JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

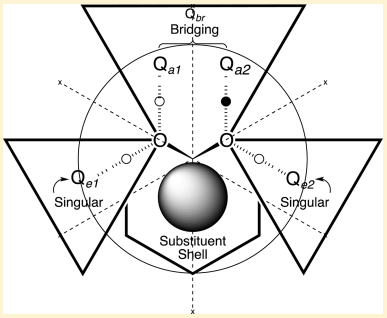
Defined-Sector Explicit Solvent in Continuum Cluster Model for Computational Prediction of pK_a: Consideration of Secondary Functionality and Higher Degree of Solvation

Rebecca A. Abramson[†] and Kim K. Baldridge*[,†]

[†]University of Zürich, OCI, Winterthurerstrasse 190, Zürich CH-8057, Switzerland

Supporting Information

ABSTRACT: Benchmark accuracy for prediction of first and second dissociation constants (pK_{a1} and pK_{a2} values) is realized with the recently developed Defined-Sector Explicit Solvent in Continuum Cluster Model. The model provides a systematic basis for inclusion of explicit solvation, essential for accurate prediction of dissociation constants using computational continuum model approaches. The DSES-CC model is demonstrated by considering the structure-to-chemical affinity relationship of the carboxyl functional group and is shown to provide predictability with mean absolute error of 0.5 pK units across a wide array of carboxylic acid functionality.



INTRODUCTION

Accurate prediction of acid dissociation constants (K_a) has seen significant progress in recent literature.[1−7] A first principles prediction of pK_a within 0.5 pK units of experimental values has been a challenge for the theory of proton transfer reactions and has therefore become a benchmark of broad interest.[8−14] The inherent challenge for QM methods is that at ambient temperature as little as 0.7 kcal/mol error in the ΔG_{diss} leads to misestimation of the pK_a by the benchmark 0.5 pK_a unit, whereas ±1.0 kcal/mol accuracy in energy is still a difficult level to achieve using *ab initio* solvent strategies. To address this challenge, we recently developed the defined-sector explicit solvent in continuum model (DSES-CC) approach, which enables a systematic approach for predictability of solvent networks based on an established preferred conformation of explicit solvent to within ±1.0 kcal/mol.[15] The model was demonstrated through consideration of the structure-to-chemical affinity relationship of the carboxyl functional group.[16]

The defined-sector model provides a systematic basis for inclusion of explicit water molecules in the molecular cavity embedded in implicit solvent, as a continuum-cluster (or explicit-implicit) method. In this method, pK_a is calculated directly from the continuum-cluster method, without using a thermodynamic cycle or means of fitting to experiment. Clusters are systematized based on a strategy for placement of the explicit solvent molecules with respect to the solute. Specific solvation states are defined according to degree of solvation (S_D) and configuration of solvation (S_C). The degree of solvation (S_D) is defined as the number of explicit solvent molecules needed, and the configuration of solvation (S_C) is defined by the specific set of principle solvation sites, secondary solvation sites, and sites within the substituent shell, where solvent is explicitly placed. For the particular case of carboxylic acid and carboxylate functionality, the principal and secondary explicit solvation sites can be illustrated as in Figure 1. Depending on the nature of the substituents on the carboxylic acid, the substituent shell will accommodate explicit solvation, S_D(N + M), where N refers to the degree of solvation of the primary carboxylic moiety and M refers to the degree of solvation of the substituent shell. Evaluation across an array of S_D's reveals patterns of limited direct solvation and provides an indication of how various S_C's affect prediction of pK_a for a set of molecules.

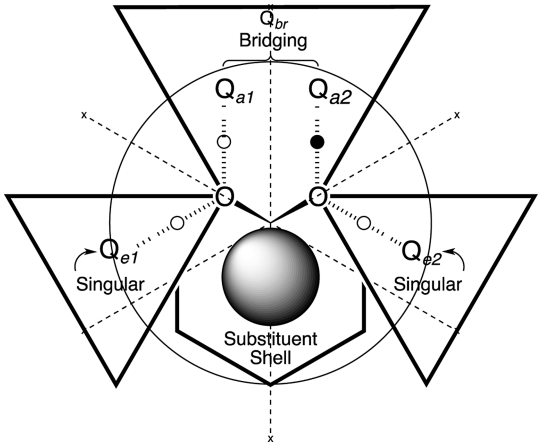


Figure 1. Depiction of principal and secondary explicit solvation sites around a carboxylic acid (or carboxylate). Small circles indicate the presence (filled) or the absence (open) of H.[15]

Received: October 28, 2012
Published: January 2, 2013

In the present work, additional development of the DSES-CC model for prediction of pK_a is illustrated across a much broader set of carboxylic acids (>32, including 9 dicarboxylic acids), thereby further substantiating the model for general use. A much broader range of electronic structure functionality is now addressed, including issues of substituent shell explicit solvation. Important to the fundamentals of the continuum model approach in general, higher degrees of solvation are explored up to $S_D(V)$, which fills all degrees of solvation for the carboxylic acid functionality (Figure 1). Finally, prediction of pK_a for dicarboxylic acids including prediction of the pK_{a2} is undertaken with the DSES-CC model.

COMPUTATIONAL METHODS

All calculations were performed with the GAMESS electronic structure program.[17] Full optimizations were carried out including effects of solvation via the DSES-CC model, using B97-D/6-311+G(2d,p)/COSab, with our most recent implementation of COSab solvation model.[18−20] Parameter optimization for several combinations of DFT functional type and basis sets have been carried out within the solvation model in previous work.[21] In our initial development of the DSES-CC model, investigation covering basis set, wave function type, thermodynamic cycle, reaction scheme, and solvent parameters were carried out.[15] As with any property, one should expect to find variation with DFT-type, basis set representation, and solvent specifications, so it is important is to choose a functional that is appropriate for the property.[12,22,23] In particular, methodology should accommodate the weak interactions present in the explicit/implicit solvent systems. The present work as well as our previous studies well supports the reliability of the B97-D functional together with a triple-ζ basis set. The dispersion enabled density functional B97-D is a reparameterization of the original B97 hybrid functional of Becke[24] and has been implemented and tested in GAMESS within the solvent model.[21] An ultrafine grid, NRAD = 96 NLEB = 1202 was specified. The triple-ζ basis set representation 6-311+G(2d,p)[25] was employed. Analytic Hessian calculations were carried out to characterize the structures and determine zero point energy corrections. Dielectric permittivity of water (ε = 78.4) was used, with cavity parameters of 1082 points for the basic grid, 92 segments on the complete sphere. Outlying charge error correction was taken into account via the double cavity approach.[20] DSES-CC representations were depicted using MacMolPlt.[26]

Consideration of contributions to the nonelectrostatic solvation term, most importantly cavitation and dispersion-repulsion, is important for calculation of accurate pK_a. Under the assumption that the differential cavitation term between carboxylic acids and carboxylates is negligible, inclusion of directed effects through explicit consideration of primary waters of hydration should enable a high level of accuracy if explicit solvation is properly handled. The defined-sector explicit solvent in the continuum cluster (DSES-CC) model relies only on solution phase computations (i.e., eliminating the use of a thermodynamic cycle or fitting schemes) together with the sector model for placement of explicit solvent molecules. This method eliminates a number of possible sources of error, making use of the reaction scheme

$$AH\cdot(H_2O)_n \rightarrow A^-\cdot(H_2O)_n + H^+$$

Both experimental and theoretical values have been used for the free energy of the proton in the literature, due to the associated difficulties for determining this quantity directly.[27] We agree with the previous thorough investigations in the use of the value −265.9 kcal/mol.[9,28−30] The gas phase energy is indisputably derived from an enthalpy contribution, 2.5RT, and an entropic contribution calculated from the Sackur-Tetrode equation, yielding a value of −6.28 kcal/mol.[31] Unique to the DSES-CC model is a greater depth of analysis involving networks of explicit solvent molecules on the individual components of the proton transfer reaction. The solvation state energy is determined for each component of the acid dissociation reaction ($HA\cdot S_C[Q..]$ or $A^-\cdot S_C[Q..]$) in a specific S_C within a given S_D. The ΔG of any specific S_C is determined by subtracting the energy of the reactant state from the product state ($\Delta G = (A^-\cdot S_C[Q..] + H^+) - HA\cdot S_C[Q..]$). The lowest energy set of S_C within a given S_D (labeled S_C*) is used to determine the thermodynamic ΔG_{diss} of acid dissociation for a given S_D. The pK_a follows directly as $\Delta G/2.3RT$,[32,33] and the calculated value is compared to the experimental value as $\Delta pK = pK_a(expt) - pK_a(calcd)$.

