Check for updates

Received: 2 December 2021 | Revised: 25 March 2022 | Accepted: 29 March 2022

DOI: 10.1002/jcc.26864

RESEARCH ARTICLE

Journal of COMPUTATIONAL CHEMISTRY WILEY

Prediction of acid pK_a values in the solvent acetone based on COSMO-RS

Niklas Sülzner[1] | Julia Haberhauer[1] | Christof Hättig[1] | Arnim Hellweg[2]

[1]Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, Bochum, 44780, Germany

[2]Dassault Systèmes Deutschland GmbH, Am Kabellager 11-15, Köln, 51063, Germany

Correspondence
Niklas Sülzner, Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany.
Email: niklas.suelzner@rub.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: Germany's Excellence Strategy - EXC 2033 - 3906778, Project Ha 2588/8-1

Abstract

In this contribution we extent the use of the conductor-like screening model for realistic solvation (COSMO-RS) to the prediction of pK_a values in acetone, a commonly used dipolar aprotic solvent. For this, we calculated the Gibbs free energy of dissociation of 120 organic acids (nine acrylic acids, 87 benzoic acids, nine phenols, and 15 benzenesulfonamides) using COSMO-RS at the two levels BP-TZVP and BP-TZVPD-FINE and determined the parameters for a linear free energy relation for the pK_a prediction by performing linear fits to experimental values. Our results suggest that the data set dissects into two groups, with the phenols being different from the other three subsets. The acrylic and benzoic acids and the sulfonamides can be treated together and yield an excellent linear correlation ($r^2 > 0.95$) with an RMSD of only ~0.3. The slope is found to be significantly smaller than the theoretical value ($1/RT\ln 10$), only 45% of it, which is in accordance with findings in the literature. The phenols, however, while similarly well correlated in their own subset with an RMSD of 1.7–1.9, exhibit a slope larger than one. We discuss both a higher uncertainty in the reference values as well as physical origins as possible reasons.

KEYWORDS
pK_a, acetone, acidity, COSMO-RS, proton transfer

1 | INTRODUCTION

Proton transfer is ubiquitous in chemistry and related fields.[1] Among countless examples, its involvement in catalytic[2] and enzymatic[3] reactions or the energy storage using fuel cells[4] are perhaps most noticeable at present. Reflecting this overall high importance, much work has been invested for a long time into the investigation of proton-transfer reactions in general,[5] and the determination of pK_a values or their associated acid dissociation constants in particular.[6] Owing to its central role as the thermodynamic measure of Brønsted acidity, the pK_a has gained widespread attention.

While pK_a determination is nowadays standard in aqueous solution due to a broad variety of well-established experimental techniques for the solvent water,[7] the situation is very different in non-aqueous solution.[8,9] Especially for aprotic organic solvents, the acidity of a compound is usually severely reduced since an aprotic solvent cannot stabilize the corresponding base effectively due to its inherent lack of H-bond donation, which largely hinders the dissociation event. The resulting extremely low acid dissociation constants in combination with other complications such as low solubilities challenge the accuracy of experimental methods.[10] The non-trivial and relatively costly experimental determination is a major reason why

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Journal of Computational Chemistry* published by Wiley Periodicals LLC.

theoretical pK_a predictions become much more important for solvents other than water.[11,12] Besides and independent from the solvent, computational methods also become crucial for such molecules, for example, reactive intermediates that elude the standard determination because of short lifetimes, difficult isolation, and so forth.

Several computational approaches for pK_a predictions have been established in the literature (e.g., see Reference [12]), including quantum chemical approaches that avoid the costly direct sampling over a statistically representative number of solvent configurations, as done in molecular dynamics simulations with explicit solvent. These studies typically utilize the polarizable continuum model (PCM) to include implicit, that is, bulk solvation effects.[13] PCM belongs to the family of dielectric continuum solvation models (DCSMs, for an overview see, for example, Reference [14]) and is closely related to the conductor-like screening model (COSMO).[15–17] When applying DCSMs to pK_a prediction, a linear free-energy relation (LFER) between the pK_a value and the Gibbs free energy change of dissociation (ΔG_{diss}) is usually employed

$$pK_a = c_1 \frac{\Delta G_{diss}}{RT\ln(10)} + c_0 \quad (1)$$

(with the ideal gas constant R, absolute temperature T, slope c_1 and intercept c_0), illustrating the need for accurate ΔG_{diss} values. One should note that this equation is an empirical generalization of the exact thermodynamic definition of pK_a, where ideally $c_0 = 1$ and $c_1 = -\lg[S]$ (with the solvent concentration $[S]$).[18] Although various, in principle equivalent, computational approaches for determining ΔG_{diss} exist, which only differ in the underlying thermodynamic cycle, they all have in common that the Gibbs free energy of solvation (ΔG_{solv}) of all involved species must be accounted for.[19]

