Aalborg Universitet

AALBORG UNIVERSITY

Improving Solvation Energy Predictions Using The SMD Solvation Method and Semiempirical Electronic Structure Methods

Kromann, Jimmy Charnley; Svendsen, Casper Steinmann; Jensen, Jan Halborg

Improving Solvation Energy Predictions Using The SMD Solvation Method and Semiempirical Electronic Structure Methods

Jimmy C. Kromann[1], Casper Steinmann[2], and Jan H. Jensen[1]

[1]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark
[2]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

The PM6 implementation in the GAMESS program is extended to elements requiring d-integrals and interfaced with the conducter-like polarized continuum model (C-PCM) of solvation, including gradients. The accuracy of aqueous solvation energies computed using AM1, PM3, PM6, and DFTB and the SMD continuum solvation model is tested using the MNSOL data set. The errors in SMD solvation energies predicted using NDDO-based methods is considerably larger than when using DFT and HF, with RMSE values of 3.4-5.9 (neutrals) and 6-15 kcal/mol (ions) compared to 2.4 and ca 5 kcal/mol for HF/6-31G(d). For the NDDO-based methods the errors are especially large for cations and considerably higher than the corresponding COSMO results, which suggests that the NDDO/SMD results can be improved by re-parameterizing the SMD parameters focusing on ions. We found the best results are obtained by changing only the radii for hydrogen, carbon, oxygen, nitrogen, and sulfur and this leads to RMSE values for PM3 (neutrals: 2.8/ions: ca 5 kcal/mol), PM6 (4.7/ca 5 kcal/mol), and DFTB (3.9/ca 5 kcal/mol) that are more comparable to HF/6-31G(d) (2.4/ca 5 kcal/mol). Though the radii are optimized to reproduce aqueous solvation energies, they also lead more accurate predictions for other polar solvents such as DMSO, acetonitrile, and methanol, while the improvements for non-polar solvents are negligible.

Introduction

Accurate yet computationally efficient models of aqueous solvation represents an important challenge to molecular modeling. Continuum solvation models such as the polarized continuum model (PCM),[1] the conductor-like screening model (COSMO),[2] and the universal SMx models[3] offer a computational efficient model of solvation for molecules treated with electronic structure methods. While the accuracy of these and related methods have been studied extensively using DFT and wavefunction-based methods,[4, 5, 6, 7, 8, 9] there has been relatively little work on corresponding studies with semiempirical methods.[2, 10, 11, 12, 13] In this paper we study the accuracy of using AM1[14], PM3[15], PM6[16], and DFTB[17] together with the SMD[9] continuum solvation model using the MNSOL data set[18, 19] and present new parameters that increase the accuracy considerably for polar solvents. We chose the SMD method over methods with system-dependent radii, since SMD generally applicable to cases, such as transition states, where such a parameterization can become ill defined.

This manuscript is organized as follows. After a presentation of our computational methodology we discuss the accuracy of the semiempirical SMD calculations using the radii optimized for *ab initio* methods. We then compare several sets of reoptimized radii and select the best set. We show that the new radii lead to more accurate pKa predictions

Computational Details

For reference we use the MNSOL[18, 19] data set, which contains experimental solvation free energies of neutral and singly-charged molecules with atom-types of H, C, N, O, F, Si, P, S, Cl, Br, and I. Here we focus mainly on the aqueous solvation subset of that data set. Because iodine is not supported by 6-31G(d) we removed compounds test2018, test4001-4 and test4007-9 from the reference data set. Compounds c091 and i091 are radicals and are not included because PM6 is only implemented for RHF in GAMESS.[20] Our final reference set of molecules then consists of 522 aqueous solvation free energies of which 81 are for anions and 60 are for cations and we refer to this as the MNSOL data set.

When parameterizing SMD Coulomb radii, we use the a subset of the MNSOL data set (the SMD data set) which was also used by Marenich et al. [9]. This subset consists of 384 solvation free energies, of which 59 are for anions and 52 are for cations. Following the original SMD implementation we used mono-aquo microsolvated species for small ions with atoms carrying large (partial) charges. The SMD method is re-parameterized using the gas-phase structures optimized at the respective levels of theory.

The integration of d-integrals for semiempirical methods and interface to the conducter-like polarizable continuum model (C-PCM),[21, 22] is implemented in a locally modified version of GAMESS using integral code donated by James Stewart. The PM6/PCM interface follows the implementation of Steinmann et al.[23] except that we use the semi-numerical gradient approach for the PCM gradient that is also used for the gas phase gradient. The DFTB calculations are done using the DFTB/PCM interface developed by Nishimoto [24] in GAMESS and using version 3ob-3-1 of the 3OB DFTB parameter set [17, 25, 26, 27]. All SMD[9] calculations were carried out with the GAMESS program, and COSMO[2] calculations were done with MOPAC2016[28].

The original SMD parameterization was done with the Gauss-Bonet[29] tessellation scheme, while the default tessellation scheme in GAMESS is FIXPVA[30]. While this difference in tessellation scheme has a negligible effect on the solvation free energies of neutral molecules, it can have a rather larger effect for ions. Thus, for calculations involving the original SMD radii we use the Gauss-Bonet tessellation scheme (`mthall=1` in the `$tescav` input group), while we use the more numerically stable FIXPVA scheme when optimizing the radii. The original SMD parameterization was also done with integral-equation-formalism PCM (IEF-PCM)[31] while we chose the C-PCM method as it is computationally significantly more efficient for semiempirical calculations. COSMO is used with AM1, PM3 and PM6 as implemented in the MOPAC program[28]. The null model is constructed by setting all the predicted solvation free energies to the average of all the reference energies and the error bars reflect 95% confidence limits [32, 33, 34].

