Prediction of the pK_a of Carboxylic Acids Using the ab Initio Continuum-Solvation Model PCM-UAHF

Gerrit Schüürmann,*,† Maurizio Cossi,‡ Vincenzo Barone,‡ and Jacopo Tomasi§

Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research, Permoserstrasse 15, D-04318 Leipzig, Germany, Department of Chemistry, University Federico II, via Mezzocannone 4, I-80134 Napoli, Italy, and Department of Chemistry and Chemical Industry, University of Pisa, via Risorgimento 35, I-56126 Pisa, Italy

Received: April 21, 1998; In Final Form: June 1, 1998

Experimental pK_a data for 16 aliphatic carboxylic acids are compared with calculated proton-transfer energies in the gas phase and in aqueous solution. The calculations are performed at the SCF and MP2 levels with inclusion of SCF-level entropic and thermochemical corrections to yield free energies of dissociation, using the basis sets 6-31G**, 6-31+G**, 6-311G(2d,2p), and 6-311+G(2d,2p) and the recently parametrized continuum−solvation method PCM-UAHF for the solvation contribution. Relative pK_a trends are reproduced well with correlation coefficients (adjusted for degrees of freedom) of up to 0.97 and standard errors down to 0.24 log units, while the computational accuracy is not sufficient for predicting absolute proton-transfer energies. The latter is mainly caused by deficiencies of the underlying gas-phase calculations, as is demonstrated by a separate analysis of the gas-phase and solution-phase contributions to pK_a.

Introduction

Proton-transfer equilibria are important for an understanding of the physicochemical and chemical behavior of ionogenic compounds in aqueous solution. A prominent example from medicinal chemistry is the ability of drugs to pass biological membranes as well as their potential to interact with intracellular receptors, both of which are affected by the readiness of the drug to undergo protonation or deprotonation.[1] Similarly, the degree of ionization is an important factor for the toxicity and fate of weak organic acids in natural waters,[2] and specific modes of toxic action, such as the uncoupling of the oxidative phosphorylation, depend directly on both the lipophilicity and the acidity of the chemical compound.[3−5]

The classical approach for calculating dissociation constants pK_a from molecular structure is given by the Hammett equation, which is based on a separation of the compound of interest into a suitable parent structure and substituents with associated increment values.[6] Despite its well-known importance for elucidating electronic substituent effects, the Hammett method suffers from its principal restriction to certain types of (previously defined) parent structures and substituents.

An interesting alternative is given by knowledge-based systems that make use of increment values for more generally defined structural features,[7,8] which have lead to a remarkable success, e.g., a predictive squared correlation coefficient (r^2) of 0.89 with a standard error (SE) of ca. 0.8 pK_a units for a total of 2464 organic acids.[7] For much smaller and congeneric series of compounds, solution-phase pK_a has also been correlated with gas-phase quantum chemical parameters,[9,10] and for 135 organic oxyacids a method for predicting dissociation constants was developed on the basis of empirically calculated atomic charges.[11]

Previous investigations of the more fundamental approach of applying quantum chemical continuum−solvation methods[12,13] for predicting solution-phase pK_a cover both semiempirical[14−16] and ab initio treatments.[17,18] The latter, however, were used for only a few compounds of small molecular size. As an overall result, acceptable correlations with relative pK_a trends were found within congeneric series of compounds, although the computational level was not sufficient to predict absolute pK_a values. In particular, comparative analysis of the performance of the semiempirical schemes SM2-AM1,[19] COSMO-AM1,[20] and MST-AM1[21] revealed systematic errors for certain functional groups,[16] suggesting room for improvement by using corresponding ab initio methods. Moreover, a previous study of the pK_a of carboxylic acids applying AM1, COSMO-AM1, and SM2-AM1 to the solutes both without and with microsolvation by three water molecules around the acidic and anionic sites, respectively, has shown that bulk polarization effects as modeled by continuum−solvation models are far more important for yielding correct trends of the dissocation energies than solute−solvent coupling effects.[15]

In the present study, the recently derived UAHF (united atom Hartree−Fock) parametrization[22] of the polarizable continuum model (PCM)[23−25] at the ab initio SCF and MP2 levels was applied to calculate dissociation energies of 16 aliphatic carboxylic acids in aqueous solution. While PCM-UAHF had already been shown to yield competitive results with regard to free energies of solvation for both neutral and charged solutes covering a variety of chemical classes, there was no experience with its performance for predicting the solution-phase energy of proton transfer from organic acids to water.

The basis sets used were of double-ζ-like and triple-ζ-like quality as implemented in the Gaussian 94 package,[26] with one or two sets of polarization functions on all atoms, and without or with diffuse functions on the heavy atoms: 6-31G**, 6-31+G**, 6-311G(2d,2p), and 6-311+G(2d,2p). Calculated energies of proton transfer to H_2O are compared with experi-

† UZF Centre for Environmental Research.
‡ University Federico II.
§ University of Pisa.

S1089-5639(98)01922-7 CCC: $15.00
Published on Web 07/29/1998

mental values derived from pK_a data taken from literature, and statistical regression analyses include bilinear correlations of pK_a with gas-phase and solution-phase portions of the free energy of proton transfer. The results are finally compared with dissociation energies calculated by a previous PCM parametrization based on standard van der Waals radii and separate scaling factors for neutral and ionic solutes, respectively.[27,28]

Theory

The dissociation of an acid AH in aqueous solution,

$$AH_{aq} + H_2O_{aq} \rightleftarrows A_{aq}^- + H_3O_{aq}^+ \quad (1)$$

is governed by the equilibrium constant K, which is related to the dissociation constant K_a according to

$$K = \frac{[A_{aq}^-][H_3O_{aq}^+]}{[AH_{aq}][H_2O_{aq}]} = \frac{K_a}{[H_2O_{aq}]} \quad (2)$$