RESULTS

Initial Predictive Set. In our first study, a set of carboxylic acids spanning several classes of functionality was investigated.[15] A training set was used to identify a thermodynamically transferable preferred solvent network, which was then applied to three categories of acid structure functionality. Evaluation criteria was based on the fact that 1/2 a pK unit is, in energy terms, only 0.68 kcal/mol, so an acceptable range of predictability was defined to be within 1 kcal/mol of the experimental value or 0.74 pK units. Within the DSES-CC model, it is possible that a range of 'acceptable' HA/A^- pairs for a given $S_D(X)$ may provide pK_a prediction within this target range; however, among any range of potentially acceptable S_C pairs, only S_C within kT of the thermodynamically favored pair need be considered, as others would not be energetically feasible.

Categorization based on electronic and resonance substituents can provide rationalizations for the best S_D and S_C's for each of the different predictive groupings; however, the ultimate goal was to provide a robust transferable cluster that provides consistent results across a large set of compounds within the target range of 0.74 pK units (±1 kcal/mol) of the experimental value. The initial findings showed that $S_D(I)$ clusters generally fail and were only found to be sufficient for electron withdrawing substituents. Although $S_D(II)$ configurations can produce accurate prediction for the small set, there is not a particular S_C that serves across all systems and consequently does not offer the desired transferability. On the other hand, a specific $S_D(III)$ cluster, $S_C*[Q_{a1}Q_{b1}Q_{a2}]$:$S_C*[Q_{a1}Q_{e1}Q_{e2}]$, enables pK_a prediction within 1 kcal mol^{-1}, or 0.74 pK units, across this entire set of carboxylic acids, as a transferable S_C (Table 1, first 8 acids). Figure 2 shows an example of the $S_D(III)$ for one of the training set of acids, acetic acid. These initial studies demonstrate that, through careful consideration of solvation networks, one can assess their predictive power as a function of the number and conformation of explicit solvent molecules, for specific classes of solutes. Such a systematic assessment offers the chance to extend the explicit solvation model to establish general methods applicable to a broader range of solutes.

To establish the proposed methodology as a general method for the positioning of explicit solvation across a broad range of carboxylic acids, the initial set was broadened to include several other substituted carboxylic acids (Predictive Sets A), additional degrees of solvation through the substituent shell (Predictive Sets B), and higher order pK_a prediction (Predictive Sets C) in

Table 1. B97-D/6-311+G(2d,p) DSES-CC-COSab Direct pK_a Prediction Using the 'Preferred' Solvent Network, S_D(III) and S_D(III+M) (M = Substituent Coverage), for Carboxylic Acids Set Compared to Experiment[a]

acid	S_D	exptl pK_a	DSES-CC pK_a	ΔpK_a
Initial Predictive Set				
acetic	S_D(III)	4.7_6	4.5_8	0.1_8
formic	S_D(III)	3.7_7	3.0_9	0.6_8
propanoic	S_D(III)	4.8_6	5.5_8	-0.7_2
isobutyric	S_D(III)	4.8_8	5.1_1	-0.2_3
trimethylacetic	S_D(III)	5.0_3	5.3_7	-0.3_4
chloroacetic	S_D(III)	2.8_1	2.2_5	0.5_6
glycolic	S_D(III)	3.8_4	3.6_5	0.1_9
benzoic	S_D(III)	4.2_0	4.7_0	-0.5_0
Expanded Predictive Set A				
butanoic	S_D(III)	4.8_3	5.2_4	-0.4_1
pentanoic	S_D(III)	4.8_4	5.2_6	-0.4_2
cyclohexanecarboxylic	S_D(III)	4.9_0	5.6_2	-0.7_2
nitroacetic	S_D(III)	1.3_2	1.4_9	-0.1_7
mandelic	S_D(III)	3.4_1	3.1_1	0.3_0
acrylic	S_D(III)	4.2_6	4.5_5	-0.2_9
crotonic	S_D(III)	4.6_9	5.0_8	-0.3_9
trans-cinnamic	S_D(III)	4.4_4	5.4_0	-0.9_6
	S_D'(III)[b]		4.6_8	-0.2_4
Predictive Set B				
o-hydroxybenzoic	S_D(III)	2.9_8	2.3_0	0.6_8
m-hydroxybenzoic	S_D(III)	4.0_8	4.5_2	-0.4_4
	S_D(III+I)		3.9_2	0.1_6
p-hydroxybenzoic	S_D(III)	4.5_8	5.0_4	-0.4_6
	S_D(III+I)		4.4_5	0.1_3
p-methoxybenzoic	S_D(III)	4.5_0	5.3_7	-0.8_7
	S_D(III+I)		5.2_4	-0.7_4
p-butylbenzoic	S_D(III)	4.4_7	5.0_2	-0.5_5
Predictive Set B				
p-aminobenzoic	S_D(III)	4.9_2	6.3_1	-1.3_9
	S_D(III+I)	4.9_2	5.8_6	-0.9_4
p-nitrobenzoic	S_D(III)	3.4_0	3.1_9	0.2_1
Predictive Set C − pK_{a1}				
carbonic[c]	S_D(III)	3.5_8	2.2_3	1.3_5
	S_D(III+I)		3.0_5	0.5_3
oxalic	S_D(III+I)	1.2_3	1.4_4	-0.2_1
malonic	S_D(III+I)	2.8_3	3.3_9	-0.5_6
succinic	S_D(III+I)	4.1_6	5.0_4	-0.8_8
	S_D(III+II)		4.9_4	-0.7_8
	S_D(III+III)		4.9_5	-0.7_9
adipic	S_D(III+I)	4.4_3	5.2_3	-0.8_0
	S_D(III+III)		5.1_8	-0.7_5
fumaric	S_D(III+1)	3.0_3	4.1_6	-1.1_3
	S_D(III+III)		3.7_8	-0.7_5
maleic	S_D(III+1)	1.8_3	2.5_7	-0.7_4
terephthalic	S_D(III+1)	3.5_1	4.0_7	-0.5_6
cyclohexanedicarboxylic	S_D(III+1)	4.1_8	4.96	-0.7_8
Predictive Set C − pK_{a2}				
carbonic[c]	S_D(V)′	10.6_0	10.9_5	-0.3_5
oxalic	S_D(III+III)	4.1_9	4.6_2	-0.4_3
malonic	S_D(III+III)	5.7_9	5.4_8	-0.0_1
adipic	S_D(III+III)	5.4_1	5.8_5	-0.4_4
succinic	S_D(III+III)	5.6_1	5.5_1	0.1_0
fumaric	S_D(III+III)	4.4_4	4.6_6	-0.2_2
maleic	S_D(III+III)	6.0_7	6.0_4	0.0_3
terephthalic	S_D(III+III)	4.4_0	5.1_9	-0.7_9
cyclohexanedicarboxylic	S_D(III+III)	5.4_2	6.1_7	-0.7_5

[a]For experimental values, see refs 30−33. [b]See text for discussion of explicit solven for cinnamic acid. [c]See text for discussion of explicit solvent for carbonic acid.