Despite their inherent disadvantages, for example, neglect of specific solute-solvent interactions or issues with short-range electrostatics, the use of DCSMs can be actually beneficial with respect to thermodynamic properties since the implicit solvation already accounts for a statistically averaged description.[17,20,21] Furthermore, these models and particularly COSMO have become widely used over the years owing to their computational efficiency and robustness. With COSMO-RS, that is, COSMO for realistic solvation, some substantial progress has been made in overcoming the inherent deficiencies of a continuum solvation model that is almost second to none. In order to improve the accuracy, this extension of COSMO reintroduces the missing contributions from local, directional interactions (e.g., H-bonds) by using an additional statistical-thermodynamics treatment of interacting surfaces (details given elsewhere).[20–23] Importantly, it was already successfully applied to the prediction of pK_a values in water,[24,25] DMSO,[24] acetonitrile,[26] and heptane.[27]

Another common dipolar, aprotic organic solvent is acetone. Chemical reactions as well as analytics are examples of high practical importance where pK_a values in acetone play a role. Only very few experimental results, for example, the compilation of pK_a values in various organic solvents by Izutsu[28,29] or the later work by Ludwig et al.[30] and Pytela et al.,[31–34] but no theoretical studies were available as a reference though.

In the current study, we extend the parameterization for pK_a calculations to acetone as an important aprotic solvent. Thereby, we followed closely the computational procedure described in References [24–26]. This is in particular of interest in light of recent studies of super photoacids in this solvent,[35,36] where previous implementations of COSMO excited-state energy gradients for efficient wave function-based methods[37,38] offer the possibility to employ these LFER parameters for COSMO in future applications on pK_a^*.

2 | METHODOLOGY

2.1 | Theoretical calculations

Our calculations of ΔG_{diss} in acetone assume the following model for the dissociation reaction

$$HA + S \rightleftharpoons A^- + HS^+, \quad (2)$$

where HA/A^- and HS^+/S refer to the conjugate Brønsted acid–base pair of the target acids HA and the solvent acetone (with $S = (CH_3)_2O$), respectively. We adapted the workflow for pK_a LFER parameterizations described in References [24–26] taking into account the recent guidelines proposed by Hellweg and Eckert.[39] Thus, the RI approximation[40] was routinely employed in all DFT[41] and COSMO-DFT[15] calculations and the conductor limit ($\epsilon = \infty$) was exclusively applied in all COSMO calculations. All quantum chemical (QC) calculations were done with the TURBOMOLE program package (version 7.3).[42] The default settings, in particular with respect to the COSMO cavity construction, were used unless stated differently.

To obtain best-quality results, we chose the recommended BP-TZVPD-FINE level (vide infra) of the COSMOtherm program[27,39] as the default for our calculations. Accordingly, the following steps were performed for each of the four species in Equation (2) to calculate the Gibbs free energy in solution:

1. Conformational analysis using the COSMOconf program (version 2020)[43] (with help of its GUI COSMOconfX [version 20.0.0]). COSMOconf automatically searches for all conformers of a given initial structure by generating a set of start structures and refining this set iteratively based on certain energetic and structural thresholds while optimizing the geometries.
2. Quantum chemical calculations for each conformer identified in step 1 at the following four different levels of theory using TURBOMOLE to generate the input required for the BP-TZVPD-FINE level of COSMOtherm (in step 3):
   a. Solution-phase geometry optimization using DFT with the BP86 functional[44,45] in combination with COSMO and the def-TZVP basis set.[46,47]
   b. A subsequent, more accurate single-point-energy calculation using the BP86 functional in combination with COSMO, in the

larger def2-TZVPD basis set,[48] and with a refined (i.e., smooth radii-based isosurface) cavity construction[49] at the COSMO-BP/TZVP geometry.

c. Gas-phase geometry optimization using the BP86 functional and the def-TZVP basis set.

d. A subsequent and more accurate single point energy calculation using the BP86 functional with the larger def2-TZVPD basis set at the BP/TZVP geometry.

