\text{D3}}$  are the default cutoff radii as determined *ab initio* for the D3 dispersion correction scheme<sup>[23]</sup> and  $Z_A$ ,  $Z_B$  are

the nuclear charges. The correction is applied for all elements up to argon. The empirical fitting parameters  $s=0.03$  and  $\gamma=0.7$  were determined to produce vanishing HF-3c total atomic forces for the B3LYP-D3(BJ)/def2-TZVPP equilibrium structures of 107 small organic molecules. The other two correction terms were included in the fitting procedure of  $E_{\text{SRB}}$ , which was carried out by minimizing the HF-3c RMS gradient for the reference geometries. The D3 and gCP parameters were kept constant at their previously optimized values in this procedure. Because the SRB correction also effects covalent bond energies, the thermochemical properties of HF-3c are different from those of HF-D3-gcp/MINIX. Some cross-checking for standard reaction energies of organic molecules showed that HF-3c performs reasonably well, but further tests which are out of the scope of this work should be conducted to validate this finding.

In summary, the HF-3c method consists of only nine empirical parameters, three for the D3(BJ) dispersion, four in the gCP scheme, and two for the SRB correction. Because the fits are done independently, this parametrization procedure was easy to perform and changes in the setup of the fit are not expected to have any major effect on the method. No element or pair-specific terms need to be determined, that is, the nine parameters apply globally for all elements considered (i.e., currently up to xenon). Total energies and 3c-components for a few molecules are given in the Supporting Information.

## Technical details

All HF/MINIX and B3LYP<sup>[41,42]</sup>-D3(BJ)/def2-TZVPP<sup>[35]</sup> calculations were performed using TURBOMOLE 6.4.<sup>[43]</sup> In case of B3LYP, the RI approximation for the Coulomb integrals<sup>[44]</sup> was applied using matching default auxiliary basis sets.<sup>[45]</sup> The numerical quadrature grid *m4* was employed for integration of the exchange-correlation contribution. The 3c-terms to energy and analytical gradient were calculated by a new code which basically merges the freely available programs *dftd3* and *gCP*.<sup>[46]</sup> For both, HF and DFT, computations of the harmonic vibrational frequencies were performed analytically using the *aoforce* code from TURBOMOLE. The 3c-contributions to the Hessian are computed numerically by two-point finite differences of analytical gradients.

All PM6<sup>[13]</sup> and PM6-DH2<sup>[47]</sup> calculations were undertaken using MOPAC 2012<sup>[48]</sup> for the calculation of energies and gradients but the *relax* or *statpt* codes from TURBOMOLE 6.4 for executing the geometry relaxation steps. Vibrational frequencies were computed numerically using MOPAC 2012.

The COSMO-RS model<sup>[49,50]</sup> was used as implemented in COSMOtherm<sup>[51]</sup> to obtain all solvation free enthalpies. Single point calculations on the default BP86<sup>[41,52]</sup>/def-TZVPP<sup>[53]</sup> level of theory were performed on the optimized gas phase geometries. All visualizations of molecules were done with USCF Chimera version 1.6.1.<sup>[54]</sup> The root mean square deviation (RMSD) of two geometries was calculated using a quaternion algorithm<sup>[55]</sup> in order to get an all atom best-fit. The HF-3c method has also been implemented into the upcoming version of the free ORCA software<sup>[56]</sup> where it is invoked simply by keyword.

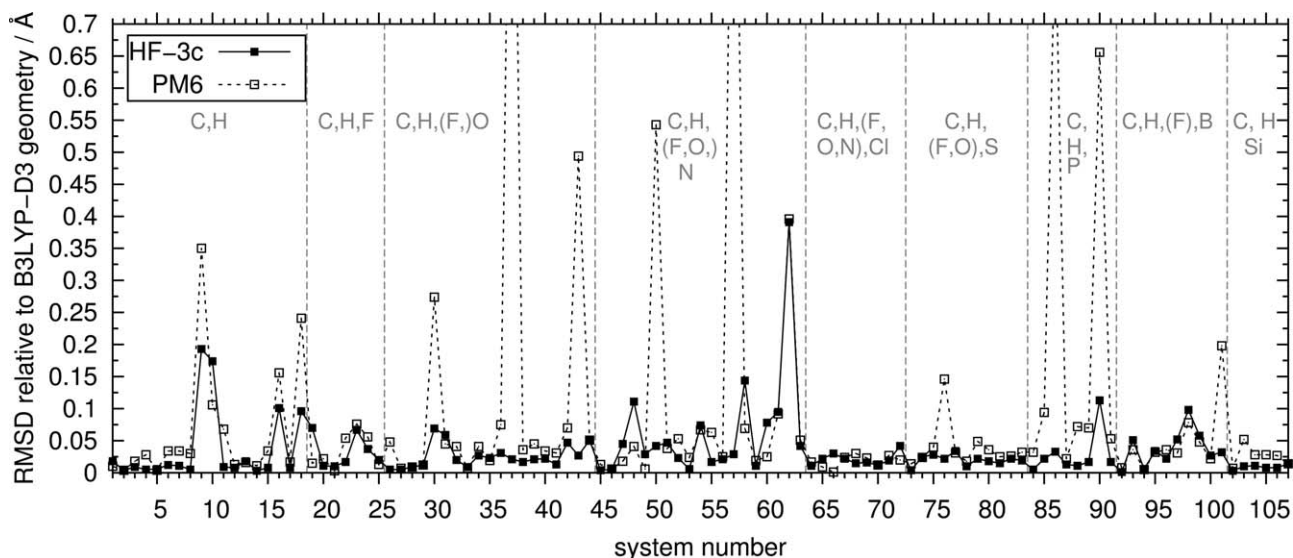


Figure 1. RMSD between HF-3c or PM6 and B3LYP-D3/def2-TZVPP geometries for 107 small organic molecules. The molecules are sorted according to the type of atoms and hence to the functional groups they contain. The atoms given in brackets are only rarely represented in the corresponding group. The lines between the data points are drawn just to guide the eye.

### Computation of free enthalpies of association

Free enthalpies of association for host and guest molecules in a solvent  $X$  at a temperature  $T$  are calculated as

$$\Delta G_a = \Delta E + \Delta G_{\text{RRHO}}^T + \Delta \delta G_{\text{solv}}^T(X). \quad (5)$$

Here,  $\Delta E$  denotes the gas phase interaction energy of the fully optimized molecules and  $G_{\text{RRHO}}^T$  is the sum of thermal corrections from energy to free enthalpy within a rigid-rotor-harmonic-oscillator approximation for each molecule in the gas phase at a given temperature  $T$ , including the zero-point vibrational energy. All harmonic frequencies are scaled with a factor of 0.86 for HF-3c. For obtaining the vibrational entropy, low-lying modes below  $\approx 100 \text{ cm}^{-1}$  are treated within a rigid-rotor model in order to reduce their error in the harmonic approximation, for details see Ref 6. The solvation free enthalpy  $\delta G_{\text{solv}}^T(X)$  is calculated for each gas-phase species by employing the COSMO-RS model.<sup>[49,50]</sup> No further (empirical) corrections are applied and the so computed values can be directly compared to experimental data.

## Results and Discussion

### Geometries of small organic molecules

The fitting set for the SRB correction of basis set deficiencies consists of 107 small organic molecules (2 to 34 atoms) containing the elements H, B, C, N, O, F, Si, P, S, and Cl. All standard functional groups are represented within this test set (for a detailed list of molecules see Supporting Information). The B3LYP-D3(BJ)/def2-TZVPP geometries, which have been proven to be reliable for organic molecules, were used as reference structures in the fitting procedure. PM6 calculations were performed to compare the HF-3c results to those from a widely used semiempirical approach.

Geometry optimization of these organic molecules using the final 3c-parameters yield an average RMSD between the HF-3c and B3LYP-D3 cartesian coordinates of 0.033 Å. This is considered to be a very good result meaning that at least for the fit set HF-3c yields structures of almost B3LYP/large-basis quality. The RMSD values for the individual molecules are shown in Figure 1. One of the rare “outliers” with a notably higher RMSD (adenine, 63) merely shows a methyl group rotated by 180° compared with the reference structure. PM6 shows more “outliers” than HF-3c and the average RMSD of 0.910 Å is much larger. Also the PM6 geometries of adenine as well as methyl acetate (43) exhibit a rotated methyl group. Furthermore, hydrogen peroxide (30) is planar whereas glyoxal (37) and urea (58) are not as they should be. Hydrazine (50), diphosphane (87), and  $\text{PH}_2\text{NH}_2$  (91) adopt the anti instead of the gauche conformation when optimized with PM6. These drastic conformational changes do not occur in optimizations with the HF-3c method.