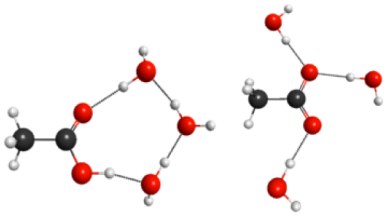


Figure 2. Depiction of S_D(III) for acetic acid HA and A^- pair. Experimental and calculated values of pK_a are 4.7_6 and 4.5_8, respectively.

addition to discussion of higher degrees of solvation, S_D(IV) and S_D(V). Each of these is discussed in detail in what follows.

Expanded Predictive Sets A. Given the results provided by the chosen training set (acetic and formic acid), our original hypothesis was that one should be able to make accurate predictions of pK_a for any carboxylic acid using the identified 'preferred' explicit solvent network, S_D(III) with S_C*[$Q_{a1}Q_{br}Q_{a2}$]:S_C*[$Q_{a1}Q_{e1}Q_{e2}$]. It was possible to show this to be the case for three predictive sets of carboxylic acids, including (I) a class with increasing steric bulk (electron donating groups), (II) a class with electronic withdrawing groups, and (III) an aromatic carboxylic acid functionality. In the present study, a greatly expanded set of acids has been included to probe further the predictability of the DSES-CC model, using the same level of theory as our initial study.[15] In particular, the initial trio of predictive sets has been now expanded to include extended substituent bulk (electron donating) in class I, additional electron withdrawing substituent groups in class II, a more extensive look into aromatic acids beyond the original single system, and a new predictive set of unsaturated functionality, set (IV).

In general, the expectation is that electron withdrawing/donating groups will influence the acidity of a carboxylic acid primarily through stabilization/destabilization of the conjugate base, i.e., inductive effects, resulting in an increase/decrease in the acidity of the acid. Additionally, in unsaturated analogues, delocalization of charge through resonance will be a further charge stabilizing effect, altering the acidity. It is the balance of inductive and resonance effects as partitioned in the mind of the chemist that must be properly modeled computationally, including the important explicit solvation interactions, for accurate prediction of pK_a in these systems.[34]

The extended series of predictive set I in the initial study illustrates the effect of additional bulk on the carboxylic moiety. The full series includes acetic, formic, propanoic, isobutyric, trimethylacetic, butanoic, pentanoic, and cyclohexane carboxylic acids. One can observe the effect of longer saturated chains on the carboxylic acid functionality within the series propanoic, butanoic, and pentanoic acids. In particular, one could imagine that the saturated tail might require additional explicit waters of solvation; however, it appears that no additional substituent shell interactions are necessary to provide pK_a within the tolerance set out. Similarly, other bulky additions to the carboxylic acid do not appear to require attention with respect to additional explicit waters around the substituent group. Data on

the preferred S_D(III) network for the full predictive set I show that this network is indeed well suited to provide predicted pK_a within the target of prediction, with deviations of $+0.1_8$, $+0.6_8$, -0.7_2, -0.2_3, -0.3_4, -0.4_1, -0.4_2, and -0.7_2 pK_a units from experiment, for the above series members, respectively (Table 1). For reference, in all cases $S_D(0)$ results show an overestimation of pK_a by ~2 pK units.

The extended series of predictive set II in the initial study, illustrating the effect of electron withdrawing groups, includes chloroacetic, glycolic, nitroacetic, and mandelic acids. The electron withdrawing groups were considered with regard to how they modify the dipolar nature of the carboxylic acid scaffold and consequently the availability of the principal and secondary solvation sites, Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, and Q_{br}. Nitroacetic acid (experimental pK_a of 1.3_2) offers an even stronger electron-withdrawing group than chloroacetic acid (see, e.g., ref 7), therefore testing the robust nature of the preferred solvent configuration $S_C*[Q_{a1}Q_{br}Q_{a2}]$:$S_C*[Q_{a1}Q_{e1}Q_{e2}]$ for carboxylic acids with very low pK_a values. In this case, pK_a prediction was only 0.1_7 pK units from the experimental value, well within the target deviation. Mandelic acid was considered as an analogue of glycolic acid. The preferred solvent configuration performs well, with the predicted pK_a value only 0.3_0 units below the experimental value of 3.4_1. The results from this predictive set of electron withdrawing substituents are important as they demonstrate that, even with very strong electron withdrawing groups that offer significant stabilization of the carboxylate charge, the solvation sites identified by the preferred configuration, Q_{a1}, Q_{a2}, Q_{br}, (acid) and Q_{a1}, Q_{e1}, Q_{e2} (anion) suffice for accurate predictions.

Predictive set IV introduces the important class of unsaturated carboxylic acids, in particular the 'ene' functionality. The inductive effect of the 'ene' functionality serves to stabilize the carboxylate relative to the acid; however, the resonance contribution can play a role in stabilizing the carboxylic acid state. This set includes acrylic, crotonic, and cinnamic acids. (Note that the general treatment of aromatic acids is treated as a separate predictive set.) In the first two acids of the series, acrylic and crotonic, application of the preferred S_D(III) configurations predicts a pK_a only slightly below that of acetic acid and is well within the target tolerance, with deviations from experimental values of 0.2_9 and -0.3_9 units, respectively (cf. Table 1).

A particularly difficult unsaturated carboxylic acid is trans-cinnamic acid (3-phenylacrylic acid), where there is an additional phenyl substitution on the 'ene' functionality. In this case, the preferred S_D(III) configuration results in a predicted pK_a just outside the target range (Δ = 0.9_6 from experiment). The associated resonance structures of cinnamic acid suggest the need to provide explicit water interactions at the Q_{e1} and Q_{a1} positions in both HA and A^-, in addition to the single explicit water at Q_{a2} and Q_{e2}, for HA and A, respectively. In fact, an S_D(III) configuration of $S_c[Q_{a1}Q_{a2}Q_{e1}]$ around the acidic species (<1 kcal/mol from thermodynamic minimum) together with the standard configuration $S_c[Q_{a1}Q_{e1}Q_{e2}]$ around the anionic species, results in a calculated pK_a (4.6_8) within 0.2_4 units of experiment. In this case, the acid is an electron deficient group that can be stabilized by resonance contributions from the alkene acting as a donor, but the carboxylate is electron rich and cannot benefit from the donor forms of the alkene. The alkene functions then as an electron withdrawing group on the carboxylate through inductive effects only, because there are no beneficial resonance forms shifting electron density from the carboxylate to the alkene. In the acid, the alkene serves as a better donor because its resonance forms are further stabilized by contributions from the phenyl ring.

The collective expanded predictive set A is shown in Table 1 (additional details in the Supporting Information), with 16 acids using the transferable S_D(III) configuration $S_c*[Q_{a1}Q_{br}Q_{a2}]$:S_c*-$[Q_{a1}Q_{e1}Q_{e2}]$. The set shows predicted pK_a well within the target range of 0.74 pK units using only the principal and secondary sites of the carboxylic acid functionality. The single exception discussed is cinnamic acid, where the preferred S_D(III) predicts a pK_a just outside the tolerance (calcd. error 0.9_6 kcal/mol), but for which an acceptable result is found with an alternative S_D(III) where the specific configuration is based on resonance considerations. The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.44 (std. dev. 0.23).

Expanded Predictive Set B, Aromatic Acids. Predictive set B greatly expands on the class of aromatic acids, which consisted of 1 aromatic acid (benzoic acid) in predictive set III of the initial study. Analogous to the 'ene' functionality in Predictive Set A, the carboxylic acid, as an electron deficient group, can be stabilized by resonance contributions from the aromatic ring acting as a donor. The electron rich carboxylate cannot benefit from the donor forms of the aromatic ring, which instead affects the carboxylate through induction. The primary resonance forms of the acid provide insight into how the functionalized aromatic substituent affects the acidity. This will in turn provide insight into first solvation shell explicit interactions, including the need for explicit solvent representation in the substituent shell of the aromatic component.