3. A COSMO-RS calculation for the free energy in solution at 25°C (298.15 K) as implemented in the COSMOtherm program (version 2020)[27] for the complete conformer set. At the BP-TZVPD-FINE level, COSMO-RS uses the solution-phase refined-cavity information (step 2b) and the gas-phase energy (step 2d), both calculated with the larger TZVPD basis set. Since the results in the smaller TZVP basis are already obtained during the preceding geometry optimizations, the corresponding value at the BP-TZVP level of COSMOtherm can be additionally computed similarly without large effort (based on steps 2a and 2c).

Both calculations, those for the gas phase and those for the solution phase, are required by COSMO-RS to predict the ΔG_{solv} contributions to the total Gibbs free energies. To conveniently apply this workflow to all investigated acids HA, we used the BP-TZVPD-FINE-COSMO+GAS-18 template provided by COSMOconfX to run the conformer searches (cf. step 1) already at the desired QC levels (cf. step 2), which directly yielded the required input for COSMO-RS (cf. step 3). Supplementary information on the structures and energies of all conformers was made available in Reference [50]. We also exemplarily checked the results after step 2 and performed additional test calculations (e.g., potential energy scans or vibrational frequency calculations) if needed to ensure that only equilibrium structures, that is, minima on the potential energy surface were included. However, vibrational zero-point energies were not explicitly taken into account, following the assumption from the preceding COSMO-RS pK_a predictions that the difference in zero-point energies between the acid and its conjugate base is generally small.[24–26] This assumption had been based on the observation that this difference was found to be fairly constant for molecules with the same functional group.[51,52] Therefore, the contribution from zero-point energies will be implicitly incorporated into the LFER parameters c_0 and c_1 of each compound class.

The effect of different conformations was taken care of by a Boltzmann average with respect to the Gibbs free energy in solution ($G_i^S = E_i^{COSMO} + \mu_i^S$) of the conformers. Since the chemical potential μ_i^S of a given conformer in the system S depends, at finite concentration, on the populations themselves, the equations for the population are solved self-consistently in the COSMOtherm program.[27] Symmetry factors, that is, conformer weights due to symmetry-degeneracies, were explicitly taken into account to avoid a biased description if only one enantiomer was included in the conformer set.

The central quantity ΔG_{diss} can be split according to

$$\begin{aligned}\Delta G_{diss} &= G_{tot}(A^-) + G_{tot}(HS^+) - (G_{tot}(HA) + G_{tot}(S)) \\ &= (G_{tot}(A^-) - G_{tot}(HA)) + (G_{tot}(HS^+) - G_{tot}(S))\end{aligned} \quad (3)$$

into a species-dependent contribution $G_{tot}(A^-) - G_{tot}(HA)$ and a constant solvent-specific contribution $G_{tot}(SH^+) - G_{tot}(S)$. Using the procedure described above, we obtain values of −1098.6 and −1103.9 kJ mol^{-1} for $G_{tot}(SH^+) - G_{tot}(S)$ at the BP-TZVP and BP-TZVPD-FINE levels, respectively. These values are in line with estimates in the literature for the solvation enthalpy of the proton (in acetone) of −1051 kJ mol^{-1} based on the DFT functional B3-LYP in combination with the IEF-PCM solvation model in the aug-cc-pVQZ basis set[53] or −1070 kJ mol^{-1} in the smaller 6-311++G** basis set,[54] considering that (i) mere enthalpies do not include entropic contributions and (ii) our estimates differ by the proton's total Gibbs free energy in the gas phase (−26.3 kJ mol^{-1})[55] from the full Gibbs free energy of solvation. Since ΔG_{diss} is subject to the linear free-energy relation in Equation (1), systematic errors due to the computational method (COSMO-RS, DFT functional, basis set, etc.) will change the LFER parameters c_1 and c_0, while the choice of the solvent species HS^+/S and the corresponding error will only affect the parameter c_0. Hence, regarding the LFER for pK_a prediction, there is no point in improving this constant solvent-specific contribution, as any errors in it will be absorbed in c_0 anyway.