(where it is assumed that concentrations can be used instead of activities). The associated free energy change ΔG_{aq} in kilojoules per mole can be written for 1 M solutions at 298 K as

$$\Delta G_{aq} = -2.3RT \log K = 5.71\ pK_a + 9.96 \quad (3)$$

and correspondingly, the solution-phase dissociation constant pK_a is given by

$$pK_a = 0.175\Delta G_{aq} - 1.74 \quad (4)$$

The free energy of dissociation in aqueous solution can be decomposed into the respective reaction energy in the gas-phase, ΔG, and a term summarizing the differences between the solvation energies of the ionized and neutral species, $\Delta\Delta G_s$:

$$\Delta G_{aq} = G(A_{aq}^-) - G(AH_{aq}) + G(H_3O_{aq}^+) - G(H_2O_{aq}) = \Delta G + \Delta\Delta G_s \quad (5)$$

with

$$\Delta G = G(A^-) - G(AH) + G(H_3O^+) - G(H_2O) \quad (6)$$

$$\Delta\Delta G_s = \Delta G_s(A^-) - \Delta G_s(AH) + \Delta G_s(H_3O^+) - \Delta G_s(H_2O) \quad (7)$$

The four terms on the right-hand side of eq 7 represent free energies of solvation and have been computed with the recently developed ab initio continuum−solvation model PCM-UAHF.[22]

For the free energy of proton transfer in the gas phase, ΔG, computational models with high accuracy like the G2 model[29,30] and complete basis set (CBS)[31] approaches would be typical methods of choice. For practical reasons, however, the following simplified approach was selected. The simplest calculated quantity is ΔE^{SCF}, the proton-transfer energy in the gas phase at the HF level. A step further is the inclusion of a part of the correlation effects at the MP2 level:

$$\Delta E^{MP2} = \Delta E^{SCF} + \Delta\Delta E^{MP2} \quad (8)$$

Both the SCF and MP2 energies of proton transfer can be supplemented by thermodynamical (entropic and thermochemical) contributions in order to obtain corresponding free energies. For simplicity, these corrections $\Delta\Delta G$ have always been calculated at the SCF geometries:

$$\Delta G^{SCF} = \Delta E^{SCF} + \Delta\Delta G^{SCF} \quad (9)$$

$$\Delta G^{MP2} \approx \Delta E^{MP2} + \Delta\Delta G^{SCF} \quad (10)$$

The solvation contribution $\Delta\Delta G_s$ (eq 7) can be computed both at the SCF and MP2 levels and then added to the gas-phase proton-transfer energies to yield solution-phase dissociation energies of the Brønsted acids AH,

$$\Delta E_{aq}^{SCF} = \Delta E^{SCF} + \Delta\Delta G_s^{SCF} \quad (11)$$

$$\Delta E_{aq}^{MP2} = \Delta E^{MP2} + \Delta\Delta G_s^{MP2} \quad (12)$$

and the corresponding free energies:

$$\Delta G_{aq}^{SCF} = \Delta G^{SCF} + \Delta\Delta G_s^{SCF} \quad (13)$$

$$\Delta G_{aq}^{MP2} = \Delta G^{MP2} + \Delta\Delta G_s^{MP2} \quad (14)$$

The relevance of gas-phase MP2 and thermodynamical corrections can thus be studied by the stepwise addition of these terms to quantify the gas-phase portion of the proton-transfer energy. The present approach thus enables an explicit test of the suitability of the solvation model PCM-UAHF together with routine SCF, MP2, and entropic−thermochemical calculations to predict solution-phase pK_a values.

Materials and Methods

The set of 16 carboxylic acids is listed in Table 1 together with experimental pK_a data taken from literature and associated free energies of dissociation in aqueous solution, ΔG_{aq}, derived from pK_a by application of eq 3.

Initial three-dimensional structures of the compounds were built using the SYBYL software,[32] followed by a preliminary semiempirical geometry optimization at the AM1 level using the MOPAC package.[33] Subsequent geometry optimization at the ab initio SCF level was performed using Gaussian 94[26] with the following four basis sets: 6-31G**, 6-31+G**, 6-311G-(2d,2p), and 6-311+G(2d,2p). For the sake of simplicity, these basis sets are referred to by the short-cut notations dsp (indicating double-ζ-like quality using a split-valence basis set with one additional set of polarization functions for all atoms), dsp+ (dsp with one set of diffuse functions on heavy atoms), ts2p (triple-ζ-like quality using a split-valence basis set and two additional sets of polarization functions on all atoms), and ts2p+ (ts2p augmented by one set of diffuse functions on heavy atoms). For bromine, the missing 6-31G basis set was replaced by a 6-41G contraction scheme as implemented in the Gaussian 94 package, and for iodine 3-21G* and 3-21+G* were used instead of the missing dsp, dsp+, ts2p, and ts2p+ basis sets. With all basis sets, both MP2 and entropic−thermochemical corrections were calculated using the SCF-optimized structures to account approximately for (gas-phase) electron correlation and entropy effects due to vibrational, rotational, and translational degrees of freedom.

Free energies of solvation of the gas-phase molecular structures were calculated with the UAHF parametrization[22] of PCM,[23,24] which was provided as a module for the Gaussian94 software.[25] Normalization of the polarization charge on the cavity surface to account for escaped tails of the wave functions[34−36] was done such that the charge difference was distributed on each tesserae according to local electronic density (option ICOMP=4 in the PCM version implemented in Gaussian 94). For all PCM-UAHF calculations, the number of initial