In the initial trio of predictive sets, the aromatic functionality on the carboxylic acid was briefly investigated with the simplest aromatic acid, benzoic acid. In this case, the ring has only a small influence on the acidity of the carboxyl unit such that benzoic acid is only a slightly stronger acid than acetic acid (4.2_0 vs 4.7_6, expt; 4.7_0 vs 4.5_8, calc.). The relatively weak effect of the phenyl substituent is a result of the additional resonance effect in the acid that is not present in the anion. The result is a weaker acid than what might be expected. Computation predicts benzoic acid to be less acidic than experimentally observed and acetic acid to be slightly more acidic than observed. However, this is a result of the computational model giving greater weight to resonance effects in the carboxylic acid compared to the inductive effects in the carboxylate. As the two systems are experimentally very close in pK value, the balance of the two effects plays an important role in predicting rank order, even if the model provides good results for each independently.

Substituents on the aromatic ring further alter the acidity of carboxylic acids through inductive and/or resonance effects, depending on the nature, type, and placement on the ring. In general, one expects an increase in acidity (lower pK_a) with addition of electron withdrawing substituents on the aromatic ring and a decrease in acidity (higher pK_a) with electron donating groups on the aromatic ring.[34] Consideration of hydroxyl-, methoxy-, amino-, butyl-, and nitrobenzoic acid derivatives enables further testing of the DSES-CC model, in terms of the transferable S_D(III) network, and illustrates the need for further explicit interactions in the substituent shell.

Investigations of the three (o-,m-,p-) isomers of hydroxybenzoic acids provide an interesting test of the DSES-CC model, as the balance of effects varies with position of substituent on the ring, resulting in significant variation in acidity of the three (exptl values 2.9_8, 4.0_8, and 4.5_8, respectively). The para-derivative is the least acidic of the three isomers relative to benzoic acid, considering only an inductive effect. In addition, the para

isomer also has an important resonance effect deriving from the hydroxyl-substituent resonating into the ring and through to the carboxylic acid. This effect stabilizes the acid form, but not the anion form, resulting in the lower acidity of the system compared to the other isomers. This resonance effect is not important in the meta-derivative, and, as such, the effect of the m-hydroxybenzene substituent on the carboxyl unit is primarily inductive in nature, resulting in an only slightly more acidic system than benzoic acid. Importantly, in the para and meta isomers, the preferred $S_D(III)$ predicts a pK_a value within the tolerance limit: -0.4_4 and -0.4_6, respectively. One might consider further the need to add a single explicit water molecule interacting with the lone pair of the hydroxyl substituent of the aromatic moiety, $S_D(III+1)$. In this case, the model predicts a pK_a value within given tolerance limits (0.1_3 and -0.1_6 for para- and meta- respectively). Therefore, these results suggest that both $S_D(III)$ with $S_c^*[Q_{a1}Q_{br}Q_{a2}]$:$S_c^*[Q_{a1}Q_{e1}Q_{e2}]$ as well as $S_D(III+1)$ with $S_c[Q_{a1}Q_{br}Q_{a2};Q_S]$:$S_c[Q_{a1}Q_{e1}Q_{e2};Q_S]$ satisfy our criteria and offer good prediction of the pK_a value.

Although one finds a similar resonance delocalization for salicylic acid (o-hydroxybenzoic acid) and an opposing inductive effect, the proximity of the hydroxyl substituent to the carboxyl units allows for a favorable intramolecular hydrogen bond to be present in the latter given the anion negative charge. The combined effect is a much stronger acid, with a predicted pK_a value of 2.3_0 (exptl, 2.9_8). In terms of explicit water interactions, the intramolecular interaction serves to reduce the number of explicit water molecules interactions needed in the first solvation shell. Coincidentally, $S_D(I)$ provides a pK_a result 0.0_3 pK_a units from the experimental value; however, the $S_D(III)$ preferred configuration results in thermodynamically favored configurations, with predicted pK_a value within the tolerance limits 0.6_8 below the experimental result.

Modification of the hydroxy substituent in p-hydroxybenzoic acid to p-methoxybenzoic acid allows a further test of the sensitivity of the DSES-CC. Calculations with the preferred configurations around the carboxylic/carboxylate moieties result in overestimated pK_a values by 0.8_7, which is slightly outside the tolerance limit. In this case, however, the availability of the methoxy lone pair is attenuated by the inductive effect of the methyl group in comparison to the hydroxyl unit. As such, addition of an explicit water molecule in the substituent shell is warranted here and, in fact, improves the calculated pK_a value to within the tolerance limit of 0.7_4 pK units. Alteration in aromatic substituent from alkoxy to alkyl, as in para-butylbenzoic acid, results in a substituent that is inductive, and predictions using the DSES-CC model with the preferred $S_D(III)$ configuration give the pK_a only 0.5_5 units above the experimental value.

Replacing the aromatic substituent with an electron-withdrawing nitro group serves to increase the acidity of the carboxylic acid, as it stabilizes the parent acid. The $S_D(III)$ preferred configuration in this case provides a prediction of pK_a value for para-nitrobenzoic acid within the tolerance limit, 0.2_1 units below the experimental value, at 3.1_9 (exptl. 3.4_0).

A more difficult case is amino-substitution, where the amino substituent is an electron-donating group through resonance and electron withdrawing through induction. In this case, it is necessary to consider substituent shell explicit solvent interactions with the lone pair of the amine group, as pK_a predictions are over 1 pK unit too acidic without consideration of explicit solvent on the amino group. Using the preferred configuration of solvation for acid and anion, plus additional solvent shell representation, the best estimate is just outside the target range at 0.9_4 pK units too basic, however, still within 1.2_0 kcal/mol of experiment, and so considered acceptable given the known difficulties in modeling amino functionality.

The collective expanded predictive set B (cf. Table 1 and Supporting Information) with 7 acids using the transferable $S_D(III)$ configuration $S_c^*[Q_{a1}Q_{br}Q_{a2}]$:$S_c^*[Q_{a1}Q_{e1}Q_{e2}]$, together with 0 or 1 additional explicit solvents on the aromatic substituent depending on the nature of the aromatic substituent, shows predicted pK_a well within the tolerance limits set out. The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.58 (std. dev. 0.24).

Expanded Predictive Set C, Diprotic Acids: Dicarboxylic Acids. Thus far, only carboxylic acids with a single ionizable group have been considered, and the pK_a value is rationalized via the DSES-CC model with respect to the various structural features of the acid. Another important test for the DSES-CC model is the class of polyprotic acids, which have presented a significant challenge for prediction of pK_a values.[4,35,36] The grouping of polyfunctional acids with the general formula, C(O)OH-R-C(O)OH, R = alkyl, alkenyl, alkynyl, aryl), is characterized by having two ionizable carboxylic acid units. There are a number of issues pertaining to the prediction of pK_a values of these acids, including (1) whether the preferred $S_D(III)$ network provides adequate prediction for the first deprotonation reaction, given the change in electronic structure due to the presence of the second carboxylic group, (2) whether the second carboxylic group should be treated as a substituent or supports also the preferred $S_D(III)$ configuration $S_C^*[Q_{a1}Q_{br}Q_{a2}]$: $S_C^*[Q_{a1}Q_{e1}Q_{e2}]$, and (3) whether calculation of the pK_a value of the second deprotonation reaction using the DSES-CC method also provides predictive results. Points (1) and (2) are addressed in this section; point (3) is addressed in the following section.