The pKa predictions are performed as described by Jensen et al. [35] using the same data set, i.e. a modified version of the one used by Eckert and Klamt [36]. Following Jensen

et al. [35], cefadroxil has been removed due to proton transfer and piroxicam has also been removed

Results and Discussion

Table 1 and Figure 1 shows the computed solvation free energy results obtained using HF/6-31G(d), AM1, PM3, PM6 and DFTB using the implicit solvation models SMD and COSMO using the MNSOL data set described in the Computational Details section. We compute the root mean square error (RMSE), mean signed error (MSE) and mean unsigned error (MUE) in order to properly quantify the accuracy of our results.
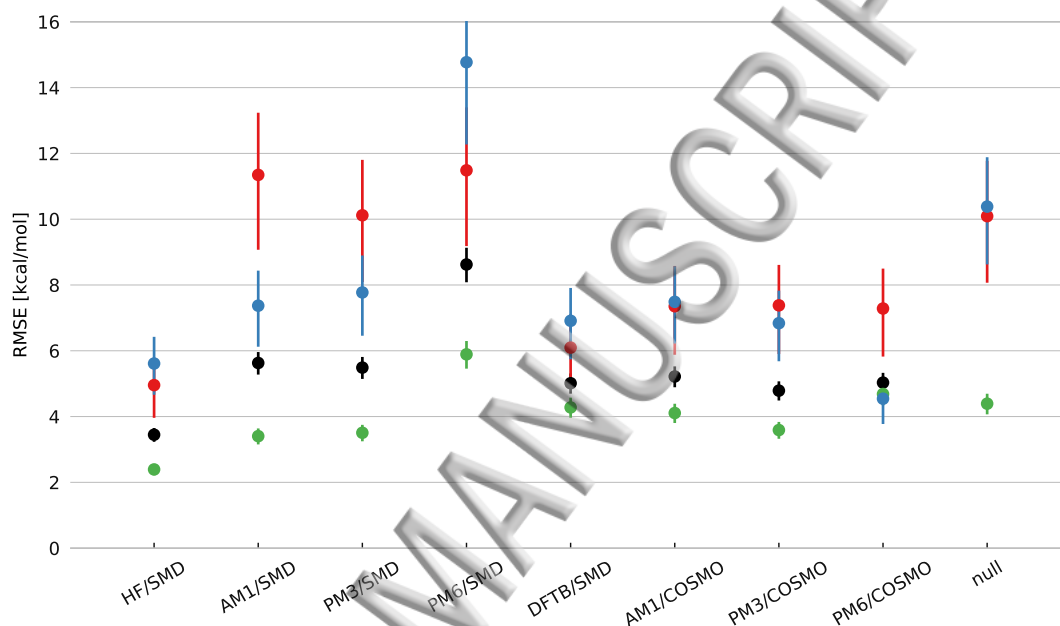

Figure 1: RMSE (in kcal/mol) of SMD aqueous solvation energies computed using HF/6-31G(d), AM1, PM3, PM6, and DFTB and the original SMD Coulomb radii for the MNSOL data set. Corresponding RMSE values for the COSMO method are included for comparison. The MNSOL data set is further split into the full set of molecules (black), neutral molecules (green), cations (red), and anions (blue).

We present the combined results (labeled "all") for all molecules in the data set, but also present individual observations on neutral and singly charged (both anions and cations) species.

SMD gives RMSE values of 3.4, 5.6, 5.5, 8.6 kcal/mol for HF, AM1, PM3, and PM6, respectively. For ions the RMSE values are 5.0 and 6.1 kcal/mol for cations, and 5.6 and 6.9 kcal/mol for anions for HF and DFTB respectively. For AM1, PM3 and PM6, the error is significantly higher with RMSE values of 11.4, 10.1, 11.5 kcal/mol for cations and 7.4, 7.8, 14.7 kcal/mol for anions, respectively. The NDDO-based method systematically underestimate the solvation free energy of anions with MSE (mean-signed-error) values of -5.1, -5.4, -12.6 kcal/mol, and overestimate the solvation free energy of cations with MSE values of 10.8, 9.2, 10.9 kcal/mol, respectively.

Table 1: RMSE, MUE, and MSE (in kcal/mol) of SMD aqueous solvation energies computed using HF/6-31G(d), AM1, PM3, PM6, and DFTB and the original SMD radii set for the MNSOL data set. COSMO solvation energies are included for comparison.

	SMD					COSMO			
	HF	AM1	PM3	PM6	DFTB	AM1	PM3	PM6	null
All									
RMSE	3.4	5.6	5.5	8.6	5.0	5.2	4.8	5.0	30.9
MUE	2.1	3.8	3.8	6.2	3.7	3.6	3.4	3.9	26.8
MSE	0.3	0.3	0.6	-3.2	-1.6	0.0	0.5	-1.8	-0.0
Max	17.8	17.9	24.5	30.2	18.3	20.1	21.3	21.1	87.5
Neutral									
RMSE	2.4	3.4	3.5	5.9	4.3	4.1	3.6	4.7	4.4
MUE	1.3	2.1	2.4	4.0	3.1	2.6	2.5	3.5	3.4
MSE	-0.9	-0.1	0.6	-3.5	-2.1	-1.8	-1.0	-3.4	0.0
Max	17.8	17.8	24.5	30.2	18.3	20.1	21.3	21.1	15.8
Cations									
RMSE	5.0	11.3	10.1	11.5	6.1	7.4	7.4	7.3	10.1
MUE	3.6	10.8	9.3	10.9	4.8	6.0	5.9	6.0	7.3
MSE	3.1	10.8	9.2	10.9	4.5	5.9	5.6	5.9	-0.0
Max	14.0	17.9	16.6	20.4	16.1	17.6	18.6	18.3	40.9
Anions									
RMSE	5.6	7.4	7.8	14.8	6.9	7.5	6.8	4.5	10.4
MUE	4.6	6.5	6.6	13.2	5.8	6.4	5.8	3.8	7.9
MSE	3.9	-5.1	-5.4	-12.6	-3.7	4.5	3.7	0.3	-0.0
Max	14.3	14.4	16.8	26.3	15.5	14.8	15.2	10.5	30.0

COSMO-predicted solvation free energies have RMSE values for the full set of 5.2, 4.8 and 5.0 kcal/mol for AM1, PM3, and PM6, respectively, which are quite similar to the corresponding SMD values for AM1 and PM3. However, the accuracy of COSMO for the ions are significantly better, especially for cations, with RMSE values of 7.3, 7.4 and 7.3 kcal/mol for cations and 7.5, 6.8, 4.5 kcal/mol for anions. The COSMO results show that is is possible to significantly improve the accuracy of the NDDO/SMD calculations and the comparatively high errors for ions indicate that the focus should be on the polar part of the solvation free energy.