TABLE 1: Compound Set with Experimental and Calculated Data[a]

no.	compd	exptl pK_a	exptl ΔG_{aq}	ΔE^{SCF}	ΔE^{MP2}	ΔG^{SCF}	$\Delta E_{aq}{}^{SCF}$	$\Delta E_{aq}{}^{MP2}$
1	formic acid	3.77	31.5	756.1	736.8	750.2	54.9	46.0
2	acetic acid	4.76	37.1	775.3	755.1	765.4	66.3	58.7
3	trimethylacetic acid	5.05	38.8	768.7	744.0	760.9	75.5	64.9
4	propionic acid	4.88	37.8	775.3	753.7	766.5	71.1	62.4
5	fluoroacetic acid	2.66	25.1	728.4	706.2	724.2	44.0	32.7
6	chloroacetic acid	2.81	26.0	717.4	702.1	712.0	44.8	40.4
7	bromoacetic acid	2.87	26.3	715.3	697.4	709.0	46.7	40.1
8	iodoacetic acid	3.13	27.8	719.1	700.7	711.7	55.2	48.3
9	α-chloropropionic acid	2.80	25.9	725.0	707.8	717.2	51.8	45.6
10	*β*-chloropropionic acid	4.10	33.4	729.8	713.0	723.3	63.0	55.8
11	dichloroacetic acid	1.30	17.4	682.2	670.7	675.1	35.0	30.7
12	trichloroacetic acid	0.89	15.0	659.8	649.2	651.7	21.1	18.4
13	cyanoacetic acid	2.44	23.9	693.0	678.5	685.9	40.0	34.5
14	nitroacetic acid	1.32	17.5	670.6	658.3	667.0	39.3	29.9
15	α-hydroxypropionic acid	3.87	32.1	722.5	695.3	721.2	58.7	43.1
16	acrylic acid	4.26	34.3	759.6	738.9	753.3	61.8	53.3
	av	3.18	28.1	724.9	755.1	718.4	51.8	44.1
	range	4.16	23.8	115.5	105.9	114.8	54.4	46.5

[a] Experimental pK_a data are taken from literature,[37] and associated free energies of dissociation in aqueous solution, ΔG_{aq}, are derived from eq 3. All energy quantities are given in kilojoules per mole, and calculated dissociation energies in the gas phase (ΔE^{SCF}, ΔE^{MP2}, ΔG^{SCF}) and solution-phase ($\Delta E_{aq}{}^{SCF}$, $\Delta E_{aq}{}^{MP2}$) refer to the proton-transfer reaction outlined in eq 1 and have been generated with 6-31+G** and PCM-UAHF//6-31+G** at the SCF and MP2 levels, respectively, with ΔG^{SCF} representing the respective gas-phase Gibbs free energy (see section Theory).

tesserae on each sphere was set to 196. Details of the PCM-UAHF approach and in particular of the definition of the atomic radii are given in the previous paper.[22]

For comparative purposes, standard PCM calculations of the solvation energies with 60 initial tesserae per atomic sphere were also performed using the following (slightly modified) Pauling set of atomic radii: H (nonpolar), 1.20 Å; H (polar), 0.9 Å; C, 1.5 Å; N, 1.5 Å; O, 1.4 Å; Cl, 1.8 Å; Br, 2.0 Å; I, 2.15 Å. Here, scaling factors of 1.25 and 1.15 were applied for the neutral and charged solutes, respectively, following recommendations for SCF calculations of solvations energies using the 6-31G* basis set.[27,28] The results of these calculations are referred to as PCM-vdW.

Proton-transfer energies were calculated at the SCF and MP2 levels in the gas phase (ΔE^{SCF}, ΔE^{MP2}) and solution phase ($\Delta E_{aq}{}^{SCF}$ and $\Delta E_{aq}{}^{MP2}$; cf. eqs 11 and 12), and addition of the gas-phase entropic−thermochemical correction $\Delta\Delta G^{SCF}$ led to the corresponding (approximate) free energies of dissociation (ΔG^{SCF}, ΔG^{MP2}, $\Delta G_{aq}{}^{SCF}$, $\Delta G_{aq}{}^{MP2}$). Moreover, the solvation part of the solution-phase dissociation was quantified at the SCF and MP2 levels using $\Delta\Delta Gs^{SCF}$ and $\Delta\Delta G_s{}^{MP2}$ as defined in eq 7.

Results and Discussion

As can be seen from Table 1, the 16 carboxylic acids cover a pK_a range of 4.16 units, with minimum and maximum values of 0.89 (trichloroacetic acid) and 5.05 (trimethylacetic acid), respectively. This corresponds to experimental ΔG_{aq} values from 15.0 to 38.8 kJ/mol, thus covering a ΔG_{aq} range of 23.8 kJ/mol (cf. eq 3). Furthermore, Table 1 contains calculated values for ΔE^{SCF}, ΔE^{MP2}, ΔG^{SCF}, $\Delta E_{aq}{}^{SCF}$, and $\Delta E_{aq}{}^{MP2}$ using the 6-31+G** (dsp+) basis set. The corresponding results with 6-31G** (=dsp), 6-31lG(2d,2p) (=ts2p), and 6-311+G(2d,2p) (=ts2p+) are not shown here in order to save space but can be obtained from the authors upon request.

With dsp+, calculated proton-transfer energies in the gas phase (ΔE^{SCF}) are around 725 kJ/mol and span a range of ca. 116 kJ/mol. The corresponding average values with the other three basis sets are 746 (dsp), 749 (ts2p), and 735 kJ/mol (ts2p+) with associated ΔE^{SCF} ranges of 133, 127, and 113 kJ/mol, respectively (data not shown in Table 1). These results indicate a considerable variation of ΔE^{SCF} with the basis set, with regard to both the absolute values and the difference in acidity between the individual compounds. Moreover, addition of diffuse functions results in a much greater lowering of the calculated proton-transfer energies (around 15 and 20 kJ/mol) than going from the double-ζ-like to the triple-ζ-like level (around 3 and 10 kJ/mol). Note further that the ΔE^{SCF} variation with the basis set is 24 kJ/mol and thus as large as the total variation of experimental ΔG_{aq} within the set of 16 carboxylic acids.