To address points (1) and (2), it is instructive to consider more specifically what constitutes the DSES-CC model for such a system (cf. Figure 1). If the second carboxylic acid group is considered as part of the substituent shell, then the number of principal and secondary explicit solvent molecules does not change, and one only needs to address any needed substituent shell explicit solvents, in much the same way as already treated in the monocarboxylic acids. If, on the other hand, one considers each of the two carboxylic acid moieties as primary sites of explicit solvation, then the number of principal and secondary solvation sites exactly doubles, Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, and Q_{br}, and consideration of potential Q_s sites in between the carboxylic acid functionalities must also be addressed. In the latter case, the relative positioning of the two carboxylic acid functionalities with respect to one another could allow for shared Q_{e1} and Q_{e2} principal sites (e.g., oxalic acid).

For the series of alkyl dicarboxylic acids, acidity is related to the chain length of the alkyl group between the two carboxylic groups. In the series considered here, carbonic acid is included as it has been used as an exemplary case in a number of studies in the prediction of pK_{a2} value.[10,37] Figure 3 gives a depiction of the principal and secondary explicit solvation sites within the DSES-CC model in this special case. For the first deprotonation reaction of carbonic acid to bicarbonate anion a selection of possible S_C's of the primary carboxylic group is shown in Figure 4 (see also the Supporting Information). Importantly, 4h is the preferred $S_D(III+1)$ configuration, which provides the predicted pK_a value within 0.5_3 of the experimental value. This example illustrates the flexibility of the DSES-CC model to treat a special case.

dx.doi.org/10.1021/ct300934q | J. Chem. Theory Comput. 2013, 9, 1027−1035

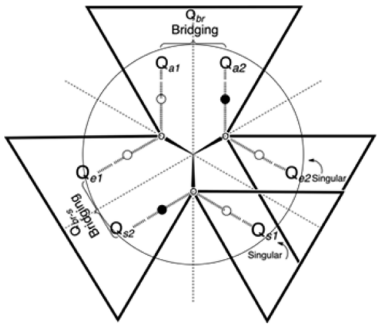


Figure 3. Depiction of the principal and secondary explicit solvation sites around dicarbonic acid (or the deprotonated forms). Small circles indicate the presence (filled) or the absence (open) of H.

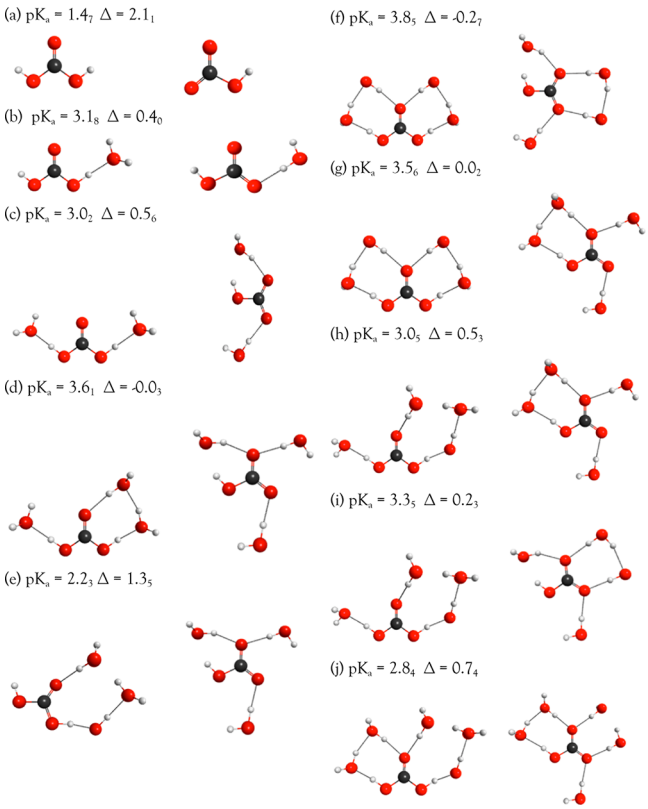


Figure 4. B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a as a function of solvation degree (S_D) and solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, Q_{br}, Q_{s1}, Q_{s2}, $Q_{br\text{-}s}$) for carbonic acid and associated anion: a) $S_D(0)$; b) $S_D(I)$ $S_c[Q_{a2}]$:$S_c[Q_{a2}]$; c) $S_D(I+I)$:$S_D(II)$, $S_c[Q_{a2};Q_{s2}]$:$S_c[Q_{e1}Q_{e2}]$; d) $S_D(II+I)$:$S_D(III)$, S_c*$[Q_{a1}Q_{a2};Q_{s2}]$:$S_c[Q_{a1}Q_{e1}Q_{e2}]$; e) $S_D(III)$, $S_c[Q_{a1}Q_{br}Q_{a2}]$:$S_c[Q_{a1}Q_{e1}Q_{e2}]$; f) $S_D(III+I)$:$S_D(IV)$, S_C*$[Q_{a1}Q_{a2}Q_{e1};Q_{s2}]$:$S_c[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$; g) $S_D(III+I)$, S_c*$[Q_{a1}Q_{a2}Q_{e1};Q_{s2}]$:$S_c$*$[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]$; h) $S_D(III+I)$, $S_c[Q_{a1}Q_{br}Q_{a2};Q_{s2}]$:$S_c$*$[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]$; i) $S_D(III+I)$:$S_D(IV)$, $S_c[Q_{a1}Q_{br}Q_{a2};Q_{s2}]$:$S_c[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$ j) $S_D(IV+I)$, $S_c[Q_{a1}Q_{br}Q_{a2}Q_{e1};Q_{s2}]$:$S_c[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}]$.

Oxalic acid, HOOCCOOH, is the shortest chain with two separate carboxylic acids. In this case, one expects the first pK_a value to be significantly lower than the typical monocarboxylic acid because formation of the monoanion is facilitated (stabilized) by the residual acid via hydrogen bonding. Prediction of the pK_a value for oxalic acid was achieved using the preferred S_D(III) explicit solvent configuration applied to one of the carboxylic units and a single explicit water applied to the second carboxylic group, thereby treating the second carboxyl unit as a substituent. The predicted pK_a value is indeed quite acidic at 1.4_4, and the result is well within the tolerance limit of the experimental value of 1.2_3 (Table 1).

When the number of carbon atoms between the carboxyl units increases, as in the series malonic, $COOHCH_2COOH$, succinic, $COOHCH_2CH_2COOH$, and adipic, $COOHCH_2CH_2CH_2CH_2COOH$, acids, geometric constraints and strong local solvation from water prevent the formation of stabilizing intramolecular H-bonds, resulting in acids that are much less acidic than oxalic acid. Computations using S_D(III+1) predicted pK_{a1} values for malonic, succinic, and adipic acids of 3.3_9 (2.8_3), 5.0_4 (4.1_6), and 5.2_3 (4.4_3), respectively, where values in parentheses are experimental results (cf. Table 1). These results suggest that the effect of the second carboxyl unit is as a substituent. In addition, the influence of applying 1, 2, or 3 explicit solvents is relatively minor but does show a convergence of results from S_D(III+1), to S_D(III+II), to S_D(III+III), such that all values come within the tolerance limits. Table 2 shows this convergence

Table 2. B97D/6-311+G(2d,p) Direct Sector Explicit Solvent in Continuum Model Results for Succinic Acid (Exptl pK_a = 4.1_6)

succinic acid S_D clusters		pK_a	ΔpK_a
S_D(III+I)			
HA	A$^-$		
S_c*$[Q_{a1}Q_{br}Q_{a2};Q_{a2}]$	S_c*$[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$	5.0_4	-0.8_8
$S_c[Q_{a1}Q_{a2}Q_{e1};Q_{a2}]$	Sc*$[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$	4.0_0	$+0.1_6$
S_D(III+II)			
HA	A$^-$		
$S_c[Q_{a1}Q_{br}Q_{a2}:Q_{a1}Q_{a2}]$	$S_c[Q_{a1}Q_{e1}Q_{e;}Q_{a1}Q_{a2}]$	4.9_4	-0.7_8
S_D(III+III)			
HA	A$^-$		
S_C*$[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	S_C*$[Q_{a1}Q_{e1}Q_{e2};Q_{a1}Q_{br}Q_{a2}]$	4.9_5	-0.7_9

in S_D for succinic acid across this set of solvent shell explicit configurations.