Optimizing SMD Coulomb Radii

Based on the results in the previous subsection we optimize the Coulomb radii for PM3, PM6, and DFTB independently, while leaving the non-polar solvation free energy contribution unchanged. We use the Nelder-Mead simplex method[37] to optimize the radii using the SMD data set with the SciPy[38] package using a custom optimization script.[39] Optimization of the radii is done by minimizing the RMSE with respect to the radii.

Initial tests showed that the solvation free energy does not vary greatly with a change in radii for neutral molecules and the flatness of the energy surface caused problems for the Nelder-Mead method. Furthermore, the error for the ions is considerably larger than for the neutral molecules. Thus, only radii of the ions are optimized but when testing the performance of the new radii neutral molecules are included. This means that the optimum parameters obtained for the ions will not necessarily lead to the lowest possible error for neutral molecules. We therefore test the effect of optimizing a subset of the radii but include all molecules when testing the overall performance. These results are presented in Tables 2 and 3 and in Figure 2.

For PM3 the results are relatively insensitive to the type of atoms chosen with RMSE* values of 4.4 to 4.7 kcal/mol (Table 2 and Figure 2). (Note that the RMSE value are computed using different numbers of charged species so the RMSE* does not necessarily decrease when more parameters are optimized.) Similarly, the RMSE values computed using all the molecules in the MNSOL data set range from 2.9 to 3.2 kcal/mol as presented in Table 3. The cations are most sensitive to the choice of atom types with an RMSE range of 5.0 to 6.2 kcal/mol and the HCON- and HCONS-optimized radii set giving the best, and nearly identical, accuracy. For anions, the RMSE for the HCON radii (4.2 kcal/mol) is lower than that for HCONS (4.5 kcal/mol) and nearly identical to the lowest RMSE observed (4.1 kcal/mol). The HCON-optimized radii are therefore the best compromise.

For PM6, the RMSE* for HCO (3.2 kcal/mol) is about 1 kcal/mol lower than for the other radii set. However, the RMSE values computed using all molecules are very similar with a range of 3.6 to 3.8 kcal/mol, with HCO and HCON tied for first place. HCO does lead to the lowest error for neutral molecules and anions, but also an RMSE that is significantly larger for cations (6.2 kcal/mol), where PM6-HCONS performs best.

When testing the PM6-HCON and PM6-HCONS radii on pKa prediction (see below) we found that several solution-phase geometry optimizations failed because some X-H bonds

Table 2: Optimized Coulomb radii (in Å) for PM3, PM6 and DFTB for different atom types and original SMD radii parameters for comparison. $RMSE_{SMD}$ denotes the RMSE using the original parameters and RMSE* denotes RMSE after optimization. The RMSE is given in kcal/mol and is only calculated for the subset of ions in the SMD subset containing those specific atom types.

H	C	N	O	F	S	Cl	Br	$RMSE_{SMD}$	RMSE*
					SMD				
1.20	1.85	1.89	1.52	1.73	2.49	2.38	3.06		
					PM3				
0.29	1.72		1.65					9.6	4.4
0.42	1.75	1.76	1.64					10.9	4.3
0.62	1.76	1.72	1.62		2.58			10.7	4.6
0.58	1.75	1.72	1.64	1.79	2.55	2.53	3.35	10.6	4.7
					PM6				
0.66	1.87		1.73					12.3	3.2
0.17	1.85	1.78	1.74					13.0	4.1
0.39	1.78	1.82	1.75		2.62			12.7	4.3
0.70	1.82	1.78	1.74	1.40	2.55	2.37	3.88	12.3	4.4
					DFTB				
0.70	1.83	1.91	1.64					9.5	3.5

broke and the proton started moving into the solvent. This indicated that the H radius might be too small, leading to overpolarization of the H atom due to the solvent, and we tested several larger H radii for PM6-HCONS since the H radius was larger than for PM6-HCON. We found that the overpolarization problem disappeared with a H radius of 0.6 Å and that using this value has a negligible 0.1 kcal/mol effect of the accuracy of the solvation energies (HCONSh in Table 3). Thus, we use PM6-HCONSh radii set going forward.
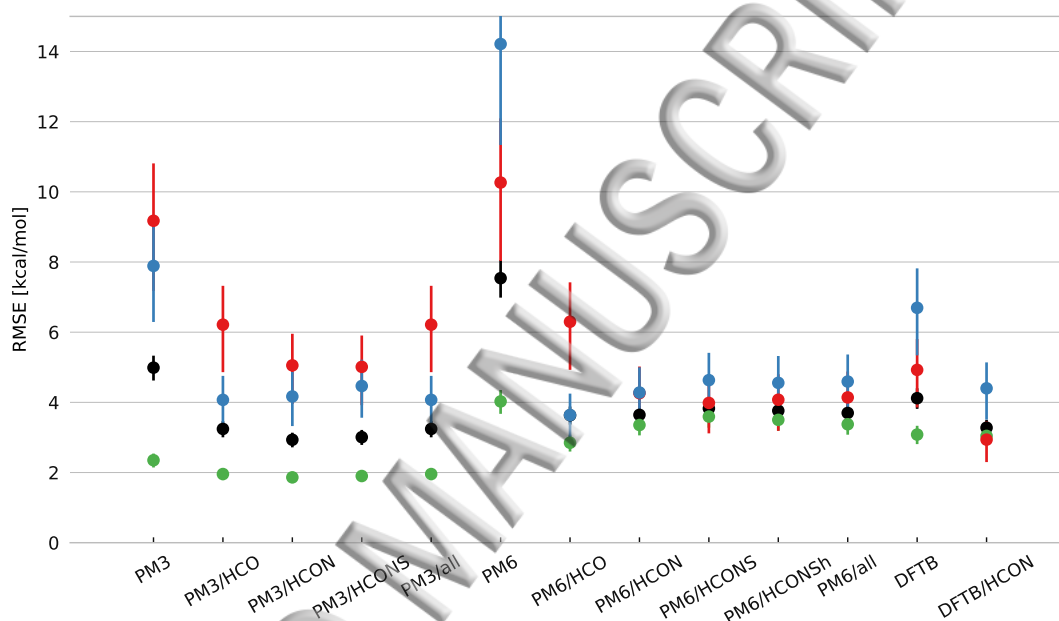
Figure 2: RMSE values (in kcal/mol) computed using SMD subset of the MNSOL data set and the Coulomb radii set shown in Table 2. HCO indicates that hydrogen, carbon, and oxygen radii parameters are changed (i.e. second and sixth entry in Table 2. The reference set is split into full set (black), neutral (green), cations (red), and anions (blue).