Comparison of the lengths for the most frequent bonds (C–C, C=C, conjugated C–C/C=C, C–H, O–H, N–H, P–H, B–H, C–F, C=O, C–O, C–N, conjugated C–N/C=N, C–S, C–Cl, C–B and C–Si) results in an overall mean deviation (MD) with respect to the reference structures of 0.012 Å for HF-3c and 0.005 Å in case of PM6. With a few exceptions (C=C, B–H and C–F) the HF-3c bond lengths tend to be slightly too long. The mean absolute deviation (MAD) for all considered bond lengths in HF-3c and PM6 structures is 0.015 Å and 0.016 Å respectively. Hence, the overall error for bond lengths is similar for both methods. Due to a better description of bond angles and dihedral angles, HF-3c geometries generally show smaller RMSD values than PM6 structures.

The accuracy as demonstrated above also results from the SRB correction. This is more clearly seen by comparing some critical bond lengths with and without this term in typical

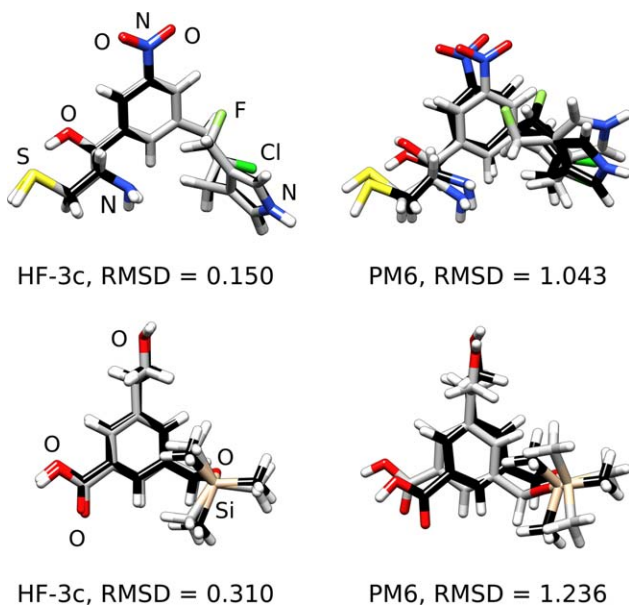
molecules. For example the C=O bond length in a ketone like acetone is 1.268 Å at the HF-D3-gCP/MINIX level (1.264 Å at HF/MINIX) which is too long by about 0.06 Å. This systematic deviation is corrected with HF-3c and the computed length of 1.206 Å is sufficiently close to the B3LYP reference value of 1.209 Å. Another example is hexafluoroethane where the corresponding values for the C—F bond length are at 1.429 Å the HF-D3-gCP/MINIX level (1.413 Å at HF/MINIX) and 1.343 Å at the HF-3c level (1.334 Å at B3LYP). A few more comparisons are given in Table 2 where in general the strong influence is seen for several bonds in polar situations.

**Table 2.** Critical bond lengths for some exemplary molecules at the HF/MINIX, HF-D3-gCP/MINIX, HF-3c and B3LYP-D3/def2-TZVPP level.

Molecule	Bond	R(HF/MINIX)	R(HF-D3-gCP/MINIX)	R(HF-3c)	R(B3LYP-D3)
Acetone	C=O	1.264	1.268	1.206	1.209
Urea	C=O	1.275	1.280	1.216	1.218
Methanimine	C=N	1.294	1.298	1.260	1.264
Ethanol	C—O	1.478	1.486	1.428	1.428
Urea	C—N	1.423	1.427	1.397	1.372
Hexafluoroethane	C—F	1.413	1.429	1.343	1.334
H <sub>2</sub> S <sub>2</sub>	S—S	2.132	2.136	2.122	2.073

All distances are given in Å.

As a cross-validation, two artificial neutral organic molecules containing a few heteroatoms were constructed in a more or less arbitrary fashion and fully optimized with all three methods taking again B3LYP-D3(BJ)/def2-TZVPP as reference. The RMSD relative to the reference structure is 0.15 Å for HF-3c and 1.043 Å for PM6 in case of the first molecule and 0.310 Å for HF-3c and 1.236 Å for PM6 in case of the second molecule (see Fig. 2). For both structures HF-3c performs significantly



**Figure 2.** Two artificially constructed organic molecules optimized with HF-3c (left grey structures) and PM6 (right grey structures). Black colored B3LYP-D3/def2-TZVPP geometries serve as reference. All RMSDs are given in Å.

better than PM6. Additionally, PM6 is not able to correctly describe the bond angle at the oxygen-atom of the silyl ether group in the second molecule but instead yields an almost linear coordination geometry.

Additionally, we performed single point calculations for 10 conformers of the tripeptide phenylalanyl-glycyl-glycine (PCONF set<sup>[57]</sup>), 15 conformers of the n-alkanes butane, pentane, and hexane (ACONF set<sup>[58]</sup>), 15 conformers of the sugar 3,6-anhydro-4-O-methyl-D-galactitol (part of the SCONF set<sup>[59]</sup>), and 10 conformers of cystein (CYCONF set<sup>[60]</sup>) as included in the GMTKN30 benchmark set.<sup>[61]</sup> The reference energies were taken from the original publications. For PCONF, SCONF, and CYCONF they were calculated on the coupled cluster with singles and doubles excitations and perturbative triples at the estimated complete basis set limit (CCSD(T)/CBS) level of theory and the ones for ACONF on the W1h-val level. The mean absolute deviation (MAD) for all conformational energies is 1.4 kcal/mol for HF-3c, which is a reasonable result in particular because this property is quite sensitive to the quality of the AO basis set. PM6-DH2 yields a much higher MAD of 2.8 kcal/mol while B3LYP-D3/def2-QZVP gives a much smaller MAD of 0.3 kcal/mol. The D3-correction contributes significantly to this good result, as plain B3LYP/def2-QZVP yields an MAD of 1.5 kcal/mol (i.e., is worse than HF-3c).

Further cross-validation studies for structures are performed on noncovalent complexes and their fragments as discussed in the next sections.

#### Geometries and interaction energies for S22 and S66 sets

In order to test the capability of the HF-3c method to describe noncovalent interactions, single-point calculations as well as geometry optimizations for the S22<sup>[62]</sup> and S66<sup>[17]</sup> test sets were carried out. Due to under representation of some interaction motifs, the S66 set was published by the Hobza group as a revised and extended version of the S22 set.<sup>[17]</sup> We also used their recently published X40 test set, which was designed to cover different halogen bonding interactions.<sup>[63]</sup> Reference values for interaction energies and geometries were taken from the original publications. The interaction energies refer to the estimated CCSD(T)/CBS level and the geometries were optimized on the MP2/cc-pVTZ(CP) or CCSD(T)/cc-pVTZ(noCP) level of theory.

Again, PM6 optimized geometries and interaction energies are used for comparison. Additionally, the DH2 correction<sup>[47]</sup> to PM6 for hydrogen-bonding and dispersion was employed which is mandatory for this kind of benchmark. Due to known problems with this correction for geometry optimizations, the scheme of calculating PM6-DH2 energies on PM6 geometries proposed by Hobza et al. was applied.<sup>[47,64]</sup>

For the S22 and S66 sets, the single-point HF-3c interaction energies are rather accurate with MADs of 0.55 kcal/mol and 0.39 kcal/mol, respectively (Table 3). These values are considerably lower than the previously published ones (0.64 and 0.51 kcal/mol) for HF/mini calculations applying just the D3 and gCP correction.<sup>[30]</sup> Thus, the modified basis set together with the SRB correction term and reparametrization gives a further

**Table 3.** MD and MAD for the single-point interaction energies of the S22, S66, and X40 test sets for the three methods HF-3c, PM6, and PM6-DH2.

	HF-3c		PM6		PM6-DH2	
	MD	MAD	MD	MAD	MD	MAD
S22	-0.01	0.55	3.39	3.39	0.13	0.39
S66	-0.09	0.38	2.68	2.68	0.35	0.65
X40	-0.80	1.44	1.19	1.73	0.35	1.46

All energies are given in kcal/mol.

significant improvement. This accuracy is comparable or even better than obtained for some density functionals at the DFT-D3/large-basis level.<sup>[18]</sup>

The MD values of -0.01 kcal/mol in case of S22 and -0.09 kcal/mol in case of S66 are almost insignificant. In case of the X40 test set both the MAD of 1.44 kcal/mol and the MD of -0.80 kcal/mol are much higher than for S22 (MAD of 0.55 kcal/mol) and S66 (MAD of 0.38 kcal/mol) but they are still reasonable for the applied theoretical level. In conclusion, it is clear that HF-3c is able to provide a qualitatively correct and quantitatively reasonable description of general noncovalent interactions. For a detailed analysis of responsible systematic error compensations see Ref. [30].