At the MP2 level, substantially lower energy differences are achieved, with average values for ΔE^{MP2}(dsp), ΔE^{MP2}(dsp+), ΔE^{MP2}(ts2p), and ΔE^{MP2}(ts2p+) of 739, 707, 749, and 718 kJ/mol, and associated ranges of variation of 129, 106, 126, and 108 kJ/mol. The MP2 correction thus corresponds to an average lowering of the SCF proton-transfer energy of 7 (dsp), 18 (dsp+), 12 (ts2p), and 17 kJ/mol (ts2p+). The basis set differences are more pronounced than at the SCF level: addition of diffuse functions lowers the calculated proton-transfer energy in both cases by ca. 30 kJ/mol on the average, while changing from dsp to ts2p yields a lowering in reaction energy of ca. 10 kJ/mol.

The entropic−thermochemical correction to convert ΔE^{SCF} values to ΔG^{SCF} values at 298 K and 1 atm (assuming ideal gas law) leads to average lowerings of the proton-transfer energy of 5 (dsp), 7 (dsp+), 5 (ts2p), and 6 kJ/mol (ts2p+), resulting in Gibbs free energy differences at the SCF level of 741 (dsp), 718 (dsp+), 745 (ts2p), and 729 kJ/mol (ts2p+), with ranges of ΔG^{SCF} variation of 131 (dsp), 115 (dsp+), 128 (ts2p), and 114 kJ/mol (ts2p+), respectively. Here, the lowering of reaction energy upon addition of diffuse functions (around 16 and 23 kJ/mol) is similar to the situation with ΔE^{SCF}, which holds also true for the effect of replacing dsp by ts2p (around 4 and 11 kJ/mol).

For each of the four basis sets, the MP2 correction is greater than the entropic−thermochemical correction for both the absolute values and the ranges of reaction energy variation, which is particularly pronounced for the basis sets with diffuse functions. For all gas-phase energy quantities ΔE^{SCF}, ΔE^{MP2}, ΔG^{SCF}, and ΔG^{MP2}, an increase of the basis set flexibility corresponds to a decrease of the calculated relative acidity

differences between the individual compounds, which is again greatest upon addition of diffuse functions and for the MP2 level.

The following observations give a useful orientation about the size of the solvation effect upon proton transfer in aqueous solution. In the gas phase, the calculated proton-transfer energies are in all cases above 700 kJ/mol and thus more than a factor of 25 greater than the corresponding dissociation energies in aqueous solution. This acidity enhancement upon solvation corresponds to more than 120 pK_a units, reflecting the well-known differences between gas-phase and solution-phase chemistry. It follows further that the term summarizing the solvation contribution to the solution-phase dissocation energy, $\Delta\Delta G_s$, should be around -650 to -700 kJ/mol, indicating that, for the present set of 16 carboxylic acids, the gas-phase and solvation contributions to the proton transfer in aqueous solution have similar absolute magnitudes. Moreover, the acidity differences (relative to the protonation $H_2O \rightarrow H_3O^+$) between the individual compounds are reduced by factors 4.5−5.5 upon solvation, reflecting a considerable attenuation of the substituent effects in aqueous solution. This solution-phase lowering of acidity differences by ca. 80−110 kJ/mol is also much greater than the average ΔG_{aq} value of 28 kJ/mol, showing again the well-known importance of solvation for the range of chemical reactivities.

Coming back to our analysis of computational methods for predicting pK_a, inclusion of the solvation effect by application of PCM-UAHF leads to ΔE_{aq}^{SCF}, ΔE_{aq}^{MP2} (see Table 1), ΔG_{aq}^{SCF}, and ΔG_{aq}^{MP2} values as defined in eqs 11−14 (again, the calculated energy values not listed in Table 1 can be obtained from the authors upon request). Comparison with the experimental data reveals that, with the presently investigated levels of theory, a direct prediction of absolute pK_a values is not possible: The average ΔG_{aq}^{SCF} values are 68 (dsp), 45 (dsp+), 93 (ts2p), and 88 kJ/mol (ts2p+) and thus exceed the experimental average by factors of 1.6−3.3, and the calculated solution-phase acidity variations of 65 (dsp), 55 (dsp+), 97 (ts2p), and 83 kJ/mol (ts2p+) are greater by factors of 2.3−4.1 as compared to the experimental range. Somewhat smaller average acidities are calculated with ΔG_{aq}^{MP2}, yielding 68 (dsp), 38 (dsp+), 83 (ts2p), and 75 kJ/mol (ts2p+) and spanning acidity differences of 61 (dsp), 47 (dsp+), 92 (ts2p), and 77 kJ/mol (ts2p+). The dissociation energies ΔE_{aq}^{SCF} and ΔE_{aq}^{MP2} as defined in eqs 11 and 12 are greater by 4.5−6.6 kJ/mol than the corresponding free energy terms (note that $\Delta G_{aq}^{SCF} - \Delta E_{aq}^{SCF} = \Delta G_{aq}^{MP2} - \Delta E_{aq}^{MP2} = \Delta\Delta G^{SCF} = \Delta G^{SCF} - \Delta E^{SCF}$).

The overestimation of absolute dissociation energies in aqueous solution is much greater than would be expected from the performance of PCM-UAHF to quantify free energies of solvation, where for 40 neutral solutes and 28 ionic solutes mean errors of ΔG_s of around 0.8 and 4.2 kJ/mol were observed.[22] It suggests that the failure to predict the correct magnitudes of solution-phase dissociation energies is mainly caused by apparent deficiencies in quantifying the gas-phase portion properly.

This is further illustrated by taking acetic acid as an example: For the molecular species CH_3COOH, H_2O, CH_3COO^-, and H_3O^+, the experimental ΔG_s values of -28.0, -26.4, -322.2, and 435.1 kJ/mol yield $\Delta\Delta G_s = -702.9$ kJ/mol. With PCM-UAHF using the dsp+ basis set, $\Delta\Delta G_s^{SCF} = -709.1$ kJ/mol and $\Delta\Delta G_s^{MP2} = -696.4$ kJ/mol, corresponding to calculation errors of -6.2 and $+6.5$ kJ/mol, respectively. The total difference between experimental and calculated dissociation energies in solution using dsp+, however, is 29.2 kJ/mol for ΔE_{aq}^{SCF}, 21.6 kJ/mol for ΔG_{aq}^{MP2}, 19.2 kJ/mol for ΔG_{aq}^{SCF}, and 11.6 kJ/mol for ΔG_{aq}^{MP2}. It shows that, except for ΔG_{aq}^{MP2}, the calculation error for ΔG_{aq} is much greater than the one for $\Delta\Delta G_s$. Moreover, the systematic error of ΔG_{aq}^{MP2} (dsp+) is 13.7 kJ/mol, which is indeed the lowest average overestimation of ΔG_{aq} and still greater by a factor of 3.3 than the above-mentioned mean error of PCM-UAHF for ionic solutes.