Table 3: RMSE, MUE and MSE values (in kcal/mol) computed using the SMD subset of the MNSOL data set and the Coulomb radii set shown in Table 2. HCO indicates that hydrogen, carbon, and oxygen radii parameters are changed (i.e. second and sixth row in Table 2.)

	PM3					PM6						DFTB	
	SMD	HCO	HCON	HCONS	all	SMD	HCO	HCON	HCONS	HCONSh	all	SMD	HCON
All													
RMSE	5.0	3.2	2.9	3.0	3.2	7.5	3.6	3.6	3.8	3.8	3.7	4.1	3.3
MUE	3.5	2.3	2.1	2.2	2.3	5.4	2.5	2.7	3.0	3.0	2.8	3.1	2.6
MSE	1.0	1.3	1.0	1.0	1.3	-2.4	-0.4	-1.3	-1.7	-1.6	-1.4	-0.9	-0.8
Max	16.8	11.6	11.3	12.6	11.6	22.6	18.4	18.6	16.9	16.9	17.6	15.5	10.9
Neutral													
RMSE	2.4	2.0	1.9	1.9	2.0	4.0	2.8	3.4	3.6	3.5	3.4	3.1	3.0
MUE	1.8	1.5	1.5	1.5	1.5	3.0	1.9	2.4	2.9	2.8	2.5	2.4	2.4
MSE	1.1	1.1	1.0	1.1	1.1	-2.6	-1.5	-2.1	-2.6	-2.6	-2.2	-1.1	-1.4
Max	8.7	8.0	7.2	7.4	8.0	16.4	18.4	18.6	16.9	16.9	17.6	10.8	10.9
Cations													
RMSE	9.2	6.2	5.1	5.0	6.2	10.3	6.3	4.3	4.0	4.1	4.1	4.9	2.9
MUE	8.5	5.5	4.0	3.9	5.5	10.0	5.8	3.6	3.4	3.5	3.5	4.0	2.4
MSE	8.4	5.3	3.2	2.5	5.3	10.0	5.8	2.9	3.0	3.1	2.8	3.6	0.2
Max	15.0	11.6	11.3	11.9	11.6	16.8	13.8	11.5	11.3	11.3	11.2	13.8	9.1
Anions													
RMSE	7.9	4.1	4.2	4.5	4.1	14.2	3.6	4.3	4.6	4.6	4.6	6.7	4.4
MUE	6.6	3.2	3.3	3.5	3.2	12.8	2.8	3.2	3.5	3.4	3.2	5.6	3.4
MSE	-6.1	-1.1	-1.3	-0.7	-1.1	-12.6	-0.6	-1.2	-1.4	-1.2	-1.3	-3.9	1.2
Max	16.8	9.2	9.9	12.6	9.2	22.6	8.8	12.1	14.9	14.7	17.3	15.5	10.4

8

From the results discussed above and presented in Table 2, optimizing for atom types beyond HCON either PM3 or PM6. Therefore the same procedure (optimizing HCON radii) was applied to DFTB. For DFTB the RMSE values of the full set of molecules, the subset of cation and the subset of anions goes from 4.1, 4.9 and 6.7 kcal/mol to 3.3, 2.9 and 4.4 kcal/mol, respectively.

In summary, we find that optimizing the SMD radii in combination with either PM3 or PM6 gives appreciable decreases in RMSE accuracy ≈ 7 kcal/mol in both cases compared to the reference results. We also find that no significant improvement is obtained when attempting to optimize the radii of other atoms than what was done with HCON subset of radii. For this subset, DFTB also saw an increase in accuracy for ions of about ≈ 6 kcal/mol. Because of this, we found the best choice was change the radii for hydrogen, carbon, oxygen and nitrogen (HCON) for PM3 and DFTB methods. For PM6 we optimized including sulfur and changed the radius of hydrogen to 0.6 Å to remove overpolarization problems. These methods are denoted as SMD† going forward.

Comparison with Other Solvation Methods

Using the new parameters for PM3/SMD†, PM6/SMD† and DFTB/SMD†, we compare the obtained RMSE with other approaches to obtain solvent free energies such as HF/6-31G(d)/SMD (one of the levels of theory used in the original SMD parameterization), PM3/COSMO and PM6/COSMO for the MNSOL data set (Table 4 and Figure 3).
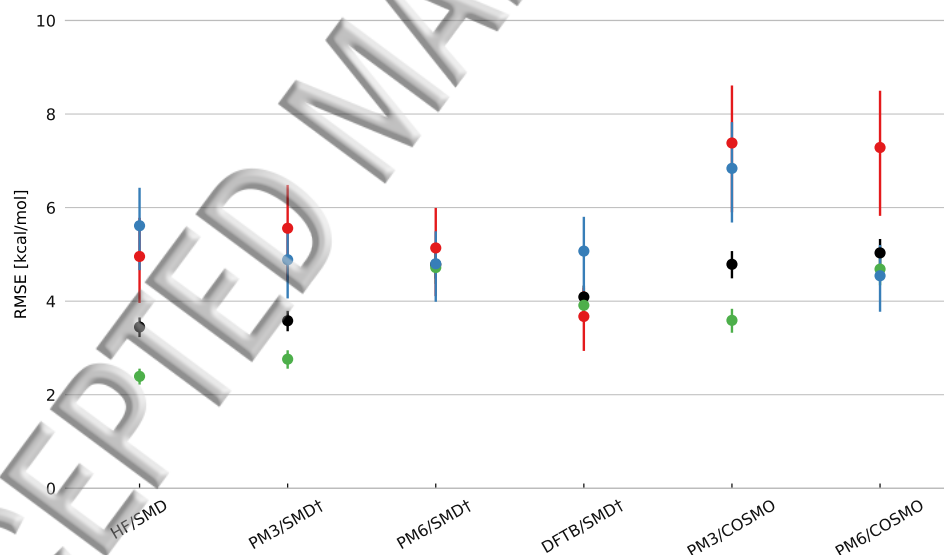