2.2 | Fit data set

To determine the LFER parameters in Equation (1), we used a set of 120 acids in the solvent acetone for which experimental reference data is available from the compilation of pK_a values in organic solvents by Izutsu[28] that combines different sources. In addition, we included new results from Ludwig et al.,[30] Pytela et al.,[31,33,34] and Kulhánek et al.[32] Only those experimental values were chosen that fulfill the condition of representing true, thermodynamic dissociation constants (i.e., in the limit of infinite dilution or ionic strength). The obtained data set consists of 9 acrylic acids,[56] 87 benzoic acids (1 Parent, 18 ortho,[31] 16 meta,[30] 16 para,[30] 10 2,6-disubst.,[32] 16 3,4-disubst.,[33] and 10 3,5-disubst.[34] derivatives), 9 phenols,[57] and 15 benzenesulfonamides[58]; with halogen atoms and amino, nitro, methyl, or methoxy groups as typical substituents.

The reported pK_a values are not directly comparable with each other due to the use of different reference values that have been used for calibration. The values for acrylic acids are given relative to a pK_a value of 13.5[56] for benzoic acid, while the values for benzoic acids and sulfonamides refer to a corresponding value of 18.2.[30,58] In turn, the measurements for phenols used picric acid (2,4,6-trinitrophenol) with a pK_a value of 6.3[57] as a reference, while the measurements on benzoic acids and sulfonamides refer to 9.2[30,58] for that compound. We prefer the values 18.2 and 9.2 over the others since (i) 18.2 for benzoic acid is both the more recent value and has later been reproduced by the same group[33] and (ii) 9.2 for picric acid has been

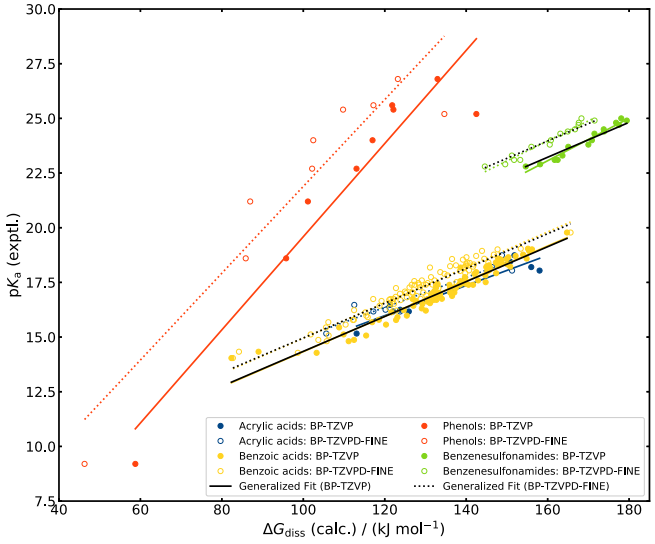


FIGURE 1 (Shifted) Experimental pK_a values [pK_a(exptl.)] plotted against calculated Gibbs free energies of dissociation [ΔG_{diss}(calc.)] in the solvent acetone at the two computational levels BP-TZVP and BP-TZVPD-FINE. Each color represents a specific compound class and linear regression lines for each class are included in the respective color. Full lines refer to the BP-TZVP and dashed lines to the BP-TZVPD-FINE level. The LFER parameters and correlation coefficients for these class-specific fits are given in Table 2. In addition, also the results from a generalized fit that combines the different subsets (details given in the text) are included as black lines and the corresponding fit results can be found in Table 3

determined via two independent methods.[59] Furthermore, both values also appear much more reasonable in the light of recent pK_a values in the solvent acetonitrile.[60,61]

Consequently, we shifted the pK_a values of the acrylic acids up by 4.7 (benzoic acid as reference) and those for the phenols up by 2.9 (picric acid as reference) in order to account for these differences in the corresponding experimental reference values. The resulting shifted pK_a values then range from 9.2 (picric acid) to 26.8 (2,6-dichlorophenol) and therefore span roughly 18 logarithmic units.

3 | RESULTS AND DISCUSSION

3.1 | Class-specific considerations

We begin our discussion with the pure ΔG_{diss} values and class-specific pK_a values before moving on to those pK_a values based on a generalized LFER. A direct comparison between the experimental pK_a values and the calculated ΔG_{diss} values in Figure 1 reveals distinct differences between the molecular classes. To further illustrate this, we also calculated pK_a values based on class-specific LFER parameter fits (cf. linear regression lines in Figure 1), referred to as $pK_a^{(calc.)}$ and given in Table 1. The corresponding LFER parameters are listed in Table 2.