TABLE 2: Statistics of Linear Regression Equations for Predicting pK_a of 16 Carboxylic Acids[a]

basis set	param	r_{adj}^2	SE	$F_{1,14}$
dsp	ΔE^{SCF}	0.86	0.48	95.1
	ΔE^{MP2}	0.81	0.56	65.9
	ΔG^{SCF}	0.87	0.47	99.5
	ΔG^{MP2}	0.82	0.54	70.8
	ΔE_{aq}^{SCF}	0.93	0.34	198.6
	ΔE_{aq}^{MP2}	0.89	0.43	123.5
	ΔG_{aq}^{SCF}	0.91	0.38	160.9
	ΔG_{aq}^{MP2}	0.91	0.39	154.7
dsp+	ΔE^{SCF}	0.90	0.42	129.4
	ΔE^{MP2}	0.86	0.48	96.0
	ΔG^{SCF}	0.90	0.41	136.3
	ΔG^{MP2}	0.88	0.46	106.9
	ΔE_{aq}^{SCF}	0.93	0.34	198.2
	ΔE_{aq}^{MP2}	0.90	0.41	133.3
	ΔG_{aq}^{SCF}	0.90	0.42	131.6
	ΔG_{aq}^{MP2}	0.91	0.38	161.9
ts2p	ΔE^{SCF}	0.86	0.49	90.2
	ΔE^{MP2}	0.79	0.60	56.5
	ΔG^{SCF}	0.85	0.49	88.8
	ΔG^{MP2}	0.79	0.60	56.7
	ΔE_{aq}^{SCF}	0.74	0.66	44.0
	ΔE_{aq}^{MP2}	0.77	0.63	50.0
	ΔG_{aq}^{SCF}	0.71	0.70	37.8
	ΔG_{aq}^{MP2}	0.75	0.65	45.7
ts2p+	ΔE^{SCF}	0.89	0.42	125.8
	ΔE^{MP2}	0.86	0.49	92.5
	ΔG^{SCF}	0.90	0.41	132.4
	ΔG^{MP2}	0.87	0.47	100.2
	ΔE_{aq}^{SCF}	0.67	0.74	31.8
	ΔE_{aq}^{MP2}	0.73	0.67	42.0
	ΔG_{aq}^{SCF}	0.63	0.79	26.4
	ΔG_{aq}^{MP2}	0.70	0.71	36.6

[a] The basis sets are given in the short-cut notations as introduced in Materials and Methods, and the statistical results of linear regression analyses are summarized using the following statistical parameters: r_{adj}^2 = squared correlation coefficient corrected for degrees of freedom, SE = standard error (often also called root-mean-squared error), and $F_{1,14}$ = Fisher test value referring to one regression variable and 14 degrees of freedom. All solution-phase parameters are calculated using PCM-UAHF.[22]

While the present level of computation is not sufficiently accurate to predict absolute compound acidities, the calculated gas-phase and solution-phase energy differences may still be useful in deriving linear regression relationships for estimating pK_a. Corresponding statistics with various gas-phase terms and solvation energies calculated by PCM-UAHF are summarized in Table 2.

With gas-phase energies as regression parameters, the following general observations can be noted: The results with dsp and dsp+ are clearly superior to the ones with ts2p and ts2p+, the SCF level yields better predictions of pK_a than the MP2 level, and ΔG^{SCF} and ΔE^{SCF} show very similar performances. The results with ΔG^{MP2} calculated approximately by adding the SCF-level entropic−thermochemical correction to ΔE^{MP2} (see eq 10) are similar to the ones with ΔE^{MP2} alone (data not given in Table 2).

The best single gas-phase parameter is ΔG^{SCF} (dsp+), yielding an explained variance (adjusted for degrees of freedom) of 90% and a standard error of 0.41 pK_a units. Interestingly, ΔE^{MP2} is clearly inferior to ΔE^{SCF} and ΔG^{SCF} for all four basis sets, and the ts2p basis set is inferior to dsp as well as to the

TABLE 3: Statistics of Bilinear Regression Equations for Predicting pK_a of 16 Carboxylic Acids[a]

gas-phase param	solution-phase param	r_{adj}^2	SE	$F_{2,13}$
ΔE^{SCF} (dsp)	$\Delta\Delta G_s^{SCF}$ (dsp)	0.93	0.35	95.0
ΔE^{MP2} (dsp)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.88	0.44	57.6
ΔG^{SCF} (dsp)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.94	0.31	128.1
ΔE^{SCF} (dsp+)	$\Delta\Delta G_s^{MP2}$ (dsp+)	0.97	0.24	209.9
ΔE^{MP2} (dsp+)	$\Delta\Delta G_s^{MP2}$ (dsp+)	0.92	0.36	92.7
ΔG^{SCF} (dsp+)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.96	0.25	196.4
ΔE^{SCF} (ts2p)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.92	0.37	84.4
ΔE^{MP2} (ts2p)	$\Delta\Delta G_s^{MP2}$ (ts2p)	0.82	0.55	35.5
ΔG^{SCF} (ts2p)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.92	0.36	87.9
ΔE^{SCF} (ts2p+)	$\Delta\Delta G_s^{MP2}$ (dsp+)	0.96	0.27	162.2
ΔE^{MP2} (ts2p+)	$\Delta\Delta G_s^{MP2}$ (ts2p+)	0.90	0.40	71.8
ΔG^{SCF} (ts2p+)	$\Delta\Delta G_s^{MP2}$ (dsp)	0.96	0.27	168.5