Figure 3: RMSE (in kcal/mol) of aqueous solvation energies computed with SMD using HF/6-31G(d), PM3, PM6, and DFTB for the MNSOL dataset. Note that † indicates that the Radii has been reoptimized. Corresponding RMSE values for the COSMO method are included for comparison. The dataset is split into full set (black), neutral (green), cations (red), and anions (blue).

The overall accuracy of PM3/SMD† is now similar to HF/SMD, although the RMSE value is 0.4 and 0.6 kcal/mol higher for neutral molecules and cations, while the RMSE

is 0.7 kcal/mol lower for anions. For PM6 the overall RMSE is 1.4 kcal/mol higher than HF, and the RMSE for cations is 2.8 kcal/mol higher. On the other hand, the

RMSE for cations is only 0.1 kcal/mol higher, while the RMSE for anions is 0.8 kcal/mol lower compared to HF. Finally, for DFTB the overall RMSE value (4.1 kcal/mol) is in between those for PM3 and PM6 as is the RMSE values for neutral molecules (3.9 kcal/mol). But for cations the RMSE (3.7 kcal/mol) is significantly lower than the other methods while the RMSE for anions (5.1 kcal/mol) is very similar.

While the solvation free energies of neutral molecules can generally be measured with an accuracy of ca 0.2 kcal/mol, the corresponding solvation free energies of ions are inferred from other experimental values with a typical accuracy about 3.0 kcal/mol.[40] Thus, the accuracy of the semiempirical SMD† is for all intents and purposes identical to HF/SMD for ions, but significantly worse for neutral ions in the case of PM6 and DFTB. The source of this error is most likely due to a poorer description of the electrostatic properties compared to HF. For example, the error in dipole moments for HCNO compounds is larger for PM6 than for PM3.[16]

Table 4 also lists values for PM3/COSMO and PM6/COSMO as implemented in MOPAC for comparison. This COSMO implementation only evaluates the polar part of the solvation free energy and will not be as accurate as properly parameterized COSMO-RS[5] values. For PM3, the SMD† results are significantly lower than the corresponding COSMO results, while for PM6 there is only a significant improvement for cations.

Table 4: RMSE, MUE and MSE (in kcal/mol) of aqueous solvation energies computed with SMD using HF/6-31G(d), PM3, PM6, and DFTB for the MNSOL dataset. Note that † indicates that the Radii has been reoptimized. Corresponding RMSE values for the COSMO method are included for comparison. The reference set is split into full set, neutral, cations, and anions.

	SMD	SMD†			COSMO	
	HF	PM3	PM6	DFTB	PM3	PM6
All						
RMSE	3.4	3.6	4.8	4.1	4.8	5.0
MUE	2.1	2.6	3.5	3.1	3.4	3.9
MSE	0.3	1.0	-1.7	-1.1	0.5	-1.8
Max	17.8	15.5	23.1	16.0	21.3	21.1
Neutral						
RMSE	2.4	2.8	4.7	3.9	3.6	4.7
MUE	1.3	2.0	3.4	2.9	2.5	3.5
MSE	-0.9	0.9	-2.9	-2.1	-1.0	-3.4
Max	17.8	15.5	23.1	16.0	21.3	21.1
Cations						
RMSE	5.0	5.6	5.1	3.7	7.4	7.3
MUE	3.6	4.4	4.2	2.9	5.9	6.0
MSE	3.1	3.6	3.8	1.0	5.6	5.9
Max	14.0	12.7	12.4	9.1	18.6	18.3
Anions						
RMSE	5.6	4.9	4.8	5.1	6.8	4.5
MUE	4.6	3.7	3.6	4.0	5.8	3.8
MSE	3.9	-0.3	-0.5	1.8	3.7	0.3
Max	14.3	14.6	14.7	14.7	15.2	10.5