At the BP-TZVP level, all compound classes show very good linear correlations between $pK_a^{(exptl.)}$ and ΔG_{diss} with squared correlation coefficients r^2 larger than 0.9, smallest for the acrylic acids (0.913) and largest for the benzenesulfonamides (0.963). The r^2 value for the benzoic acids is also very close (0.940). Despite their different experimental origins, the pK_a values of the acrylic acids fit very well to those for the benzoic acids, demonstrating that a shift of the pK_a scale is indeed justified in this case. Although the pK_a values of the benzoic acids and sulfonamides have been determined by the same authors,[30–34,58] the latter do not match those for the carboxylic acids, but appear to be vertically shifted by a constant offset instead. Since a difference in the experimental reference values should not be the cause here, this behavior rather appears to have real, physical origin (vide infra). In addition, the phenols appear quite well linearly correlated ($r^2 > 0.91$), but characteristically different from the other classes.

The LFER parameter for the slope, c_1, is smallest for the acrylic acids (0.393), medium-large for the benzoic acids (0.460) and sulfonamides (0.555), and largest for the phenols (1.216). The statistical errors are largest for the acrylic acids and phenols (~10%), and decrease over the sulfonamides (~5%) toward the benzoic acids (~3%), reflecting the increasing numbers of compounds in the classes that increase the statistical significance. This flexibility in the slope is also reflected in the uncertainty of the intercepts, c_0. Their statistical errors amount to about 1 for the acrylic acids and sulfonamides, almost 3 for the phenols, but only 0.3 for the benzoic acids. Within these error bars, the overall agreement between the carboxylic acids and the sulfonamides is actually already fairly good. In analogy to the slope, the phenols also clearly differ from the other classes with respect to the intercept.

Largest deviations between the class-specific fits and the experimental data are observed for the phenols and benzoic acids, while the former will be discussed later. Still, due to the large number of benzoic acids, it is worthwhile to have a look at the most significant outliers. The largest absolute error is found for the ortho-SH derivative (1.130), followed by the ortho-COOH (0.884), and meta-$COCH_3$ (0.739) derivatives. If these compounds are excluded, the correlation further improves ($r^2 = 0.981$) and the regression results in the coefficients $c_0 = 5.106$ and $c_1 = 0.507$ at an RSMD of 0.044. The poor agreement for ortho-SH benzoic acid seems to be caused by a qualitatively wrong deprotonation at the SH group (i.e., proton exchange in the carboxylate), while for the ortho-COOH derivative the rather large proton delocalization involving intra-molecular hydrogen bonding might be the origin, as already indicated for dicarboxylic acids in acetonitrile.[26]

In general, similar observations are also made for the computational level BP-TZVPD-FINE, but the computed ΔG_{diss} values are systematically smaller, that is, horizontally shifted to the left compared to the BP-TZVP level. This shift is largest for the phenols and sulfonamides and much smaller for the carboxylic acids. We suspect that at least two factors contribute to this shift. First, we found a difference of about ~5 kJ mol^{-1} between these two levels for the constant solvent-dependent term contributing to ΔG_{diss}. Second, due to the negative charge, we anticipate a stronger energetic stabilization of the anionic bases relative to the neutral acids when moving to the higher computational level since the main difference lies in the larger basis

TABLE 1 Comparison of experimental pK_a values ($pK_a^{(exptl.)}$) and calculated ΔG_{diss} values of 120 acids in acetone using COSMO-RS at both the BP-TZVP and BP-TZVPD levels