In contrast, PM6 single-point calculations result in equal values for the MD and MAD of 3.39 kcal/mol for the S22 and 2.68 kcal/mol for the S66 test set which indicates a systematical underbinding. This error can be reduced by applying the DH2 correction which accounts for dispersion and H-bonding. PM6-DH2 yields an MD of 0.13 kcal/mol and an MAD of 0.39 kcal/mol in case of the S22 and an MD of 0.35 kcal/mol and an MAD of 0.65 kcal/mol for the S66 set. Again, for the X40 set the deviations are much higher (MAD of 1.46 kcal/mol, MD of 0.35 kcal/mol). Altogether, the HF-3c method performs slightly better than PM6-DH2 in reproducing the interaction energies.

For the S22 set HF-3c geometry optimizations lead to an MD of 0.42 kcal/mol and an MAD of 0.94 kcal/mol for the interaction energies. Optimizations on the PM6 level of theory results in much higher values of 3.11 kcal/mol for both MD and MAD. Except for complex 10, which shows an imaginary vibrational mode for methyl rotation on the HF-3c level of theory, all optimized complexes are minima on the corresponding potential energy surface (PES) for both methods when started straightforwardly from the reference coordinates. In various cases, the convergence criteria for energy and gradient and the step size for the numerical PM6 frequency calculations had to be adjusted in order to remove small artificial imaginary frequencies. Similar numerical problems do not occur in HF-3c calculations. PM6-DH2 single-point calculations on PM6 geometries yield an MD of 0.1 kcal/mol and an MAD of 0.76 kcal/mol which are slightly lower than the corresponding values for HF-3c although the inconsistencies in the PM6 optimizations should be kept in mind.

Comparison of the resulting geometries with the reference structures yields an average RMSD of 0.21 Å in case of HF-3c and 0.45 Å for PM6. As shown in Figure 3(a), there are more

outliers for PM6 than for HF-3c geometries. The HF-3c geometries of both, the T-shaped benzene dimer (20) and the T-shaped benzene-indole complex (21) show structures in between a T-shaped and parallel-stacked one. The rings of two parallel stacked systems, namely the benzene dimer (11) and the benzene-indole complex (14), are rotated towards each other compared with the reference structures. Altogether, the general structural motifs of the S22 complexes can be reproduced well with HF-3c keeping in mind the flatness of the corresponding PES. In contrast, PM6 seems to systematically disfavor parallel stacked geometries. Instead of a parallel stacking the benzene dimer (11) shows a T-shaped stacking, the uracil dimer (14) an H-bonded geometry and the benzene-indole complex (14) a structure between parallel-stacked and T-shaped. Furthermore, the orientation of the monomers in PM6 optimized geometry of the methane dimer (8) differs from the one in the reference structure. Overall, the HF-3c geometries in the S22 set match the reference structures better than the PM6 ones.

The results for the S66 set reveal a similar picture. Geometry optimizations of the complexes yield an MD of 0.08 kcal/mol and an MAD of 0.59 kcal/mol for the interaction energy in case of HF-3c and again the same value for the MD and MAD of 2.33 kcal/mol for PM6. The PM6-DH2 single-point calculations on PM6 geometries result in an MD of 0.33 kcal/mol and an MAD of 0.81 kcal/mol which are slightly higher than the values for HF-3c. Similar to the S22 set there are more outliers for PM6 than for HF-3c geometries (Fig. 3b) compared to the reference. The average structural RMSD is 0.20 Å in case of HF-3c and 0.68 Å for PM6. All structures were proven to be minima on the corresponding PES though PM6 again shows problems with numerical noise. In general, HF-3c geometries reproduce the reference structures very well. The acetamide dimer (21) shows a rotated methyl group and the rings of the parallel stacked benzene-uracil complex (28) are differently rotated towards each other compared to the reference structures. In all cases, the basic interaction motifs are preserved in the HF-3c geometries which is a very important result.

PM6 geometries of the acetic acid dimer (20), acetamide dimer (21), and the ethyne-acetic acid complex (60) feature a rotated methyl group. As already observed for the S22 set PM6 prefers T-stacked geometries over parallel stacked ones. Almost every parallel stacked reference geometry shows T-shaped binding when optimized with PM6. Furthermore, the pyridine-uracil complex (29) shows an H-bonded geometry instead of parallel stacking and the H-bonded pyridine-methylamine complex (66) does not exhibit an H-bond at all.

Overall the HF-3c method reproduces the reference geometries of the S22 and S66 sets better than PM6. The RMSD is smaller and the general interaction motives are preserved in all cases indicating robustness in practical applications. The MDs and MADs for HF-3c interaction energies derived from optimized structures are similar to single-point values indicating that the HF-3c and reference PES are reasonably parallel to each other. The accuracy for HF-3c computed noncovalent interaction energies approaches that of dispersion corrected DFT but is less than the best DFT-D3/large-basis variants.

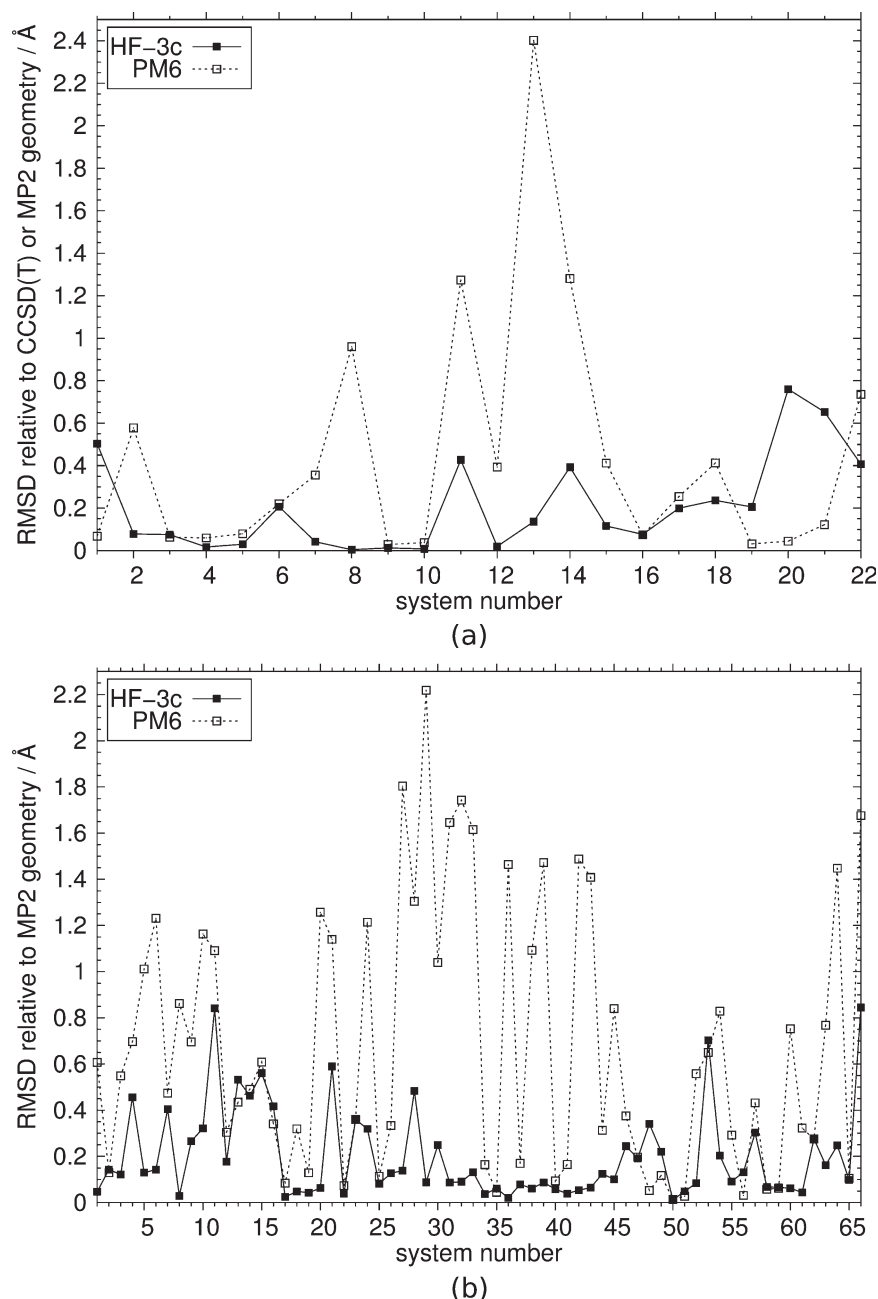


Figure 3. RMSD between HF-3c or PM6 and CCSD(T)/cc-pVTZ(noCP) or MP2/cc-pVTZ(CP) reference geometries for S22 (a) and S66 (b). The lines between the data points are drawn just to guide the eye.