The set of dicarboxylic acids was extended further to consider more complex bridges than the simple alkyl linkage between the carboxylic acid units. In this category, fumaric and maleic acids, C(O)OH-CHCH-C(O)OH, have intervening unsaturated units in the trans- and the cis-conformations, respectively, terephthalic acid has an intervening aromatic ring, C(O)OH-Ar-C(O)OH, and cyclohexanedicarboxylic acids has an intervening saturated ring unit, C(O)OH-C_6H_{10}-C(O)OH.

Initial predictions for fumaric acid using the preferred S_D(III+1) resulted in an overestimation of the first pK_a value by 1.1_3 units (exptl. pK_a = 3.0_3). A comprehensive DSES-CC analysis for this acid was therefore conducted (Table 3 and Supporting Information). Comparing to succinic acid as the unsaturated analogue, the unsaturated bond and second carboxylic group offer substantial stabilization of the charge of the anionic species of fumaric acid substantially lowering the pK_a. Comparison across the series S_D(III+I), S_D(III+II), and S_D(III+III) shows a convergence of results, with S_D(III+III) providing a balanced explicit distribution and prediction of a pK_a value on the edge of the tolerance limit with respect to the experimental value. The cis isomer, maleic acid has a significantly lower pK_a due to stabilization of the anion through formation of an intramolecular hydrogen bond between the two carboxylic groups in this conformation. The predicted pK_a value using the preferred S_D(III+1) is 2.5_7, which is within an acceptable tolerance of the experimental value.

dx.doi.org/10.1021/ct300934q | J. Chem. Theory Comput. 2013, 9, 1027−1035

Table 3. B97D/6-311+G(2d,p) Direct Sector Explicit Solvent in Continuum Model Results for Fumaric Acid (Exptl pK_{a1} = 3.0_3)

S_D cluster assignment		pK_a	ΔpK_a
$S_D(0)$		2.7_2	0.3_1
$S_D(I)$			
HA	A^-		
$S_c[Q_{a2}]$	$S_c[Q_{a2}]$	3.6_1	-0.5_8
$S_D(I+I)$			
HA	A^-		
$S_c[Q_{a2};Q_{a2}]$	$S_c[Q_{a2} + Q_{a2}]$	3.9_3	-0.9_0
$S_D(II+II)$			
HA	A^-		
$S_c[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$	$S_c*[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$	3.8_2	-0.7_9
$S_c[Q_{a1}Q_{a2};Q_{a1}Q_{a2}]$	$S_c[Q_{a1}Q_{a2};Q_{a1}Q_{e2}]$	4.2_0	-1.1_7
$S_D(III+I)$			
HA	A^-		
$S_c*[Q_{a1}Q_{br}Q_{a2};Q_{a2}]$	$S_c*[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$	4.1_6	-1.1_3
$S_c[Q_{a1}Q_{a2}Q_{e1};Q_{a2}]$	$S_c*[Q_{a1}Q_{e1}Q_{e2};Q_{a2}]$	2.8_2	$+0.2_1$
$S_D(III+II)$			
HA	A^-		
$S_C[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{a2}$	$S_C[Q_{a2}Q_{e1}Q_{e2};Q_{a1}Q_{a2}$	3.7_1	-0.6_8
$S_D(III+III)$			
HA	A^-		
$S_c*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	$S_c[Q_{a1}Q_{e1}Q_{e2};Q_{a1}Q_{br}Q_{a2}]$	3.7_9	-0.7_6
$S_c*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	$S_c*[Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{br}Q_{a2}]$	3.7_8	-0.7_5

Terephthalic acid, a para-substituted benzoic acid, is analogous to the para-substituted derivatives in predictive set B. As an electron withdrawing substituent, the COOH-Ar- substituent is expected to make the acid somewhat more acidic than benzoic acid. Results using the preferred $S_D(III+I)$ shows a predicted pK_a, value of 4.0_7, which is within the defined tolerance of the experimental value (-0.5_6) and more acidic than benzoic acid (calcd. 4.7_0). Finally, in the case of cyclohexanedicarboxylic acid, one finds that the preferred $S_D(III)$ with one additional substituent shell explicit solvation (i.e., $S_D(III+I)$, provides a pK_a value within the tolerance limit, at 4.9_6 (expt: 4.1_8).

The collective expanded predictive set C of pK_{a1} values with 9 diacids using the transferable $S_D(III)$ configuration $S_c*[Q_{a1}Q_{br}Q_{a2}]$:$S_c*[Q_{a1}Q_{e1}Q_{e2}]$ with 3 substituent explicit solvents, together with carbonic acid, are reported in Table 1 (additional details in the Supporting Information). The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.71 (std. dev. 0.27).

Dicarboxylic Acids, Second Protonation States. The remaining point to be addressed in this section involves the ability of the DSES-CC model to predict multiple acidic protons. In particular, the second acid dissociation constants, pK_{a2}, are of interest for the class of dicarboxylic acids, as also recently explored in the literature.[4] In general, one expects the second protonation state in water to be much weaker (larger pK_a values), since it is more difficult to remove a proton from an anion than from an uncharged molecule. However, the structure of the intervening R group of the COOH − R − COOH will be important in determining the relative acid strength of the remaining proton. In particular, one expects that, as the distance between the two carboxylic units increases, the acidity of the second proton increases.

For all of the diprotic acids except carbonic acid, prediction of pK_{a2} is achieved within the tolerance limit with the preferred DSES-CC configuration around both carboxylic(ate) groups (Figure 5). In all cases, pK_{a2} is indeed less acidic than pK_{a1}. In particular, oxalic acid has a predicted pK_{a2} that is considerably less acidic than pK_{a1} due to the fact that the second acid proton is held more tightly via an intramolecular hydrogen bond, as facilitated by the proximity of the carboxyl units.

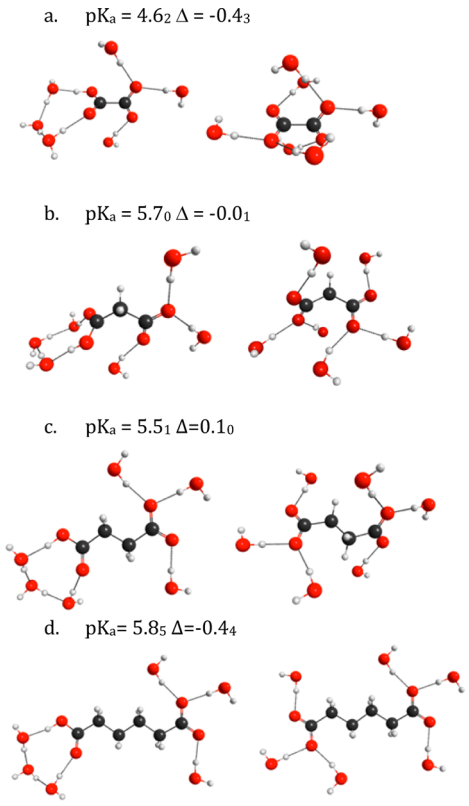


Figure 5. B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a for the second deprotonation reaction with the preferred configuration $S_D(III+III)$, $Sc(Q_{a1}Q_{br}Q_{a2};Q_{a1}Q_{e1}Q_{e2}]$:$Sc[Q_{a1}Q_{e1}Q_{e2};Q_{a1}Q_{e1}Q_{e2}]$, for both carboxylic/carboxylate groups of (a) oxalic, (b) malonic, (c) succinic, and (d) adipic acids.