[a] For each of the gas-phase parameters of a given basis set, the best PCM-UAHF[22] solution-phase parameter covering all four basis sets (see Materials Methods) was selected by applying stepwise regression with eq 17 (for the explanation of the statistical parameters, see Table 2).

basis sets with diffuse functions. With ΔE^{MP2} (ts2p), the statistics are even worse than with the semiempirical AM1 method that yielded $r_{adj}^2 = 0.83$ and SE = 0.54 for the same compound set.[15]

Among the solution-phase parameters tested, ΔE_{aq}^{SCF} is generally better than ΔG_{aq}^{SCF} as a predictor for pK_a and performs better than all other single gas-phase and solution-phase parameters with the dsp and dsp+ basis sets. Comparison of the respective regression equations,

$$pK_a = (0.071 \pm 0.005)\,\Delta E_{aq}^{SCF}(dsp) - (2.0 \pm 0.4) \quad (15)$$

$$pK_a = (0.087 \pm 0.006)\,\Delta E_{aq}^{SCF}(dsp+) - (1.3 \pm 0.3) \quad (16)$$

with eq 4 shows that the regression coefficients are only around half the theoretical value of 0.175. However, the statistics are superior to the ones using the semiempirical continuum−solvation models COSMO-AM1 and SM2-AM1, which both yielded r_{adj}^2 and SE values of 0.90 and 0.41 for the present set of 16 carboxylic acids.[15]

The lower statistical quality of the calculated solution-phase proton-transfer energies with the basis sets ts2p and ts2p+ can be explained by the fact that PCM-UAHF was parametrized using 6-31G* for neutral solutes and 6-31+G* for anions.[22] Thus it is likely that, with ts2p and ts2p+, the solvation calculations suffer particularly from problems with charges escaping the solute cavity.

A way to avoid such computational deficiencies is to split the dissociation energy into the gas-phase and solution-phase portions according to eqs 5−7, which enables application of different basis sets and levels of theory for the evaluation of ΔG and $\Delta\Delta G_s$. The statistical results of corresponding two-parameter regression analyses according to

$$pK_a = a\,\Delta E + b\,\Delta\Delta G_s + c \quad (17)$$

are listed in Table 3. It should be noted that the regression equations were derived in a stepwise manner, starting with the gas-phase term for a given basis set and selecting among the various solution-phase terms of all basis sets the parameter yielding the best increase in correlation with pK_a. As with Table 2, the results with ΔG^{MP2} (representing the formally highest level of gas-phase theory) have been omitted from the table beause they are very similar to the ones with ΔE^{MP2}.

Inspection of Table 3 reveals some interesting trends: Calculation of the gas-phase portion with dsp+ yields better results than with the greater basis set ts2p+ and much better than with ts2p. The prediction performance of the latter is inferior to all of the other three basis sets, including the substantially smaller double-ζ-like scheme dsp. Calculation of the gas-phase portion at the MP2 level leads to an MP2-level solution-phase term as the best second parameter, but at the same time gas-phase MP2 is clearly inferior to gas-phase SCF for predicting pK_a with all four (gas-phase) basis sets. In contrast, the best solution-phase contribution is at the MP2 level in the majority of cases.

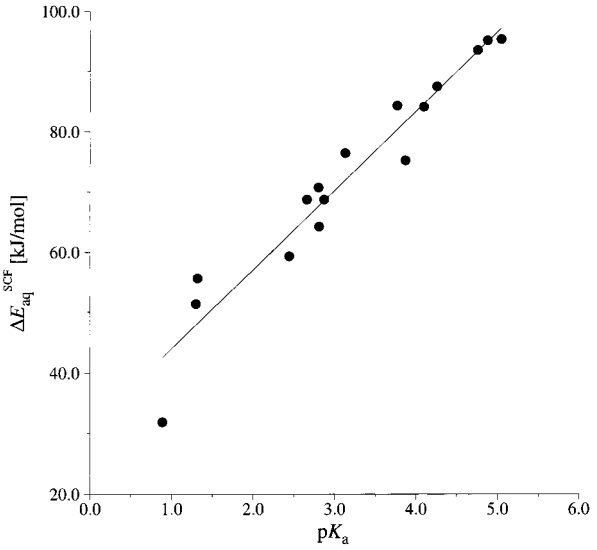

Figure 1. Calculated dissociation energy in aqueous solution (ΔE_{aq}^{SCF}) versus experimental pK_a using PCM-UAHF//dsp (see Materials and Methods) together with the linear regression line according to eq 15.

The two best bilinear relationships read as follows, where it may be useful to note that the $\Delta\Delta G_s^{SCF}$ ($\Delta\Delta G_s^{MP2}$) terms can be easily derived from the corresponding ΔE_{aq}^{SCF} and ΔE^{SCF} (ΔE_{aq}^{MP2} and ΔE^{MP2}) values according to eqs 11 and 12, respectively, the latter of which are listed in Table 1 for the dsp+ basis set:

$$pK_a = (0.066 \pm 0.006)\,\Delta E^{SCF}(dsp+) + (0.056 \pm 0.010)\,\Delta\Delta G_s^{MP2}(dsp+) - (7.9 \pm 2.8) \quad (18)$$

$$pK_a = (0.061 \pm 0.006)\,\Delta G^{SCF}(dsp+) + (0.045 \pm 0.009)\,\Delta\Delta G_s^{MP2}(dsp) - (10.6 \pm 2.6) \quad (19)$$

For regression equations 15 and 18, the data distributions of calculated versus experimental data are shown in Figures 1 and 2. Interestingly, the compounds with greatest calculation errors differ for the two regression equations: With eq 15, the greatest overestimation and underestimation of pK_a are found for nitroacetic acid (+0.65 pK_a units) and trichloroacetic acid (−0.61 pK_a units), while the greatest errors with eq 17 are observed with α-chloropropionic acid (+0.42 pK_a units) and cyanoacetic acid (−0.33 pK_a units).