pKa Prediction

Part of the impetus for this work were the relatively large error in pKa values predicted using PM3/COSMO compared to PM3/SMD with the original SMD radii observed by Jensen et al. [35], so we use the new SMD radii to compute the pKa value using the same data set. The data is presented in Figure 4 and Table 5 and shows that the RMSE drops from 1.5±0.3 to 0.9±0.3 pH units for PM3/SMD† and 2.0±0.4 to 1.6±0.3/0.4 pH units for DFTB/SMD†. In both cases the improvement is primarily a result of the decrease in mean error (ME), i.e the original radii led to a fairly systematic underestimation of the pKa values that has now been removed by the reparameterization. For DFTB the increased accuracy is also due to a larger number of low error predictions as is evident in Figure 4. In the case of PM3/SMD† the accuracy is now very similar to PM3/COSMO. The RMSE for PM6/SMD† is similar to PM6/COSMO and larger than for PM3/SMD in analogy with the corresponding results for COSMO.
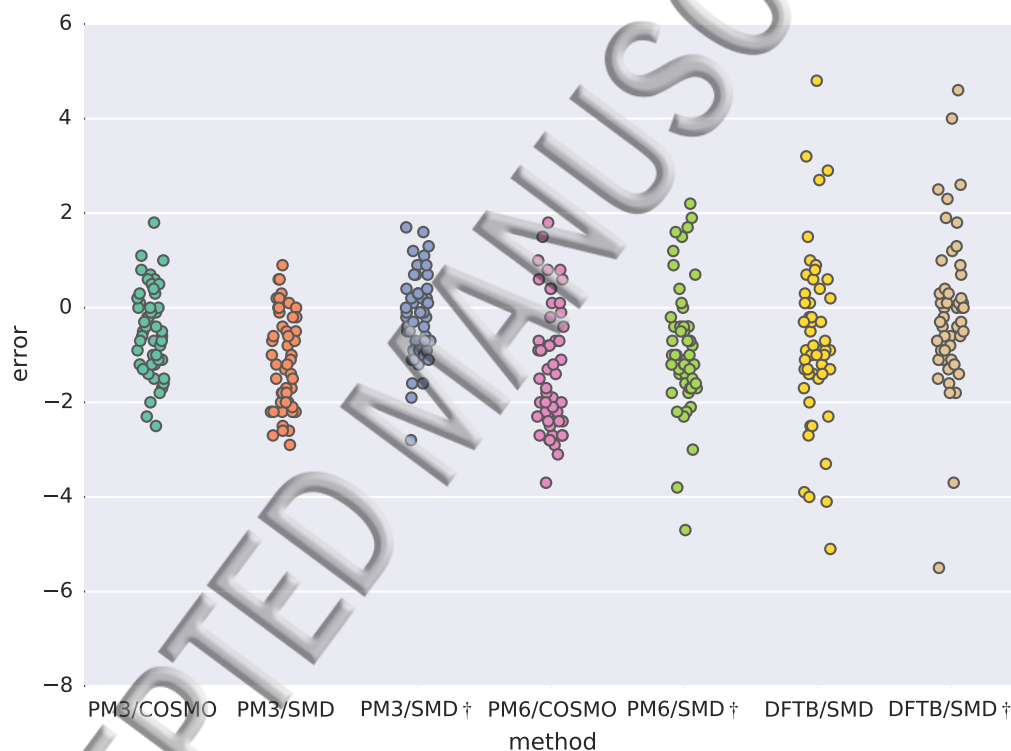

Figure 4: Plot of the errors in the predicted pKa values ($pK_a - pK_a^{\mathrm{exp}}$)

Table 5: Root-Mean-Square-Error (RMSE), man error (ME) and Pearson correlation (r) for predicted pKa values relative to experiment, together with their statistical uncertainty (95% confidence limits)

	PM3/COSMO	PM3/SMD	PM3/SMD†	PM6/COSMO	PM6/SMD†	DFTB/SMD	DFTB/SMD†
RMSE	1.0	1.5	0.9	1.8	1.6	2.0	1.6
95 % conf	± 0.2/0.2	± 0.3/0.3	± 0.2/0.2	± 0.3/0.4	± 0.3/0.4	± 0.4/0.4	± 0.3/0.4
ME	-0.5 ± 0.2	-1.2 ± 0.3	-0.2 ± 0.2	-1.3 ± 0.4	-0.8 ± 0.4	-0.7 ± 0.5	-0.1 ± 0.4
r	0.95	0.93	0.93	0.90	0.86	0.87	0.86
95 % conf	± 0.02/0.04	± 0.03/0.05	± 0.03/0.05	± 0.04/0.07	± 0.06/0.09	± 0.05/0.08	± 0.06/0.09

Non-aqueous Solvents

The new radii were optimized explicitly for water, so it is not *a priori* clear whether they offer more accurate results for other solvents. Table 6 shows the RMSE values for all, neutral, cations, and anions for the other polar solvents in the MNSOL data set, acetonitrile, DMSO and methanol for the original SMD radii and the corresponding re-optimized radii (†). In addition, corresponding combined RMSE values are given for the remaining 88 mostly non-polar solvents in the MNSOL data set, for which only data is available for neutral molecules. It is clear from this data that the new radii sets lowers the error of the predicted solvation free energies for the other polar solvents and has a negligible effect on the accuracy for the non-polar solvents.

As for acetonitrile and DMSO, the improvement is greatest for the ions where the RMSE decreases by 3-12 kcal/mol. The improvement is more modest (0.5-3 kcal/mol) for the neutral molecules. For methanol, which contains only ions, the RMSE is similarly decreased with 2-13 kcal/mol.

We thus recommend that the new radii optimized for water also be used for other polar solvents when doing semiempirical SMD calculations. For non-polar solvents either set of radii can be used.

Table 6: RMSE (in kcal/mol) of non-aqueous solvation energies computed with SMD using HF/6-31G(d), PM3, PM6, and DFTB for the MNSOL dataset. Note that † indicates that the radii has been reoptimized. Corresponding RMSE values for the COSMO method are included for comparison. The reference set is split into full set, neutral, cations, and anions.

	All	Neu	Cat	Ani
acetonitrile				
HF/SMD	11.2	2.9	10.5	13.1
PM3/SMD	19.4	4.1	16.2	24.6
PM3/SMD†	14.0	3.7	11.3	18.2
PM6/SMD	21.9	3.3	17.7	28.4
PM6/SMD†	13.0	1.2	11.4	16.0
DFTB/SMD	15.0	2.3	11.9	19.7
DFTB/SMD†	10.9	2.2	9.0	14.0
PM3/COSMO	13.0	3.1	12.8	14.7
PM6/COSMO	13.7	2.3	13.5	15.5
dimethylsulfoxide				
HF/SMD	11.9	3.0	10.6	12.6
PM3/SMD	20.2	4.2	15.3	21.4
PM3/SMD†	17.0	3.5	9.0	18.2
PM6/SMD	25.5	4.5	17.2	27.2
PM6/SMD†	17.7	1.7	9.9	18.9
DFTB/SMD	16.9	2.3	10.8	18.1
DFTB/SMD†	12.5	2.1	7.6	13.4
PM3/COSMO	12.5	3.4	10.7	13.2
PM6/COSMO	15.0	3.1	11.9	15.9
methanol				
HF/SMD	4.0	0.0	4.1	3.9
PM3/SMD	12.6	0.0	10.8	13.5
PM3/SMD†	6.6	0.0	6.0	6.9
PM6/SMD	16.3	0.0	12.0	18.3
PM6/SMD†	5.2	0.0	5.4	5.2
DFTB/SMD	8.4	0.0	6.3	9.4
DFTB/SMD†	4.7	0.0	4.4	4.9
PM3/COSMO	5.9	0.0	6.8	5.4
PM6/COSMO	5.4	0.0	6.7	4.6
other				
HF/SMD	4.2	4.2	0.0	0.0
PM3/SMD	5.1	5.1	0.0	0.0
PM3/SMD†	5.1	5.1	0.0	0.0
PM6/SMD	3.5	3.5	0.0	0.0
PM6/SMD†	3.4	3.4	0.0	0.0
DFTB/SMD	3.0	3.0	0.0	0.0
DFTB/SMD†	2.9	2.9	0.0	0.0
PM3/COSMO	3.6	3.6	0.0	0.0
PM6/COSMO	2.7	2.7	0.0	0.0