		BP-TZVP					BP-TZVPD-FINE				
			Class-specific		Generalized			Class-specific		Generalized	
Compound	$pK_a^{(exptl.)}$	ΔG_{diss}	$pK_a^{(calc.)}$	ΔpK_a	$pK_a^{(calc.)}$	ΔpK_a	ΔG_{diss}	$pK_a^{(calc.)}$	ΔpK_a	$pK_a^{(calc.)}$	ΔpK_a
Acrylic acids											
Parent	18.45	150.8	18.10	0.35	18.40	0.05	148.1	18.28	0.17	18.78	−0.33
2-F	16.17	125.9	16.39	−0.22	16.41	−0.24	123.8	16.62	−0.45	16.85	−0.68
2-Cl	16.25	123.8	16.24	0.01	16.24	0.01	120.2	16.37	−0.12	16.56	−0.31
2-Br	16.17	122.1	16.12	0.05	16.10	0.07	117.2	16.17	0.00	16.32	−0.15
2-CH_3	18.74	151.8	18.17	0.57	18.48	0.26	148.8	18.33	0.41	18.84	−0.10
3-C_6H_5	18.20	156.0	18.46	−0.26	18.81	−0.61	146.4	18.16	0.04	18.64	−0.44
2-CN,3-C_6H_5	16.47	121.3	16.07	0.40	16.04	0.43	112.5	15.85	0.62	15.95	0.52
3-(2-Furyl)	18.04	158.0	18.60	−0.56	18.97	−0.93	151.3	18.50	−0.46	19.03	−0.99
2,3-Dibromo-3-formyl	15.16	113.1	15.50	−0.34	15.39	−0.23	105.6	15.38	−0.22	15.40	−0.24
Benzoic acids											
Parent	18.20	144.9	17.95	0.25	17.93	0.27	139.6	18.16	0.04	18.11	0.09
2-Substituted (ortho)											
2-F	17.44	139.9	17.54	−0.10	17.52	−0.08	133.0	17.62	−0.18	17.58	−0.14
2-Cl	17.01	133.4	17.02	−0.01	17.01	0.00	125.2	16.99	0.02	16.96	0.05
2-Br	16.84	132.0	16.91	−0.07	16.90	−0.06	122.7	16.78	0.06	16.76	0.08
2-I	16.98	131.4	16.86	0.12	16.85	0.13	123.7	16.87	0.11	16.84	0.14
2-CH_3	18.39	148.7	18.26	0.13	18.23	0.16	140.9	18.26	0.13	18.21	0.18
2-OH	14.81	111.1	15.23	−0.42	15.23	−0.42	105.7	15.41	−0.60	15.41	−0.60
2-OCH_3	18.94	155.4	18.80	0.14	18.77	0.17	143.9	18.50	0.44	18.45	0.49
2-COOH	14.33	89.0	13.45	0.88	13.47	0.86	84.2	13.67	0.66	13.70	0.63
2-$COOCH_3$	17.39	132.9	16.98	0.41	16.97	0.42	126.6	17.10	0.29	17.07	0.32
2-$OCOCH_3$	17.39	135.9	17.22	0.17	17.21	0.18	130.1	17.38	0.01	17.35	0.04
2-NH_2	18.58	148.0	18.20	0.38	18.18	0.40	144.4	18.55	0.03	18.49	0.09
2-$NHCH_3$	18.28	145.6	18.01	0.27	17.98	0.30	137.5	17.98	0.30	17.94	0.34
2-$NHCOCH_3$	15.79	118.6	15.83	−0.04	15.83	−0.04	112.4	15.95	−0.16	15.94	−0.15
2-NO_2	15.91	123.1	16.19	−0.28	16.19	−0.28	115.8	16.23	−0.32	16.21	−0.30
2-SCH_3	17.97	142.5	17.75	0.22	17.73	0.24	133.6	17.67	0.30	17.63	0.34
2-$SCOCH_3$	17.08	132.1	16.92	0.16	16.91	0.17	124.9	16.97	0.11	16.94	0.14
2-SH	14.04	82.3	12.91	1.13	12.94	1.10	82.7	13.55	0.49	13.58	0.46
2-SO_2NH_3	15.44	108.8	15.04	0.40	15.05	0.39	105.6	15.40	0.04	15.40	0.04
3-Substituted (meta)											
3-F	17.40	138.5	17.44	−0.04	17.42	−0.02	131.7	17.52	−0.12	17.48	−0.08
3-Cl	16.88	136.5	17.27	−0.39	17.26	−0.38	127.8	17.20	−0.32	17.17	−0.29
3-Br	17.03	136.1	17.24	−0.21	17.22	−0.19	126.7	17.11	−0.08	17.08	−0.05
3-I	17.27	136.0	17.23	0.04	17.22	0.05	126.1	17.06	0.21	17.03	0.24
3-CN	16.69	131.9	16.90	−0.21	16.89	−0.20	123.9	16.88	−0.19	16.85	−0.16
3-CH_3	18.40	148.3	18.22	0.18	18.19	0.21	140.3	18.21	0.19	18.16	0.24
3-OH	18.41	147.3	18.14	0.27	18.12	0.29	145.4	18.63	−0.22	18.57	−0.16
3-OCH_3	18.18	145.8	18.02	0.16	18.00	0.18	139.4	18.14	0.04	18.09	0.09
3-OC_6H_5	17.39	140.6	17.60	−0.21	17.58	−0.19	127.4	17.16	0.23	17.13	0.26
3-COOH	17.24	137.4	17.35	−0.11	17.33	−0.09	133.2	17.64	−0.40	17.60	−0.36