It is instructive to compare the PCM-UAHF performance for predicting pK_a with the one of PCM-vdW//dsp, using standard van der Waals radii with different values for polar and nonpolar hydrogen and scaling factors of 1.25 and 1.15 for neutral and ionic solutes as recommended for double-ζ-like basis sets.[27,28] As can be seen from Table 4, the explained pK_a variance (adjusted for degrees of freedom) is below 70% with both ΔE_{aq}^{SCF} (dsp) and ΔE_{aq}^{MP2} (dsp) in the PCM-vdW parametrization. Inspection of the data distributions (not shown), however, reveals that these statistics are greatly influenced by the presence of two outliers: The pK_a of nitroacetic acid is overestimated

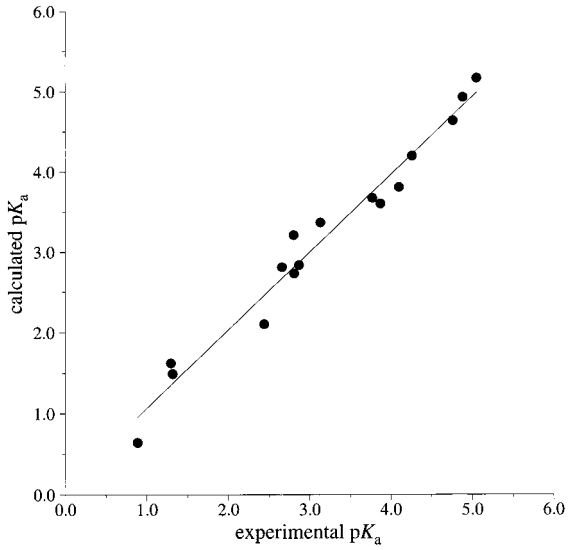

Figure 2. Calculated versus experimental pK_a using ΔE^{SCF} and $\Delta\Delta G_s^{MP2}$ calculated with PCM-UAHF//dsp+ (see Materials and Methods) and the associated regression line according to eq 18.

TABLE 4: Statistics of Linear Regression Equations Based on PCM-vdW for Predicting pK_a of 16 and 14 Carboxylic Acids[a]

solution-phase param	no. of compds	r_{adj}^2	SE	$F_{1,14}$ or $F_{1,12}$
ΔE_{aq}^{SCF} (dsp)	16	0.69	0.72	34.5
	14	0.93	0.33	177.1
ΔE_{aq}^{MP2} (dsp)	16	0.64	0.78	27.8
	14	0.91	0.38	137.4

[a] The subset of 14 compounds is derived from the list of 16 carboxylic acids as given in Table 1 by omission of nitroacetic acid and α-hydroxyacetic acid (for the explanation of statistical parameters see Table 2).

by 2.13 (SCF) and 1.89 (MP2) units, and the one of α-hydroxypropionic acid is underestimated by 1.11 (SCF) and 1.71 (MP2) units. The third largest calculation error is observed for dichloroacetic acid with an overestimation of pK_a by 0.64 (SCF) and 0.80 (MP2). Omission of the two outliers leads to significantly improved regression results with r_{adj}^2 values of 0.93 (SCF) and 0.91 (MP2), respectively, where now α-hydroxypropionic acid yields the largest deviation by 0.64 (SCF) and 0.68 (MP2) pK_a units. These results suggest particular deficiencies of the PCM-vdW parametrization with the functional groups NO_2 and aliphatic OH, but a more definite evaluation would certainly need a greater compound set. It should be further noted that, for this smaller data set of 14 acids, the corresponding statistics with PCM-UAHF//dsp are $r_{adj}^2 = 0.95$ and SE = 0.28 at both the SCF and MP2 levels.

Recently, a new implementation of PCM analytical gradients became available that allows solution-phase geometry optimization within the PCM-UAHF framework and includes also nonelectrostatic contributions to the solvation energy.[38] Corresponding calculations were undertaken for a subset of five compounds at the PCM-UAHF/dsp level using the normalization option ICOMP=2 (scaling of polarization charges with a constant factor),[36] and the resultant proton-transfer energies are compared in Table 5 with the ones of single-point PCM-UAHF calculations using the more sophisticated normalization option ICOMP=4 (distribution of additional surface charges according to the solute electronic density),[36] which cannot be used for calculating analytical derivatives.[38]

As can be seen from the table, there is a considerable variation between the different series of ΔE_{aq}^{SCF} values. For this subset

TABLE 5: Calculated Proton-Transfer Energies Using Gas-Phase and Solution-Phase Geometries[a]

	ΔE_{aq}^{SCF} (kJ/mol)		
	gas-phase geom	solution-phase geom	
compd no.	ICOMP=4	ICOMP=4[b]	ICOMP=2
1	84.5	90.3	81.8
3	95.4	85.8	98.8
5	68.8	57.1	69.1
11	51.4	39.4	67.5
16	87.6	75.9	88.9

[a] The proton-transfer energies refer to eq 1 with molecular geometries optimized at the SCF-dsp (6-31G**) level. The solution-phase geometry optimization was performed using the polarization charge normalization with a constant factor (option ICOMP=2), while the more sophisticated normalization procedure with additional surface charges distributed according to the solute electronic density (option ICOMP=4) was applied for single-point PCM-UAHF calculations. [b] Single-point PCM-UAHF calculation with ICOMP=4 using molecular geometries optimized at the PCM-UAHF/dsp level with ICOMP=2.

of five carboxylic acids, the proton-transfer energies derived from PCM-UAHF//dsp using gas-phase geometries are on the average ca. 4 kJ/mol below the ones from PCM-UAHF/dsp (ICOMP=2). Interestingly, application of single-point calculations with ICOMP=4 on the geometries optimized in solution (with ICOMP=2) yields proton-transfer energies that are lower than the ones from PCM-UAHF//dsp by ca. 8 kJ/mol on the average. These results show that also with inclusion of solution-phase geometry optimization there is still a considerable systematic overestimation of the experimental ΔG_{aq} values for the proton transfer from carboxylic acids to water. However, more experience with the inclusion of solution-phase geometry optimizations for pK_a predictions will be needed before this approach can be better evaluated as compared to the simplified approach using gas-phase geometries.