### Thermal corrections to Gibbs free energies for small organic molecules and noncovalent complexes

Vibrational frequency calculations and the corresponding zero-point energy and thermal corrections to Gibbs free energies are supposed to be a main area of application of HF-3c. We randomly chose ten molecules out of 107 from the geometry fitting set, four complexes from S22, and six from the S66 test set. For these 20 molecules the  $E \rightarrow G(298)$  corrections were calculated using HF-3c, PM6, and B3LYP-D3/def2-TZVPP as reference. The scaling factors for the harmonic vibrational frequencies were set to 0.86 for HF-3c, 1.0 for PM6 and 0.97 for B3LYP. Low-lying modes below  $\approx 100 \text{ cm}^{-1}$  were treated within

a rigid-rotor model<sup>[6]</sup> in order to reduce their error in the harmonic approximation when obtaining the vibrational entropy. The final thermal corrections for all 20 molecules are listed in the Supporting Information.

Comparison of HF-3c with the B3LYP reference values shows a good agreement with an MD of 0.8 kcal/mol and an MAD of 1.9 kcal/mol (corresponding to about 3% relative error). For most molecules, the deviations range from only  $-1.3$  to 2.7 kcal/mol. The four molecules with the highest deviations are tetramethylsilane, the ethane-pentane complex, and the cyclopentane-neopentane complex where the HF-3c thermal corrections are 4.2 to 7.4 kcal/mol too large and the T-shaped benzene dimer for which the HF-3c value is 7 kcal/mol too

small. The large error for the benzene dimer can be attributed to the very shallow potential energy surface. In case of PM6, the thermal corrections for all regarded molecules except ammoniaborane are too small. The MD with respect to the B3LYP-D3/def2-TZPP values is  $-7.0$  and the MAD is  $7.2$  kcal/mol, that is, significantly worse than for HF-3c.

### Geometries and association free enthalpies of supramolecular complexes

Recently, we compiled a set of 12 supramolecular complexes (S12L set) and compared calculated free enthalpies of association with experimental data.<sup>[6]</sup> This set was very recently used to benchmark various dispersion corrections to DFT<sup>[15]</sup> and will be taken in this work for cross-validation of the HF-3c method on large realistic systems.

The investigated complexes are two “tweezer” complexes with tetracyanoquinone and 1,4-dicyanobenzene (1a and 1b measured in  $\text{CHCl}_3$ ),<sup>[65]</sup> two “pincer” complexes of organic  $\pi$ -systems (2a and 2b in  $\text{CH}_2\text{Cl}_2$ ),<sup>[66]</sup> the fullerenes  $\text{C}_{60}$  and  $\text{C}_{70}$  in a “buckycatcher” (3a and 3b in toluene),<sup>[67]</sup> complexes of an amide macrocycle (mcycle) with glycine anhydride and benzoquinone (4a and 4b in  $\text{CHCl}_3$ ),<sup>[68]</sup> complexes of cucurbit[6]uril (CB6) with butylammonium ( $\text{BuNH}_3$ ) and propylammonium ( $\text{PrNH}_3$ ) (5a and 5b in a 1:1 mixture of formic acid and water)<sup>[69]</sup> and complexes of cucurbit[7]uril (CB7) with a dicationic ferrocene derivative (FECF) and 1-hydroxyadamantane (6a and 6b in water).<sup>[70]</sup>

Computations at the PW6B95-D3(BJ)/def2-QZVP//TPSS-D3(BJ)/def2-TZVP level for gas phase interaction energies  $\Delta E$  together with a rigid rotor harmonic oscillator model for thermodynamical corrections  $\Delta G_{\text{RRHO}}$  and the COSMO-RS model for solvation free enthalpies  $\Delta\delta G_{\text{solv}}$  are able to reproduce the experimental values for association free enthalpies for these complexes with good accuracy. The MAD from experimental data was about 2 kcal/mol.<sup>[6]</sup> These results were used as a reference to test the performance of HF-3c for geometries and free enthalpies of association of the S12L set of supramolecular complexes. Again, PM6-DH2//PM6 calculations are performed for comparison.

Figure 4a shows the magnitudes of the contributions to the association free enthalpy ( $\Delta E$ ,  $\Delta G_{\text{RRHO}}$ , and  $\Delta\delta G_{\text{solv}}$ ) for HF-3c, PW6B95-D3//TPSS-D3 as reference and PM6 or PM6-DH2//PM6, respectively. The HF-3c gas phase interaction energy tends to be lower than the PW6B95-D3 energy, the deviation for the complexes 1a, 1b, 2a, 2b, 4a, 4b, and 6b is 0.5 to  $-2$  kcal/mol. For  $\text{C}_{60}$ @Catcher (3a) and  $\text{C}_{70}$ @Catcher (3b) HF-3c is overbinding by 5 to 6 kcal/mol, for  $\text{BuNH}_3$ @CB6 (5a) and  $\text{PrNH}_3$ @CB6 (5b) by 10 kcal/mol and for FECF@CB7 (6a) by 12.6 kcal/mol. The result for FECF@CB7 is not surprising since HF is known to describe transition metal complexes in general badly. Additionally, the complex has a double positive charge, which is challenging for a small basis set method due to large polarization effects. Consistent with this, the two complexes 5a and 5b with a larger error also carry a positive charge. These errors demonstrate that HF-3c is well-behaved and performs as expected.

Overall, the HF-3c gas phase interaction energies have an MD of  $-4.2$  and an MAD of 4.4 kcal/mol compared with the

PW6B95-D3//TPSS-D3 reference values. The MD indicates a small systematical overbinding and the MAD is similar to various dispersion corrected DFT methods employing large AO basis sets.<sup>[6]</sup>

All PM6 interaction energies are much higher than the reference values, the deviation ranges from 3 up to 30 kcal/mol. Applying the PM6-DH2//PM6 approach, the deviations decrease but remain larger than for HF-3c (6.1 kcal/mol compared to 4.4 kcal/mol). Exceptions are  $\text{C}_{70}$ @Catcher (3b) and FECF@CB7 (6a) with an error of  $-3.6$  and  $-7.6$  kcal/mol, respectively. Except for complexes 1a and 1b, PM6-DH2 overbinds and the MD ( $-5.6$  kcal/mol) is absolutely larger than for HF-3c.

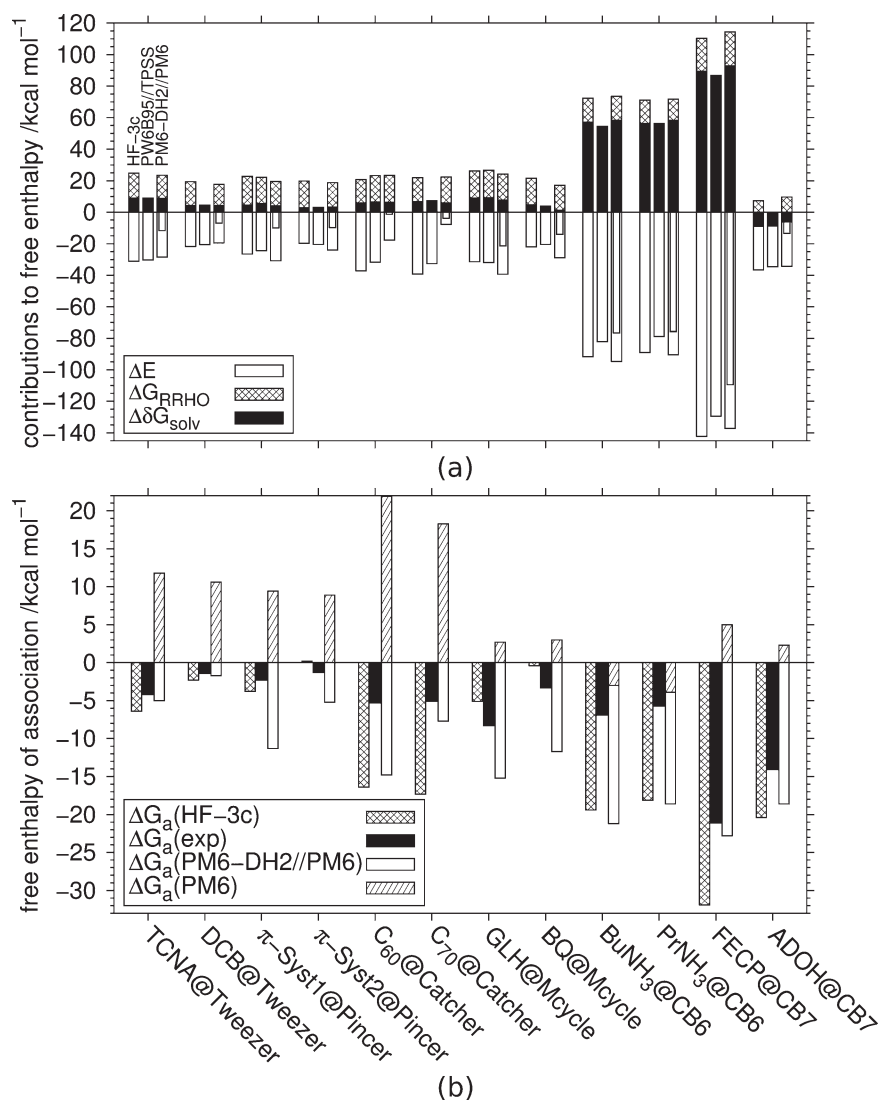
Comparison of the HF-3c geometries with the TPSS-D3 reference structures yield a minimal RMSD of 0.04 Å for the complex  $\text{C}_{60}$ @Catcher (3a) and a maximal RMSD of 0.48 Å for  $\pi$ -Syst1@Pincer (2a). The average RMSD is 0.19 Å. The corresponding values for PM6 are 0.11 Å, 0.97 Å and 0.45 Å. For both methods, the complexes  $\text{BuNH}_3$ @CB6 (5a) and  $\text{PrNH}_3$ @CB6 (5b) show a slightly different coordination of the guest molecule compared with the reference geometries. Similar to the small noncovalent complexes, the HF-3c method reproduces the reference structures better than PM6.