(Continues)

TABLE 1 (Continued)

		BP-TZVP					BP-TZVPD-FINE				
			Class-specific		Generalized			Class-specific		Generalized	
3-$COCH_3$	18.18	138.6	17.44	0.74	17.42	0.76	131.1	17.47	0.71	17.43	0.75
3-$NHCOCH_3$	18.05	145.6	18.01	0.04	17.98	0.07	140.7	18.24	−0.19	18.19	−0.14
3-NO_2	16.52	129.2	16.69	−0.17	16.68	−0.16	122.3	16.75	−0.23	16.73	−0.21
3-SH	17.51	144.8	17.94	−0.43	17.92	−0.41	137.6	17.99	−0.48	17.95	−0.44
3-SO_2CH_3	16.75	134.0	17.07	−0.32	17.06	−0.31	128.8	17.28	−0.53	17.25	−0.50
3-SO_2NH_2	17.00	133.5	17.03	−0.03	17.02	−0.02	132.0	17.54	−0.54	17.50	−0.50
4-Substituted (para)											
4-F	17.89	142.0	17.72	0.17	17.70	0.19	134.6	17.75	0.14	17.71	0.18
4-Cl	17.47	139.4	17.51	−0.04	17.49	−0.02	130.5	17.42	0.05	17.38	0.09
4-Br	17.47	138.8	17.46	0.01	17.44	0.03	128.3	17.24	0.23	17.20	0.27
4-I	17.56	138.6	17.44	0.12	17.42	0.14	127.3	17.15	0.41	17.12	0.44
4-CN	16.74	130.5	16.79	−0.05	16.78	−0.04	121.4	16.68	0.06	16.66	0.08
4-CH_3	18.33	148.2	18.22	0.11	18.19	0.14	140.0	18.19	0.14	18.14	0.19
4-*tert*-Butyl	18.57	149.9	18.35	0.22	18.33	0.24	136.5	17.90	0.67	17.86	0.71
4-OH	19.04	155.2	18.77	0.27	18.75	0.29	153.4	19.27	−0.23	19.20	−0.16
4-OCH_3	18.77	154.0	18.68	0.09	18.65	0.12	146.8	18.74	0.03	18.68	0.09
4-$COCH_3$	17.28	137.2	17.33	−0.05	17.31	−0.03	128.7	17.27	0.01	17.24	0.04
4-NH_2	19.78	164.7	19.55	0.23	19.51	0.27	165.6	20.26	−0.48	20.17	−0.39
4-NO_2	16.43	127.0	16.51	−0.08	16.50	−0.07	118.7	16.46	−0.03	16.45	−0.02
4-SH	17.92	147.4	18.15	−0.23	18.12	−0.20	139.8	18.17	−0.25	18.12	−0.20
4-SCH_3	18.22	150.8	18.42	−0.20	18.40	−0.18	141.7	18.32	−0.10	18.27	−0.05
4-SO_2CH_3	16.82	132.3	16.93	−0.11	16.92	−0.10	125.4	17.01	−0.19	16.98	−0.16
4-SO_2NH_2	16.90	133.6	17.04	−0.14	17.02	−0.12	130.2	17.39	−0.49	17.36	−0.46
2,6-Disubstituted											
2-CH_3,6-CH_3	17.74	139.7	17.53	0.21	17.51	0.23	131.4	17.49	0.25	17.45	0.29
2-CH_3,6-Cl	16.74	126.9	16.49	0.25	16.49	0.25	117.2	16.34	0.40	16.32	0.42
2-CH_3,6-NO_2	16.13	121.6	16.07	0.06	16.07	0.06	115.3	16.18	−0.05	16.17	−0.04
2-CH_3,6-OCH_3	18.24	145.1	17.97	0.27	17.95	0.29	136.0	17.87	0.37	17.82	0.42
2-Cl,6-NO_2	15.12	109.9	15.13	−0.01	15.13	−0.01	101.8	15.09	0.03	15.10	0.02
2-Cl,6-Cl	15.68	116.9	15.69	−0.01	15.69	−0.01	106.1	15.45	0.23	15.45	0.23
2-NO_2,6-NO_2	14.28	103.3	14.60	−0.32	14.61	−0.33	98.7	14.84	−0.56	14.85	−0.57
2-OCH_3,6-Cl	17.24	132.6	16.95	0.29	16.94	0.30	122.8	16.79	0.45	16.77	0.47
2-OCH_3,6-NO_2	16.65	127.6	16.55	0.10	16.54	0.11	119.2	16.50	0.15	16.48	0.17
2-OCH_3,6-OCH_3	18.65	150.4	18.39	0.26	18.36	0.29	140.1	18.19	0.46	18.14	0.51
3,4-Disubstituted											
3-Br,4-CH_3	17.59	139.8	17.54	0.05	17.52	0.07	129.1	17.30	0.29	17.27	0.32
3-Br,4-OCH_3	17.82	144.7	17.93	−0.11	17.91	−0.09	135.4	17.81	0.01	17.77	0.05
3-CH_3,4-CH_3	18.71	152.1	18.53	0.18	18.50	0.21	142.4	18.38	0.33	18.33	0.38
3-CH_3,4-Cl	17.59	142.9	17.79	−0.20	17.77	−0.18	132.2	17.56	0.03	17.52	0.07
3-CH_3,4-NO_2	16.55	131.1	16.84	−0.29	16.83	−0.28	121.3	16.67	−0.12	16.65	−0.10
3-CH_3,4-OCH_3	19.01	156.2	18.85	0.16	18.82	0.19	147.7	18.81	0.20	18.75	0.26
3-Cl,4-Cl	16.70	132.7	16.96	−0.26	16.95	−0.25	121.1	16.66	0.04	16.63	0.07
3-Cl,4-NO_2	15.78	122.8	16.17	−0.39	16.16	−0.38	112.2	15.94	−0.16	15.93	−0.15
3-NO_2,4-NO_2	15.07	115.4	15.57	−0.50	15.58	−0.51	106.2	15.45	−0.38	15.45	−0.38
3-NO_2,4-CH_3	16.75	131.3	16.85	−0.10	16.84	−0.09	122.1	16.74	0.01	16.72	0.03