Conclusions

The potential of ab initio continuum−solvation methods to predict solution-phase proton-transfer equilibria is mainly governed by the level of theory of the underlying gas-phase calculation. With basis sets of double-ζ-like and triple-ζ-like quality at the SCF and MP2 levels, including SCF-level entropic and thermochemical corrections to account for vibrational, rotational, and translational degrees of freedom in the gas phase, the precision of calculated free energies of dissociation is not sufficient for the prediction of absolute pK_a values. However, the results with 16 aliphatic carboxylic acids suggest that, within chemical classes, experimental trends of pK_a can be well-reproduced when using PCM-UAHF for the solvation contribution to compound acidity, which is particularly superior to previous PCM parametrizations based on scaled van der Waals radii fixed for all atom types.

Particular attention should be given to the surprising observation that the gas-phase portion of the proton-transfer energy is apparently better described with 6-31G** and 6-31+G** than with the considerably greater basis sets 6-311G(2d,2p) and 6-311+G(2d,2p), respectively, and better at the SCF and SCF-free energy level than with MP2. Further investigation with other compound sets will be needed to address these aspects in a more definite way.

Decomposition of the dissociation energy in solution into the gas-phase and solvation portions enables combination of routine continuum−solvation calculations with more elaborate gas-phase calculations and thus appears to be a promising tool for deriving predictive regression equations for various chemical classes.

From the viewpoint of practical applications, however, an important question will be whether or not the regression coefficients turn out to be sufficiently similar to combine different compound classes into one (empirically derived) equation. To this end, corresponding investigations with further compound sets are on the way and will be reported in due course. With regard to the level of solvation calculations, there is still room for improvement by extending the geometry optimization to the solution phase, which will be more important for ionic species than for neutral solutes.

References and Notes

(1) Hansch, C., Sammes, P. G., Taylor, J. B., Eds. *Comprehensive Medicinal Chemistry. Quantitative Drug Design;* Pergamon Press: Oxford, U.K., 1990; Vol. 4.

(2) Schüürmann, G. In *Ecotoxicology;* Schüürmann, G., Markert, B., Eds.; John Wiley and Spektrum Akademischer Verlag: New York, 1998; pp 665−749.

(3) Terada, H. *Environ. Health Perspect.* **1990**, *87*, 213.

(4) Schüürmann, G.; Somashekar, R. K.; Kristen, U. *Environ. Toxicol. Chem.* **1996**, *15*, 1702.

(5) Schüürmann, G.; Segner, H.; Jung, K. *Aquat. Toxicol.* **1997**, *38*, 277.

(6) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. pK_a *Prediction for Organic Acids and Bases;* Chapman and Hall: Cambridge, U.K., 1981.

(7) Klopman, G.; Fercu, D. *J. Comput. Chem.* **1994**, *15*, 1041.

(8) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348.

(9) Sotomatsu, T.; Murata, Y.; Fujita, T. *J. Comput. Chem.* **1989**, *10*, 94.

(10) Ozment, J. L.; Schmiedekamp, A. M. *Int. J. Quantum. Chem.* **1992**, *43*, 783.

(11) Dixon, S. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci* **1993**, *14*, 1460.

(12) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(13) Cramer, C. J.; Truhlar, D. G. In *Reviews in Computational Chemistry 6;* Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995.

(14) Ford, G. P.; Wang, B. *J. Mol. Struct. (THEOCHEM)* **1993**, *283*, 49.

(15) Schüürmann, G. *Quant. Struct.-Act. Relat.* **1996**, *15*, 121.

(16) Schüürmann, G. In *Quantitative Structure-Activity Relationships in Environmental Sciences VII;* Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 225−242.

(17) Lim, C.; Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 5610.

(18) Andzelm, J.; Kölmel, C.; Klamt, A. *J. Chem. Phys.* **1995**, *103*, 9312.

(19) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213.

(20) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(21) Luque, F. J.; Bachs, M.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 847.

(22) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(23) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(24) Miertuš, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239.

(25) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. *Chem. Phys. Lett.* **1996**, *225*, 327.

(26) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. J. P.; Head-Gordon, M.; Gonzalez, C., Pople, J. A. *Gaussian 94,* Revision D.4; Gaussian, Inc.: Pittsburgh, PA, 1995.

(27) Bachs, M.; Luque, F. J.; Orozco, M. *J. Comput. Chem.* **1994**, *15*, 446.

(28) Orozco, M.; Luque, F. J. *Chem. Phys.* **1994**, *182*, 237.

(29) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.

(30) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 1293.

(31) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A., Jr., *J. Chem. Phys.* **1996**, *104*, 2598.

(32) *SYBYL Molecular Modelling Software 6.3;* Tripos Asociates Inc.: St. Louis, MO, 1993.

(33) *MOPAC 93 Rev. 2;* Fujitsu Limited and Stewart Computational Chemistry: Chiba, Japan, and Colorado Springs, CO, 1994.

(34) Klamt, A.; Jonas, V. *J. Chem. Phys.* **1996**, *105*, 9972.

(35) Chipman, D. *J. Chem. Phys.* **1997**, *106*, 10194.

(36) Cossi, M.; Mennucci, B.; Pitarch, J.; Tomasi, J. *J. Comput. Chem.* **1998**, *19*, 833.

(37) Christen, H. R.; Vögtle, F. *Organische Chemie-Von den Grundlagen zur Forschung* (Organic Chemistry-From Fundamentals to Research) Otto Salle Verlag: Frankfurt, Germany, 1988, Vol. 1, p 419.

(38) Barone, V.; Cossi, M.; Tomasi, J. *J. Comput. Chem.* **1998,** *19*, 404.