Since the geometry enters the COSMO-RS calculation, the better performance of HF-3c is also reflected in the solvation free enthalpies  $\Delta\delta G_{\text{solv}}$  of the complexes. The  $\Delta\delta G_{\text{solv}}$  values based on the HF-3c geometries deviate from the reference values in the range from only  $-0.5$  to  $+2.6$  kcal/mol whereas the deviation based on PM6 geometries ranges from  $-2.7$  to  $+6.1$  kcal/mol.

Because of the high computational cost, the thermodynamic correction  $\Delta G_{\text{RRHO}}$  on the TPSS-D3/def2-TZVP level of theory has been computed only for three complexes (2a, 3a, and 4a).<sup>[6]</sup> Both simpler methods match the three reference values relatively well. The highest deviation is 1.5 kcal/mol in case of HF-3c and 1.3 kcal/mol for PM6 corresponding to about 5–10% of  $\Delta G_{\text{RRHO}}$ . Because the number of comparisons is very small we can only guess that both methods might perform equally well.

The sum of all these contributions, the association free enthalpy  $\Delta G_a$ , is shown in Figure 4b in comparison to the experimental values. Since the gas phase interaction energy is the largest contribution and also most sensitive to the quality of the underlying electronic structure method, the error in  $\Delta G_a$  mainly reflects the error in  $\Delta E$ . Therefore, HF-3c yields  $\Delta G_a$  values which are too low (overbinding). Nevertheless, the calculated  $\Delta G_a$  values from HF-3c are surprisingly good regarding the simplicity of the method and an MD of  $-5.2$  and an MAD of 6.2 kcal/mol seems to be very respectable. The PM6-DH2//PM6 values are even lower and hence, the overbinding is even stronger than for HF-3c in most cases. The only significant exception is the complex FECF@CB7, whose  $\Delta G_a$  (PM6-DH2//PM6) matches the reference value much better than the HF-3c one. Since the HF-3c geometries are quite accurate and the derived values for  $\Delta\delta G_{\text{solv}}$  and  $\Delta G_{\text{RRHO}}$  in particular are reasonable, a single point DFT-D3/large-basis calculation on the HF-3c geometries is suggested for improved performance. For screening applications or scanning of supramolecular potential





**Figure 4.** a) Contributions to free enthalpy of association (interaction energy  $\Delta E$ , RRHO free enthalpy correction  $\Delta G_{RRHO}$  and solvation free enthalpy  $\Delta\delta G_{solv}$ ). PW6B95-D3/def2-QZVP//TPSS-D3/def2-TZVP values are taken from Ref. 6 and are shown for comparison. The left bar for each complex always presents the HF-3c values, the bar in the middle the PW6B95-D3//TPSS-D3 values and the right bar the PM6-DH2//PM6 (pure PM6 results for  $\Delta E$  are shown with narrower bars) values. Not all  $\Delta G_{RRHO}$  have been computed at the DFT level. b) Total free enthalpy of association  $\Delta G_a$  for all supramolecular complexes on the HF-3c, PM6 and PM6-DH2//PM6 levels of theory. Experimental values are taken from Refs. 65–70 and are shown for comparison.

energy surfaces, however, HF-3c seems to be sufficiently accurate.

### Geometries of small proteins

Recently, Martinez et al. composed a set of 58 small proteins with 5 to 35 residues in length and total charges ranging from  $-2$  to  $+2$ .<sup>[29]</sup> To test the performance of HF-3c, these proteins were fully optimized starting from the experimental geometries, which were taken from the Protein Databank (PDB).<sup>[71]</sup> Eight structures were excluded due to problems with the original PDB file (residues were missing or charges could not be assigned according to Ref. [29]). In case of multiple protein structures in one PDB file, the first one was always used. Again, PM6 optimizations were performed for comparison.

During the HF-3c geometry optimization procedure of almost all proteins, the charged termini of the protein

backbone neutralize via proton transfer from the protonated amino group to the carboxylate, if they are in close proximity or close to a lysine and aspartic or glutamic acid. This was also observed when two of those amino acids are too close. The protonation states and final charges were determined with USCF Chimera, which uses an empirical procedure for adding hydrogen atoms to the protein structure and AMBER ff99SB parameters<sup>[72]</sup> to assign the overall charge. Hence, it is not completely sure whether this is the same protonation state the protein would adopt in its natural environment. Six final HF-3c geometries (1T2Y, 2I9M, 2NX6, 2NX7, 2RLJ, 2RMW) exhibit a very small imaginary vibrational frequency below  $-22 \text{ cm}^{-1}$ , all other structures are true minima on the PES. In case of PM6, this hydrogen transfer is observed for only a few proteins. Contrary to the unproblematic HF-3c calculations, the PM6 optimization of ten proteins showed convergence problems which could not be solved. Additionally, 13 optimized

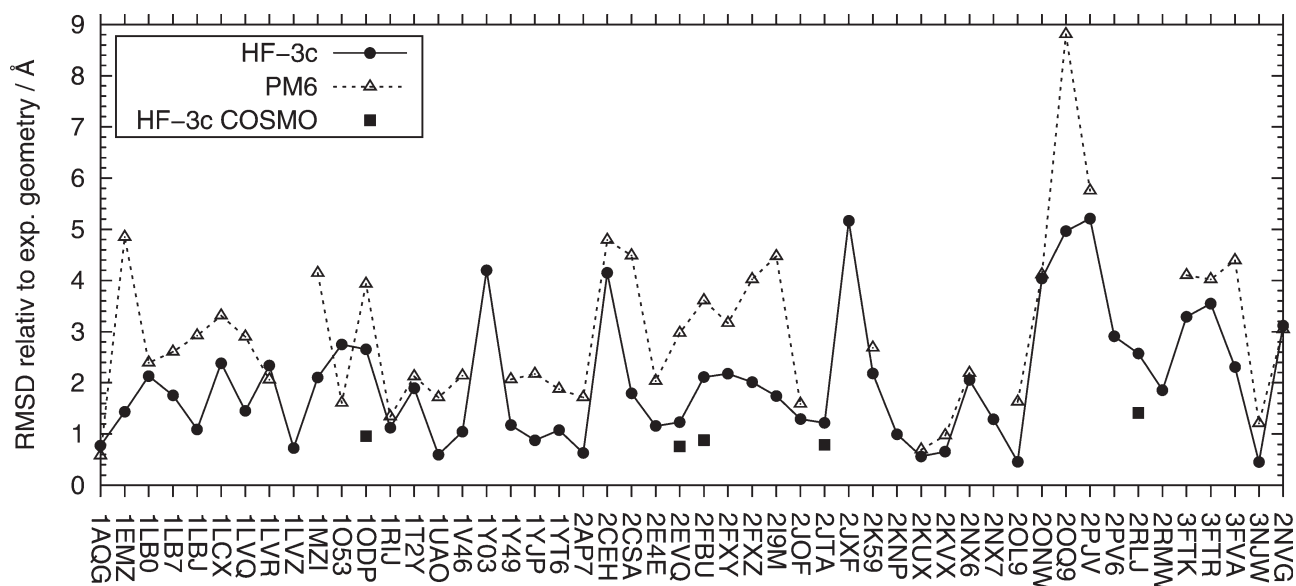


Figure 5. Backbone RMSD for all optimized protein structures on the HF-3c and PM6 level of theory relative to the experimental starting structure. The lines between the data points are drawn just to guide the eye.

structures exhibit persistent imaginary frequencies. Nevertheless, all structures also with imaginary frequencies are included in the geometry analysis.