Prediction of pK_{a2} for carbonic acid is considered a special case just as in prediction of pK_{a1}, which due to its small size technically has only principal solvent sites (Figure 3). As was done for the assignment of S_D for pK_{a1}, it is more instructive to refer to the sum of the explicit molecules, rather than the components. The findings from defined sector model study of the training sets reveals the second deprotonation reaction, from carbonate to bicarbonate, to be quite sensitive to explicit placement. However, a converged result is found with a total of five explicit solvent molecules, as shown in Figure 6d.

The collective predictive set C of pK_{a2} values for the 9 diacids using the transferable $S_D(III)$ configuration $S_c*[Q_{a1}Q_{br}Q_{a2}]$:S_c*-$[Q_{a1}Q_{e1}Q_{e2}]$ together with 3 substituent explicit solvents, and the special case of carbonic acid, are reported in Table 1 (additional details in the Supporting Information). The mean absolute error (MAE) for calculated pK_a across all acids in set A is 0.35 (std. dev. 0.29).

Alternative $S_D(III)$'s. Across all systems, a preferred $S_D(III)$ with $S_C*[Q_{a1}Q_{br}Q_{a2}]$:$S_C*[Q_{a1}Q_{e1}Q_{e2}]$, together with the inclusion of 1−3 explicit solvents in the substituent shell where warranted, appears to satisfy pK_a value prediction within the tolerance limits set out; however, one might expect other possibilities could exist. The key is that any 'preferred' S_C needs to be transferable among a large set of structures and within kT of the thermodynamic minimum, in order to be a faithful representation of

dx.doi.org/10.1021/ct300934q | J. Chem. Theory Comput. 2013, 9, 1027−1035

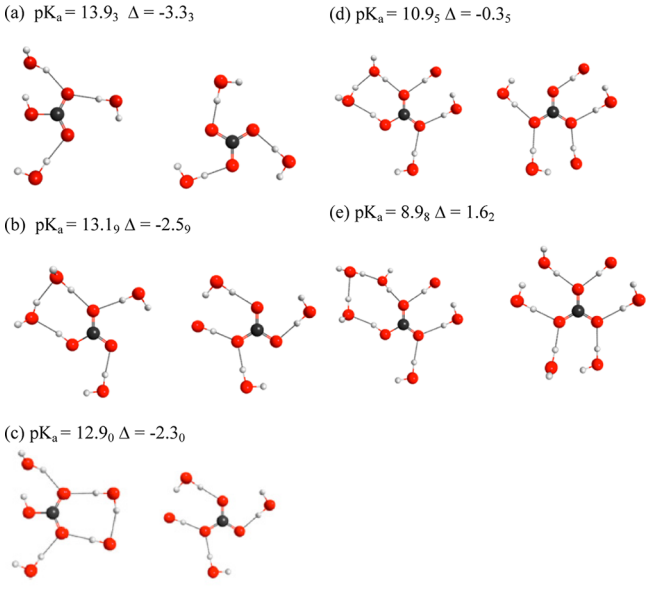


Figure 6. B97-D/6-311+G(2d,p) DSES-CC-COSab pK_a as a function of solvation degree (S_D) and solvation sites (Q_{a1}, Q_{a2}, Q_{e1}, Q_{e2}, Q_{br}, Q_{s1}, Q_{s2}, $Q_{br\text{-}s}$) for carbonate and associated anion, bicarbonate: a) S_D(III):S_D(II+I) $S_C[Q_{a1}Q_{e1}Q_{e2}]$:$S_C[Q_{a1}Q_{e2};Q_{s2}]$; b) S_D(III+I):S_D(II+II) S_C*$[Q_{a1}Q_{e1}Q_{e2};Q_{s2}]$:$S_C[Q_{a2}Q_{e1};Q_{s1}Q_{s2}]$; c) S_D(IV):S_D(II+II) $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2}]$:$S_C[Q_{a1}Q_{a2}Q_{e2};Q_{s1}]$; d) $_D$(IV+I):S_D(III+II) $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}]$:$S_C[Q_{a1}Q_{a2}Q_{e2};Q_{s1}Q_{s2}]$; e) S_D(IV+II) $S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s2}Q_{br\text{-}s}]$:$S_C[Q_{a1}Q_{a2}Q_{e1}Q_{e2};Q_{s1}Q_{s2}]$.

the ensemble. For example, one can find a second S_D(III) with the same anion configuration as the preferred anion S_C but with an alternative acid configuration of $S_C[Q_{a1}Q_{a2}Q_{e1}]$, which also provides excellent prediction of pK_a values. However, while in several cases the new acid configuration is <0.5 kcal/mol of the preferred acid configuration, there are also several cases where the difference is quite large (e.g., nearly 2 kcal/mol). As such, this alternative S_D(III) does not appear to be a transferable S_D(III) (see., e.g., Supporting Information provided). In the set of acids considered in this study, only one 'preferred' S_D was found, that being S_C*$[Q_{a1}Q_{br}Q_{a2}]$:S_C*$[Q_{a1}Q_{e1}Q_{e2}]$, and, across the entire set of acids considered, this provided pK_{a1} predictions with MAE of 0.50 (std. dev. 0.28).

Higher Degrees of Solvation – Substituent Shell. While generally, within a any particular acid, prediction of pK_a value converged toward the experimental value with principal and secondary explicit solvation sites represented by a 'preferred' S_D(III), one might question whether higher degrees of principal solvation show convergence of predicted pK_a, given that 4 principal and 1 secondary explicit solvent sites are present in the carboxyl unit (cf. Figure 1); however, consideration of S_D(IV) was already observed to result in unsatisfactory results for the training set.[15] In this work, a further look into both S_D(IV) and S_D(V) was undertaken for a larger grouping of carboxylic acids (see Supporting Information provided), to explore more fully whether the S_D(IV) results are anomalous or whether the carboxylic and carboxylate systems are always fully satisfied with S_D(III) in the preferred configuration, S_c*$[Q_{a1}Q_{br}Q_{a2}]$:S_c*$[Q_{a1}Q_{e1}Q_{e2}]$.

In all cases considered, prediction of pK_a with a fully saturated solvation shell (i.e., occupation of all 4 principal and 1 secondary solvent sites, Figure 1), S_D(V), as well as S_D(IV) is quite poor and well outside the target tolerance (e.g., Table 4 and Supporting Information). The question then arises as

Table 4. B97D/6-311+G(2d,p) Direct Sector Explicit Solvent in Continuum Model S_D(V) Results for Acetic Acid (Exptl pK_a = 4.7$_6$) and Formic Acid (Exptl pK_a = 3.7$_7$)

S_D(V) cluster		pK_a	ΔpK_a
acetic acid			
HA	A^-		
$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	2.5$_1$	2.2$_5$
formic acid			
HA	A^-		
$S_C[Q_{a1}Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	$S_C[Q_{a1}\ Q_{br}Q_{a2}Q_{e1}Q_{e2}]$	0.7$_4$	3.0$_3$

to why the higher degrees of solvation, S_D(IV) and S_D(V), generally provide poor representations of solution state of carboxylic acids. One might presume that, when the addition of explicit solvent molecules disturbs the "natural" charge distribution of the solute, the predicted pK_a value will be out of the acceptable range of accuracy. It appears that S_D(IV) and higher degrees of solvation tends to overcrowd the solute systems with more directed interaction in the first solvation shell than would be realistic in a dynamic solution environment. Consequently, the additional explicit solvents begin to constitute the bulk, which not only introduces further challenges but also does not provide accurate pK_a prediction. On the other hand, it is conceivable that these results indicate a fundamental inadequacy in the continuum model approach itself, which is a subject of our future investigations.