Conclusion

The PM6 method in GAMESS was extended to elements requiring d-integrals and interfaced with the polarized continuum model (PCM) of solvation, including semi-numerical gradients. However, the accuracy of aqueous solvation energies computed using AM1, PM3, PM6, and DFTB and the SMD continuum solvation model was tested using a subset of molecules from the MNSOL data set which showed that the errors in SMD solvation energies predicted using NDDO-based methods was considerably larger than when using DFT and HF, with RMSE values of 5.0 to 8.6 kcal/mol compared to 3.4 kcal/mol for HF/6-31G(d). For the NDDO-based methods the errors were especially large for cations and, in the case of PM6, also for anions, with RMSE values of 10.1 to 14.8 kcal/mol in comparison with to 5.0 to 5.7 kcal/mol for HF/6-31G(d)/SMD. Corresponding COSMO results (where the maximum RMSE is 7.5 kcal/mol) suggested that the NDDO/SMD results could be improved by re-parameterizing the SMD Coulomb radii for the NDDO methods.

The fact that the NDDO/SMD errors are largest for ions suggest that the problem is with the polar solvation energy, so we focus on optimizing the values of the Coulomb radii while leaving all parameters associated with the non-polar solvation free energy unchanged. We optimize the radii only for the ionic species on a subset of the MNSOL data set previously used to parametrize RHF/6-31G(d)/SMD, but include neutral molecules when testing how well SMD with the new Coulomb radii perform. We found the best results are obtained by changing only the radii for hydrogen, carbon, oxygen, nitrogen, and sulfur and this leads to RMSE values for PM3 (neutrals: 2.8/ions: ca 5 kcal/mol), PM6 (4.7/ca 5 kcal/mol), and DFTB (3.9/ca 5 kcal/mol) that are more comparable to HF/6-31G(d) (2.4/ca 5 kcal/mol). We note that the SMD parameterization is done using mono-aquo microsolvated species for small ions with atoms carrying large (partial) charges, while the MNSOL data set also contains data for the non-microsolvated equivalents. Comparison to the HF/6-31G(d) MSEs reported by Marenich et al. [8] reported for the selectively microsolvated ions indicates that the inclusion of non-microsolvated species leads an MSE increase of about 1 kcal/mol.

Though the radii are optimized to reproduce aqueous solvation energies, they also lead more accurate predictions for other polar solvents such as DMSO, acetonitrile, and methanol, while the improvement for non-polar solvents are negligible.

Supplementary Material

See supplementary material for Tables S1-S3 and Figures S1-S3, which are referred to in the text. In addition we also provide a table of dielectric constants used for the calculations and a table of aqueous solvation free energies obtained using the optimized parameters.

Acknowledgments

References

[1] J. Tomasi, B. Mennucci, and R. Cammi, "Quantum mechanical continuum solvation models," *Chem. Rev.*, vol. 105, pp. 2999–3093, 2005.

[2] A. Klamt and G. Schuurmann, "COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient," *J. Chem. Soc., Perkin Trans.*, vol. 2, pp. 799–805, 1993.

[3] C. J. Cramer and D. G. Truhlar, "A universal approach to solvation modeling," *Accounts of Chemical Research*, vol. 41, no. 6, pp. 760–768, jun 2008, doi: 10.1021/ar800019z.

[4] V. Barone, M. Cossi, and J. Tomasi, "A new definition of cavities for the computation of solvation free energies by the polarizable continuum model," *J. Chem. Phys.*, vol. 107, no. 8, pp. 3210–3221, aug 1997, doi: 10.1063/1.474671.

[5] F. Eckert and A. Klamt, "Fast solvent screening via quantum chemistry: COSMO-RS approach," *AIChE J.*, vol. 48, no. 2, pp. 369–385, feb 2002, doi: 10.1002/aic.690480220.

[6] C. Curutchet, A. Bidon-Chanal, I. Soteras, M. Orozco, and F. J. Luque, "MST continuum study of the hydration free energies of monovalent ionic species," *J. Phys. Chem. B.*, vol. 109, no. 8, pp. 3565–3574, mar 2005, doi: 10.1021/jp047197s.

[7] Y. Takano and K. N. Houk, "Benchmarking the conductor-like polarizable continuum model (CPCM) for aqueous solvation free energies of neutral and ionic organic molecules," *J. Chem. Theory Comput.*, vol. 1, no. 1, pp. 70–77, jan 2005, doi: 10.1021/ct049977a.

[8] A. V. Marenich, R. M. Olson, C. P. Kelly, C. J. Cramer, and D. G. Truhlar, "Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges," *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2011–2033, nov 2007, doi: 10.1021/ct7001418.

[9] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, "Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions," *J. Phys. Chem. B.*, vol. 113, no. 18, pp. 6378–6396, may 2009, doi: 10.1021/jp810292n.

[10] G. E. Chudinov, D. V. Napolov, and M. V. Basilevsky, "Quantum-chemical calculations of the hydration energies of organic cations and anions in the framework of a continuum solvent approximation," *Chem. Phys.*, vol. 160, no. 1, pp. 41 – 54, 1992, doi: 10.1016/0301-0104(92)87090-V.

[11] M. J. Negre, M. Orozco, and F. J. Luque, "A new strategy for the representation of environment effects in semi-empirical calculations based on dewar's hamiltonians," *Chem. Phys. Lett.*, vol. 196, no. 1, pp. 27–36, 1992.