As a first examination, the backbone RMSD between the calculated and the starting experimental geometries was evaluated using USCF Chimera.<sup>[54]</sup> The results are shown in Figure 5. All  $C^\alpha$  atom pairs were included, even if the calculated secondary structure strongly deviates from the reference one. In this way, the RMSD value gives a hint how good the computed secondary structure is. The minimal RMSD for the HF-3c geometries is 0.45 Å for 3NJW, the maximal value is 5.21 Å for 2PJV and the average RMSD is 2.02 Å. The average RMSD between different models of solution NMR structures in the whole set of 58 proteins is 1.73 Å.<sup>[29]</sup> Hence, the average RMSD for the HF-3c geometries is acceptable. In most cases the general secondary structure is preserved. Figure 6 shows four protein geometries with a very small RMSD in comparison to the experimental structures. We consider 13 protein structures which exhibit a backbone RMSD higher than 2.5 Å (arbitrarily chosen threshold) as some kind of outliers and these are now discussed in more detail.

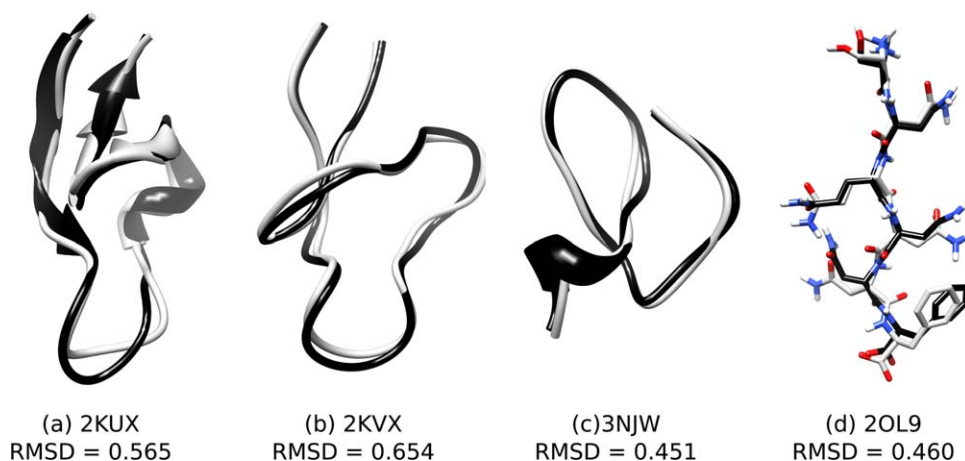
Figure 7 shows the HF-3c geometries of four proteins with a high RMSD and the experimental structure in comparison. The experimentally determined  $\alpha$ -helix of 1Y03 is bent but straight in the HF-3c calculation (Fig. 7a). The opposite applies for 2JXF (Fig. 7b) and 2OQ9, where the experimental structure exhibits a straight helix and the calculated geometry a bent one. In case of 2PJV (Fig. 7c), 2PV6, 1ODP, and 1O53 the  $\alpha$ -helix is strongly distorted compared with the experimental geometry. For 2ONW (Fig. 7d), 3FTK, 3FTR, and 3NVG the backbone of the experimental structures is more or less linear whereas it is folded in HF-3c optimized geometries. Protein 2CEH neither has a  $\alpha$ -helix nor a  $\beta$ -sheet structure and the HF-3c geometry is disordered in a different way than the experimental one. 2RLJ exhibits a larger helix part

when optimized with HF-3c compared to the experimentally obtained geometry.

In case of PM6, the minimal backbone RMSD is 0.58 Å for 1AQQ and the maximal value is 8.81 Å for 2OQ9. The average backbone RMSD of 2.96 Å is much higher than for the HF-3c optimized geometries. For more than half of the investigated proteins, the PM6 structure yields an RMSD larger than 2.5 Å and in most cases PM6 is not able to reproduce the general secondary structure.

Standard health checks to characterize the protein structures were used as described in Refs. [73–75]: (1) clashcores or steric overlaps greater than 0.4 Å per 1000 atoms, (2) percentage of bad side-chain dihedrals or rotamers, (3) number of  $\beta$ -carbon deviations greater than 0.25 Å from the expected position based on the backbone coordinates, (4) percentage of backbone dihedrals that fall into a favored region on a Ramachandran plot and (5) percentage of those, which are Ramachandran outliers, (6) percentage of bad bonds, and (7) percentage of bad angles. These health checks were performed for the calculated as well as the starting experimental structures. No structural improvements, for example, allowing Asn/Gln/His flips, were made. To provide one single number that represents the quality of a protein structure, the MolProbity score was defined as a logarithmic-weighted combination of clashcores, percentage of Ramachandran outliers and percentage of bad side-chain rotamers.<sup>[73]</sup> The averaged results are shown in Table 4, the individual values for each protein are provided in the supporting information.

The health check data for the HF-3c structures match the values obtained for the experimental geometries very well. The values for clashcores and bad angles are only slightly higher. The most defective health criterion is the percentage of bond outliers. Compared to the values published by Martinez et al.<sup>[29]</sup> for HF-D3/mini the application of the geometrical counterpoise correction and the additional short-range term in



**Figure 6.** HF-3c structures (gray) for four proteins with a small backbone RMSD in comparison to experimental ones (black). The RMSDs are given in Å. Hydrogens at carbon atoms in structure (d) are omitted for clarity.

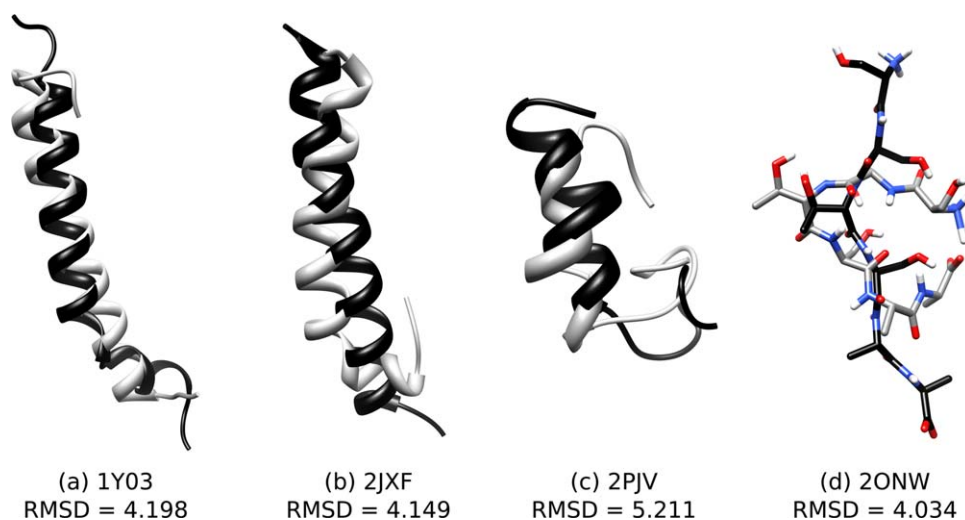
the HF-3c method gives an improvement for all health criteria. This is particularly obvious for the percentage of bond outliers, which is much smaller for the HF-3c geometries than for the ones obtained with HF-D3/mini. Compared to the results from the original publication for HF and the 6-31G basis set, the HF-3c health criteria are almost compatible. The highest deviation is found again in the percentage of bond outliers. Additionally, the number of clashcores is substantially smaller for HF/6-31G than for both HF-3c and experiment. Overall, we conclude that HF-3c is able to yield good geometries for the tested proteins. Because the method includes only minor empiricism and was not parameterized specifically for protein structures, we think that this conclusion holds in general and suggest it as a tool in structural biochemistry.

The health checks for PM6 geometries give worse results than those for HF-3c for most criteria. The number of clashcores and the percentage of poor rotamers is higher and the percentage of favored Ramachandran dihedrals is much smaller. The results for bond and angle outliers are slightly

better than for HF-3c but overall the PM6 structures are not as good as the HF-3c ones. Additionally, in many cases the positively charged guanidinium group of the amino acid arginine is not planar when optimized with PM6.

In general, HF-3c seems to predict too many hydrogen bonds (Fig. 8). On average, the calculation yields six hydrogen bonds too much compared to the corresponding experimental structures. PM6 shows on average four excessive hydrogen bonds. The hydrogen bond search was done with USCF Chimera<sup>[54]</sup> applying default criteria.