In the context of the present DSES-CC model, one can assert that accurate predictions of pK_a for a general carboxylic acid can be realized using the identified 'preferred' explicit solvent network, S_D(III) with S_C*$[Q_{a1}Q_{br}Q_{a2}]$:S_C*$[Q_{a1}Q_{e1}Q_{e2}]$. This degree of solvation appears to adequately capture the principal, secondary, and substituent shell directed interactions between solute and solvent, with the continuum model capturing the essentials of the bulk.

CONCLUSIONS

One of the most fundamental reactions in chemistry and biochemistry involves the protolytic reaction of acids and bases, as illustrated by the volume of natural and synthetic organic compounds with acidic or basic functionality. Determining systematic effects of polar substituents on ionization of acids enables establishment of fundamental structure/reactivity relationships. Theoretical prediction of pK_a has been quite challenging and tends to vary widely in chemical accuracy depending on methodology and class of compounds. In particular, for continuum models, a significant challenge has been inclusion of explicit first shell solvation interactions, necessary for accurate prediction of pK_a. The DSES-CC model has been presented as an important step for determining explicit solvation in the first solvation shell. The model has been demonstrated for prediction of both pK_{a1} and pK_{a2} values across a broad range of carboxylic acids, a relatively challenging class of functionality.

In the relative comparison of acid strengths among a series of carboxylic acids, entropy factors are not considered but are found to make only minor contribution. The relative translational and rotational degrees of freedom between acid and anion are similar for all acids being compared, so that enthalpy factors become the most important factor for prediction of the relative acidities.[6] In this way, a straightforward approach using only the continuum model plus the appropriate defined-sector model is found to be needed for prediction of acid dissociation constants. Through careful consideration of solute solvent

surfaces, the model has enabled generalizations that indicate number and conformation of explicit solvent molecule networks for classes of solutes and associated functionality. For the class of carboxylic acid structure (32 acids, including 9 dicarboxylic acids), a 'preferred' network conformation, consisting of S_D(III), with S_C*[$Q_{a1}Q_{br}Q_{a2}$]:S_C*[$Q_{a1}Q_{e1}Q_{e2}$] plus substituent explicit solvation when necessary, is found to provide pK_a within the tolerance set out, with a MAE of 0.50 pK units (0.7 kcal/mol) accuracy. Moreover, the model shows equal reliability for prediction of pK_{a2} values of dicarboxylic acids.

Future studies will investigate a) the general applicability of the DSES-CC model for other classes of functionality, b) ways to automate the method for S_D/S_C choice and solvent placement, and c) the fundamental nature of the transition from higher degrees of explicit solvation to the continuum model. Extension of the DSES-CC model for other functionality formally requires designation of principal and secondary explicit solvation sites around the relevant functional groups (e.g., amine, alcohol, carbon acid, etc.), as in Figure 1 for carboxylic acid functionality. The present study demonstrates how the DSES-CC model addresses other functionality through the treatment of the substituent shell component of the carboxylic acid sector model. In this way, the DSES-CC sectors for a variety of functionality are illustrated. Future studies should detail the different degrees and configurations of solvation and preferred solvent networks for other classes of functionality.

■ ASSOCIATED CONTENT

Supporting Information

Details of computational methodology, depictions of specific solvation states in accord to degree of solvation (S_D) and configuration of solvation (S_C), with associated pK_a data for all systems considered. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kimb@oci.uzh.ch.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the University of Zürich and the Swiss National Science Foundation for support of this research. We are grateful to Prof. Donald Truhlar, Jay S. Siegel, and Andreas Klamt for helpful discussions.

■ REFERENCES

(1) Casasnovas, R.; Fernandez, D.; Ortega-Castro, J.; Frau, J.; Donoso, J.; Munõz, F. *Theor. Chem. Acc.* **2011**, *130*, 1−13.
(2) Du, D.; Qin, M.; Zhou, Z.-Y.; Fu, A. *Int. J. Quantum Chem.* **2011**, *112*, 351−258.
(3) Ho, J.; Coote, M. L. *WIREs* **2011**, *1*, 649.
(4) Marenich, A. V.; Ding, W. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys.Chem. Lett.* **2012**, *3*, 1437−1442.
(5) Sutton, C. C. R.; Franks, G. V.; da Silva, G. *J. Phys. Chem. B* **2012**, *116*, 11999−12006.
(6) Zhang, S. *J. Comput. Chem.* **2011**, *33*, 517−526.
(7) Zheng, Y. C.; Chen, X.; Zhao, D.; Li, H.; Zhang, Y.; Xiao, X. *Fluid Phase Equilib.* **2012**, *313*, 148−155.
(8) Eckert, F.; Diedenhofen, M.; Klamt, A. *Mol. Phys.* **2009**, *108*, 1.
(9) Ho, J.; Coote, M. *Theor. Chem. Acc.* **2010**, *125*, 3−21.
(10) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 2493−2499.
(11) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. *J. Phys. Chem. A* **2003**, *107*, 9380−9386.
(12) Liu, J.; Kelly, C. P.; Goren, A. C.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G.; Zhan, C.-G. *J. Chem. Theor. Comput.* **2010**, *6*, 1109−1117.
(13) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2010**, *6*, 2829−2844.
(14) Pliego, J. R.; Riveros, J. M. *J. Phys. Chem. A* **2002**, *106*, 7434−7439.
(15) Abramson, R. A.; Balridge, K. K. *Mol. Phys.* **2012**, *110*, 2401−2412.
(16) Klebe, G. In *Structure Correlation*; Bürgi, H.-B., Dunitz, J. D., Eds.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 1994; Vol. *2*, pp 543−603.
(17) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Elbert, S. T. *J. Comput. Chem.* **1993**, *14*, 1347.
(18) Baldridge, K. K.; Klamt, A. *J. Chem. Phys.* **1997**, *106*, 6622−6633.
(19) Baldridge, K. K.; Jonas, V. *J. Chem. Phys.* **2000**, *113*, 7511.
(20) Gregerson, L. N.; Baldridge, K. K. *Helv. Chem. Acta* **2003**, *86*, 4112.
(21) Peverati, R.; Baldridge, K. K. *J. Chem. Theory Comput.* **2009**, *5*, 2772−2786.
(22) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157−167.
(23) Marenich, A.; Cramer, C.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 877−887.
(24) Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554−8560.
(25) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
(26) Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Modell.* **1999**, *16*, 133−138.
(27) Camaioni, D. M.; Schwerdtfeger, C. A. *J. Phys. Chem. A* **2005**, *109*, 10795−7.
(28) Kelly, C.; Cramer, C.; Truhlar, D. *J. Phys. Chem. B* **2006**, *110*, 16066−16081.
(29) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 2493−9.
(30) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R., Jr. *J. Phys. Chem. A* **1998**, *102*, 7787−7794.
(31) Schmidt am Busch, M.; Knapp, E.-W. *ChemPhysChem* **2004**, *5*, 1513−1522.
(32) Hasselbalch, K. A. *Biochem. Z.* **1917**, *78*, 112−144.
(33) Henderson, L. J. *Am. J. Physiol.* **1906**, *21*, 173−179.
(34) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*; VCH: Weinheim, 1988.
(35) Lee, T. B.; McKee, M. L. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10258−10269.
(36) Smiechowski, M. *J. Mol. Struct. (Theochem)* **2009**, *924*−*926*, 170−174.
(37) Wang, X.-X.; Fu, H.; Du, D.-M.; Zhou, Z.-Y.; Zhang, A.-G.; Su, C.-F.; Ma, K.-S. *Chem. Phys. Lett.* **2008**, *460*, 339−342.