[12] F. J. Luque, M. J. Negre, and M. Orozco, "An am1-scrf approach to the study of changes , no. 17, pp. 4386–4391, 1993.

[13] M. Caricato, B. Mennucci, and J. Tomasi, "Solvent effects on the electronic spectra: An extension of the polarizable continuum model to the zindo method," *J. Phys. Chem. A.*, vol. 108, no. 29, pp. 6248–6256, 2004.

[14] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, "Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model," *J. A. Chem. Soc.*, vol. 10, pp. 3902–3909, 1985, doi: 10.1021/ja00299a024.

[15] J. J. P. Stewart, "Optimization of Parameters for Semi-Empirical Methods I-Method," *J. Comp. Chem.*, vol. 10, pp. 209–220, 1989, doi: 10.1002/jcc.540100208.

[16] J. J. P. Stewart, "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements," *J. Mol. Model*, vol. 13, pp. 1172–1213, 2007.

[17] M. Gaus, Q. Cui, and M. Elstner, "DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)," *J. Chem. Theory Comput.*, vol. 7, no. 4, pp. 931–948, apr 2011, doi: 10.1021/ct100684s.

[18] A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. P. Winget, C. J. Cramer, and D. G. Truhlar, "Minnesota solvation database," 2012.

[19] "Minnesota solvation database," accessed November 24th, 2017. [Online]. Available: https://comp.chem.umn.edu/mnsol

[20] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J.A.Montgomery, "General atomic and molecular electronic structure system," *J. Comput. Chem.*, vol. 14, pp. 1347–1363, 1993.

[21] V. Barone and M. Cossi, "Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model," *The Journal of Physical Chemistry A*, vol. 102, no. 11, pp. 1995–2001, mar 1998, doi: 10.1021/jp9716997.

[22] M. Cossi, N. Rega, G. Scalmani, and V. Barone, "Energies, structures, and electronic properties of molecules in solution with the c-PCM solvation model," *Journal of Computational Chemistry*, vol. 24, no. 6, pp. 669–681, apr 2003, doi: 10.1002/jcc.10189.

[23] C. Steinmann, K. L. Blædel, A. S. Christensen, and J. H. Jensen, "Interface of the polarizable continuum model of solvation with semi-empirical methods in the gamess program," *PLoS ONE*, vol. 8, no. 7, p. e67725, 07 2013, doi: 10.1371/journal.pone.0067725.

[24] Y. Nishimoto, "DFTB/PCM applied to ground and excited state potential energy surfaces," *J. Phys. Chem. A*, vol. 120, no. 5, pp. 771–784, feb 2016, doi: 10.1021/acs.jpca.5b10732.

[25] M. Gaus, X. Lu, M. Elstner, and Q. Cui, "Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications," *J. Chem. Theory Comput.*, vol. 10, no. 4, pp. 1518–1537, apr 2014, doi: 10.1021/ct401002w.

[26] X. Lu, M. Gaus, M. Elstner, and Q. Cui, "Parametrization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 1062–1082, jan 2015, doi: 10.1021/jp506557r.

[27] M. Kubillus, T. Kubař, M. Gaus, J. Řezáč, and M. Elstner, "Parameterization of the DFTB3 Method for Br Ca, Cl, F, I, K, and Na in Organic and Biological Systems," *J. Chem. Theory Comput.*, vol. 11, no. 1, pp. 332–342, jan 2015, doi: 10.1021/ct5009137.

[28] J. J. P. Stewart, "MOPAC2012," 2016, Stewart Computational Chemistry, Colorado Springs, CO, USA. [Online]. Available: http://openmopac.net

[29] J. L. Pascual-Ahuir, E. Silla, J. Tomasi, and R. Bonaccorsi, "Electrostatic interaction of a solute with a continuum. improved description of the cavity and of the surface cavity bound charge distribution," *Journal of Computational Chemistry*, vol. 8, no. 6, pp. 778–787, sep 1987, doi: 10.1002/jcc.540080605.

[30] P. Su and H. Li, "Continuous and smooth potential energy surface for conductorlike screening solvation model using fixed points with variable areas," *The Journal of Chemical Physics*, vol. 130, no. 7, p. 074109, feb 2009, doi: 10.1063/1.3077917.

[31] J. Tomasi, B. Mennucci, and E. Cancès, "The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level," *Journal of Molecular Structure: THEOCHEM*, vol. 464, no. 1-3, pp. 211–226, may 1999, doi: 10.1016/s0166-1280(98)00553-3.

[32] A. Nicholls, "Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals," *J. Comput. Aided Mol. Des.*, vol. 28, no. 9, pp. 887–918, Sep 2014, doi: 10.1007/s10822-014-9753-z.

[33] A. Nicholls, "Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods," *J. Comput. Aided Mol. Des.*, vol. 30, no. 2, pp. 103–126, Feb 2016, doi: 10.1007/s10822-016-9904-5.

[34] J. H. Jensen, "Which method is more accurate? or errors have error bars," *PeerJ Preprints*, vol. 5, p. e2693v1, 2017, doi: 10.7287/peerj.preprints.2693v1.

[35] J. H. Jensen, C. J. Swain, and L. Olsen, "Prediction of pKa values for druglike molecules using semiempirical quantum chemical methods," *The Journal of Physical Chemistry A*, vol. 121, no. 3, pp. 699–707, jan 2017, doi: 10.1021/acs.jpca.6b10990.

[36] F. Eckert and A. Klamt, "Accurate prediction of basicity in aqueous solution with COSMO-RS," *Journal of Computational Chemistry*, vol. 27, no. 1, pp. 11–19, 2005, doi: 10.1002/jcc.20309.

[37] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965, doi: 10.1093/comjnl/7.4.308.

[38] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001, accessed November 24th, 2017. [Online]. Available: http://www.scipy.org/"

[39] J. Charnley, 2017, accessed November 24th, 2017. [Online]. Available: https://github.com/charnley/optimize_gamess_parameters

[40] C. P. Kelly, C. J. Cramer, and D. G. Truhlar, "Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton," *The Journal of Physical Chemistry B*, vol. 110, no. 32, pp. 16 066–16 081, aug 2006, doi: 10.1021/jp063552y.