To test the influence of the solvent (i.e., artificially neglected water molecules) on the observed hydrogen transfer and the excess of hydrogen bonds, five proteins (1ODP, 2EVQ, 2FBU, 2JTA, and 2RLJ) were optimized with HF-3c using the COSMO model<sup>[76]</sup> for continuum solvation. The dielectric constant  $\epsilon$  was set to 78 for pure water. For all optimizations including COSMO, considerably less hydrogen transfers are observed. 1ODP and 2RLJ do not show a hydrogen transfer at all. For the other three proteins, the number of transferred hydrogens



**Figure 7.** HF-3c structures (gray) for four proteins with a high backbone RMSD in comparison to experimental ones (black). The RMSDs are given in Å. Hydrogens at carbon atoms in structure (d) are omitted for clarity. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

**Table 4.** Averaged health criteria for the HF-3c (50 proteins) and PM6 (41 proteins) optimized structures as well as the experimental starting geometries (50 proteins). Values for HF-D/mini and HF/6-31G were taken from Ref. 30] for comparison (all 58 proteins).

	Exp.	HF-3c	PM6	HF-D3/mini	HF/6-31G
Clashcore/1000 atoms	29	34	54	43	8
Bad side-chain rotamers	19%	13%	21%	18%	10%
C <sup>β</sup> deviations	0.2	0.2	0.0	0.5	0.3
Ramachandran outliers	5%	6%	8%	7%	3%
Ramachandran favored	81%	81%	71%	77%	86%
Bad bonds	0.5%	8%	3%	79%	1%
Bad angles	1%	4%	1%	10%	1%
MolProbity score	2.7	3.3	3.9	3.1	1.9

is reduced from two in case of 2EVQ and 2JTA and four in case of 2FBU to just one. Regarding the hydrogen bonds, only the 2FBU structure exhibits more H-bonds in the HF-3c-COSMO optimization than with plain HF-3c. The other four proteins exhibit two or three H-bonds less when optimized with COSMO. Nevertheless, the number of computed hydrogen bonds is still higher compared to the experiment. Because HF-3c performs very well for the structures and energies of all hydrogen bonded systems in S22 and S66, it is not clear in how far this conclusion is based on biased experimental data instead of errors of the theoretical model.

The geometries of all five proteins improve regarding all health checks when using COSMO in the optimization (for explicit values see Supporting Information). In particular, the number of clashcores is reduced and the percentage of Ramachandran favored dihedrals is increased. Also the backbone RMSD relative to the experimental geometry is much smaller, that is, it decreases by a factor of about two. The largest improvement was observed for 1ODP, its RMSD is reduced from 2.656 Å to only 0.956 Å. Thus, inclusion of the COSMO model in the optimization yields a further improvement to already good HF-3c protein “gas phase” structures.

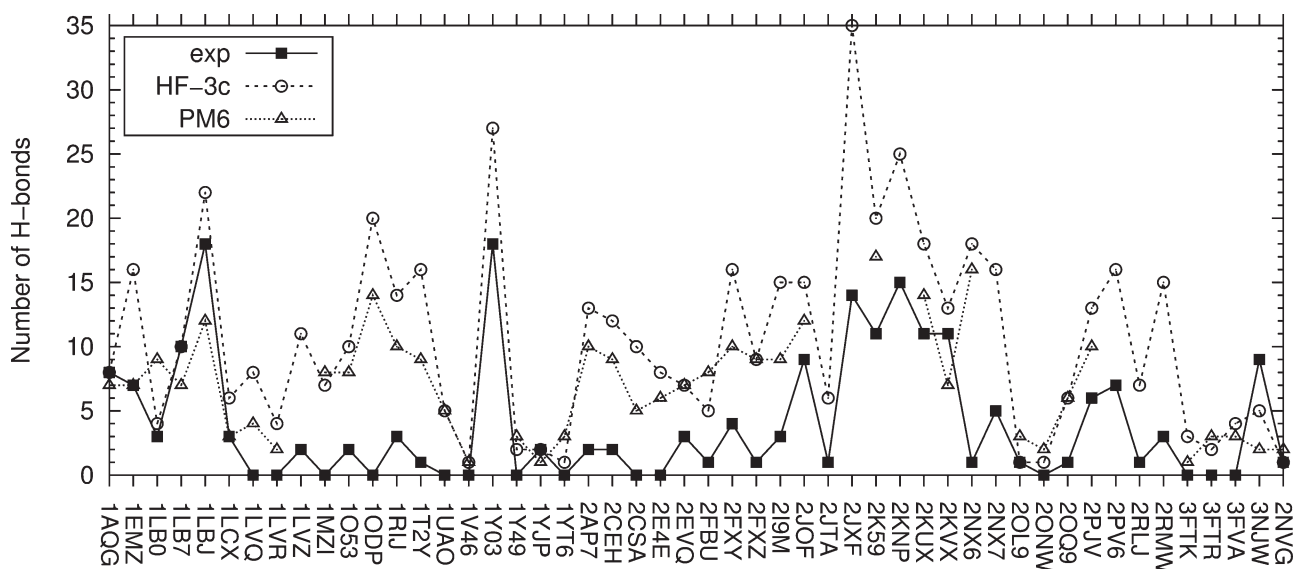
## Conclusions

A fast method based on a Hartree-Fock calculation with a small (in part minimal) basis set is presented (dubbed HF-3c from now on). Three corrections, namely the D3 scheme to include London dispersion, a geometrical counterpoise correction to handle intramolecular and intermolecular BSSE and a short-range term to correct basis set deficiencies for bond lengths are added to improve the plain HF energy. Detailed benchmarks for a variety of molecular properties were presented.

The method is able to yield good geometries for small covalently bound organic molecules, small noncovalent complexes as included in the S22 and S66 test sets as well as large supramolecular complexes. Fully optimized geometries of small proteins with up to 550 atoms yield good results in standard protein structure health checks and reasonable RMSD agreement compared to experimental structures.

By construction, the method gives a physically sound description of noncovalent interactions which is reflected in accurate interaction energies for a variety of systems. The MAD of the interaction energies compared with theoretical reference values is only 0.55 kcal/mol for the S22 and 0.38 kcal/mol for the S66 test set. For 12 supramolecular complexes, the fully *ab initio* computed association free enthalpy has an MAD of 6.2 kcal/mol with respect to experimentally obtained values. The MAD for the corresponding gas phase interaction energies is 4.4 kcal/mol. To put this into perspective, dispersion corrected DFT methods yield MADs in the range 2–5 kcal/mol while MP2/CBS yields an MAD of 16 kcal/mol<sup>[6]</sup> for the same set of realistic complexes. For the S66 set the MAD for the best DFT-D3/large basis variants and MP2/CBS are 0.2–0.3 and 0.45 kcal/mol, respectively.<sup>[18]</sup>

Compared to widely used semiempirical approaches (PM6 and PM6-DH2 used here as typical examples), the presented Hartree-Fock based method is slower but generally more



**Figure 8.** Number of hydrogen bonds for the experimental, HF-3c and PM6 protein structures. The lines between the data points are drawn just to guide the eye.


accurate, robust and numerically stable. It is easier to handle in large-scale geometry optimizations as shown by the protein studies. The method can be used routinely even on small desktop computers to optimize systems with hundreds of atoms and in parallel it can be applied to those with a few thousand atoms. Analytical vibrational frequency calculations are straightforward and the derived statistical thermodynamic corrections seem to be reasonable. Thus, the HF-3c methods might be able to fill the gap between semiempirical and DFT methods in terms of cost and accuracy and is recommended as a standard quantum chemical tool in biomolecular or supramolecular simulations. Current work in our laboratory investigates its applicability for the computation of molecular crystals.

## Acknowledgement

The authors thank Dr. Holger Kruse and Dr. Andreas Hansen for their help with the implementation of HF-3c into the ORCA program suit.

**Keywords:** Hartree-Fock · London dispersion energy · counterpoise-correction · noncovalent interactions · protein structures · supramolecular systems

How to cite this article: R. Sure, S. Grimme, *J. Comput. Chem.* **2013**, *34*, 1672–1685. DOI: 10.1002/jcc.23317

 Additional Supporting Information may be found in the online version of this article.

- [1] J.-M. Lehn, *Supramolecular chemistry: Concepts and perspectives*; VCH, Weinheim, **1995**.
- [2] J. L. Atwood, J. Steed, *Supramolecular Chemistry*, 2nd ed.; Wiley, **2009**.
- [3] S. Grimme, *WIREs Comput. Mol. Sci.* **2011**, *1*, 211.
- [4] J. Klimes, A. Michaelides, *J. Chem. Phys.* **2012**, *137*, 120901.
- [5] J. Antony, S. Grimme, *J. Comput. Chem.* **2012**, *33*, 1730.
- [6] S. Grimme, *Chem. Eur. J.* **2012**, *18*, 9955.
- [7] Y. Zhang, W. Yang, *J. Chem. Phys.* **1998**, *109*, 2604.
- [8] O. Gritsenko, B. Ensing, P. R. T. Schipper, E. J. Baerends, *J. Phys. Chem. A* **2000**, *104*, 8558.
- [9] S. Grimme, W. Hujo, B. Kirchner, *Phys. Chem. Chem. Phys.* **2012**, *14*, 4875.
- [10] E. Rudberg, *J. Phys. Condens. Matter* **2012**, *24*, 072202.
- [11] M. Gaus, A. Goez, M. Elstner, *J. Chem. Theory Comput.* **2013**, *9*, 338.
- [12] W. Weber, W. Thiel, *Theor. Chem. Acc.* **2000**, *103*, 495.
- [13] J. J. P. Stewart, *J. Mol. Mod.* **2007**, *13*, 1173.
- [14] J. R. Reimers, Ed., *Computational Methods for Large Systems*; Wiley: Hoboken, New Jersey, **2011**.
- [15] T. Risthaus, S. Grimme, *J. Chem. Theory Comp.* **2013**, *9*, 1580.
- [16] M. Korth, *Chem. Phys. Chem* **2011**, *12*, 3131.
- [17] J. Řezáč, K. E. Riley, P. Hobza, *J. Chem. Theory Comput.* **2011**, *7*, 2427.
- [18] L. Goerigk, H. Kruse, S. Grimme, *Chem. Phys. Chem* **2011**, *12*, 3421.
- [19] E. D. Murray, K. Lee, D. C. Langreth, *J. Chem. Theory Comput.* **2009**, *5*, 2754.
- [20] J. A. Pople, *Modern Theoretical Chemistry*, Vol. 4; Plenum: New York, **1976**.
- [21] E. R. Davidson, D. Feller, *Chem. Rev.* **1986**, *86*, 681.
- [22] W. Kolos, *Theor. Chim. Acta* **1979**, *51*, 219.
- [23] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
- [24] S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.* **2011**, *32*, 1456.
- [25] U. R. Fogueri, S. Kozuch, A. Karton, J. M. L. Martin, *J. Phys. Chem. A* **2013**, *117*, 2269.
- [26] A. Bauza, D. Quinero, P. M. Deya, A. Frontera, *Phys. Chem. Chem. Phys.* **2012**, *14*, 14061.
- [27] A. Antony, C. Hakanoglu, A. Asthagiri, J. F. Weaver, *J. Chem. Phys.* **2012**, *136*, 054702.
- [28] J. Granatier, M. Pitoňák, P. Hobza, *J. Chem. Theory Comput.* **2012**, *8*, 2282.
- [29] H. J. Kulik, N. Luehr, I. S. Ufimtsev, T. J. Martinez, *J. Phys. Chem. B* **2012**, *116*, 12501.
- [30] H. Kruse, S. Grimme, *J. Chem. Phys.* **2012**, *136*, 154101.
- [31] F. Jensen, *J. Chem. Theory Comput.* **2010**, *6*, 100.
- [32] A. Galano, J. R. Alvarez-Idaboy, *J. Comput. Chem.* **2006**, *27*, 1203.
- [33] H. Tatewaki, S. Huzinaga, *J. Comput. Chem.* **1980**, *1*, 205.
- [34] A. Schäfer, H. Horn, R. Ahlrichs, *J. Chem. Phys.* **1992**, *97*, 2571.
- [35] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
- [36] K. A. Peterson, D. Figgen, E. Goll, H. Stoll, M. Dolg, *J. Chem. Phys.* **2003**, *119*, 11113.
- [37] A. D. Becke, E. R. Johnson, *J. Chem. Phys.* **2005**, *123*, 154101.
- [38] E. R. Johnson, A. D. Becke, *J. Chem. Phys.* **2005**, *123*, 24101.
- [39] S. Boys, F. Bernardi, *Mol. Phys.* **1970**, *19*, 553.
- [40] H. Kruse, L. Goerigk, S. Grimme, *J. Org. Chem.* **2012**, *77*, 10824.
- [41] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.
- [42] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.
- [43] TURBOMOLE 6.4: R. Ahlrichs, M. K. Armbruster, M. Bär, H.–P. Baron, R. Bauernschmitt, N. Crawford, P. Deglmann, M. Ehrig, K. Eichkorn, S. Elliott, F. Furche, F. Haase, M. Häser, C. Hättig, A. Hellweg, H. Horn, C. Huber, U. Huniar, M. Kattannek, C. Kölmel, M. Kollwitz, K. May, P. Nava, C. Ochsenfeld, H. Öhm, H. Patzelt, D. Rappoport, O. Rubner, A. Schäfer, U. Schneider, M. Sierka, O. Treutler, B. Unterreiner, M. von Arnim, F. Weigend, P. Weis and H. Weiss. Universität Karlsruhe **2012**. See also: <http://www.turbomole.com>.
- [44] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, R. Ahlrichs, *Chem. Phys. Lett.* **1995**, *242*, 652.
- [45] F. Weigend, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.
- [46] Available at: <http://www.thch.uni-bonn.de/>. Last accessed May 6, 2013.
- [47] M. Korth, M. Pitoňák, J. Řezáč, P. Hobza, *J. Chem. Theory Comput.* **2010**, *6*, 344.
- [48] J. J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, **2012**. Available at: <http://OpenMOPAC.net>. Last accessed May 6, 2013.
- [49] A. Klamt, *J. Chem. Phys.* **1995**, *99*, 2224.
- [50] F. Eckert, A. Klamt, *AIChE J.* **2002**, *48*, 369.
- [51] F. Eckert, A. Klamt, COSMOtherm, Version C2.1, Release 01.11; COSMOlogic GmbH & Co. KG, Leverkusen, Germany, **2010**.
- [52] J. P. Perdew, *Phys. Rev. B* **1986**, *33*, 8822.
- [53] A. Schäfer, C. Huber, R. Ahlrichs, *J. Chem. Phys.* **1994**, *100*, 5829.
- [54] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605.
- [55] E. A. Coutsas, C. Seok, K. A. Dill, *J. Comput. Chem.* **2004**, *25*, 1849.
- [56] F. Neese, ORCA—an ab initio, density functional and semiempirical program package, Ver. 2.9 (Rev 0), Max Planck Institute for Bioinorganic Chemistry, Germany, **2011**.
- [57] D. Reha, H. Valdés, J. Vondrášek, P. Hobza, A. Abu-Riziq, B. Crews, M. S. de Vries, *Chem. Eur. J.* **2005**, *11*, 6803.
- [58] D. Gruzman, A. Karton, J. M. L. Martin, *J. Phys. Chem. A* **2009**, *113*, 11974.
- [59] G. I. Csonka, A. D. French, G. P. Johnson, C. A. Stortz, *J. Chem. Theory Comput.* **2009**, *5*, 679.
- [60] J. J. Wilke, M. C. Lind, H. F. Schaefer, A. G. Császár, W. D. Allen, *J. Chem. Theory Comput.* **2009**, *5*, 1511.
- [61] L. Goerigk, S. Grimme, *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670.
- [62] P. Jurecka, J. Sponer, J. Cerny, P. Hobza, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- [63] J. Řezáč, K. E. Riley, P. Hobza, *J. Chem. Theory Comput.* **2012**, *8*, 4285.
- [64] J. Řezáč, J. Janfřlík, D. Salahub, P. Hobza, *J. Chem. Theory Comp.* **2009**, *5*, 1749.
- [65] M. Kamieth, U. Burkert, P. S. Corbin, S. J. Dell, S. C. Zimmerman, F.-G. Klärner, *Eur. J. Org. Chem.* **1999**, 2741.
- [66] J. Gratton, J.-Y. Le Questel, B. Legouin, P. Uriac, P. van de Weghe, D. Jacquemin, *Chem. Phys. Lett.* **2012**, *522*, 11.

- [67] C. Mück-Lichtenfeld, S. Grimme, L. Kobryn, A. Sygula, *Phys. Chem. Chem. Phys.* **2010**, *12*, 7091.
- [68] C. Allott, H. Adams, C. A. Hunter, J. A. Thomas, P. L. Bernad Jr., C. Rotger, *Chem. Commun.* **1998**, 2449.
- [69] W. L. Mock, N. Y. Shih, *J. Am. Chem. Soc.* **1989**, *111*, 2697.
- [70] S. Moghaddam, C. Yang, M. Rekharsky, Y. H. Ko, K. Kim, Y. Inoue, M. K. Gilson, *J. Am. Chem. Soc.* **2011**, *133*, 3570.
- [71] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.* **1977**, *112*, 535.
- [72] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- [73] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, *Acta Cryst. D* **2010**, *66*, 12.
- [74] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, D. C. Richardson, *Nucleic Acids Res.* **2007**, *35*, W375.
- [75] I. W. Davis, L. W. Murray, J. S. Richardson, D. C. Richardson, *Nucleic Acids Res.* **2004**, *32*, W615.
- [76] A. Klamt, G. J. Schüürmann, *Chem. Soc. Perkin Trans.* **1993**, *2*, 799.

---

Received: 28 February 2013

Revised 28 March 2013

Accepted: 3 April 2013

Published online on 14 May 2013