TABLE 1 (Continued)

		BP-TZVP					BP-TZVPD-FINE				
			Class-specific		Generalized			Class-specific		Generalized	
3-NO_2,4-Cl	15.98	125.4	16.38	−0.40	16.37	−0.39	115.5	16.20	−0.22	16.19	−0.21
3-NO_2,4-OCH_3	17.04	137.1	17.32	−0.28	17.30	−0.26	129.5	17.33	−0.29	17.30	−0.26
3-OCH_3,4-OCH_3	18.59	154.8	18.75	−0.16	18.72	−0.13	147.8	18.82	−0.23	18.76	−0.17
3-OCH_3,4-CH_3	18.32	150.1	18.37	−0.05	18.35	−0.03	141.9	18.34	−0.02	18.29	0.03
3-OCH_3,4-Cl	17.39	141.7	17.69	−0.30	17.67	−0.28	132.6	17.58	−0.19	17.55	−0.16
3-OCH_3,4-NO_2	16.21	130.1	16.75	−0.54	16.74	−0.53	122.8	16.80	−0.59	16.77	−0.56
3,5-Disubstituted											
3-Br,5-Br	16.23	124.7	16.32	−0.09	16.31	−0.08	112.5	15.96	0.27	15.95	0.28
3-Cl,5-NO_2	15.57	120.3	15.97	−0.40	15.97	−0.40	110.4	15.79	−0.22	15.78	−0.21
3-CH_3,5-CH_3	18.54	148.6	18.24	0.30	18.22	0.32	139.2	18.12	0.42	18.07	0.47
3-CH_3,5-Cl	17.40	138.3	17.41	−0.01	17.40	0.00	128.4	17.25	0.15	17.21	0.19
3-CH_3,5-NO_2	16.62	130.5	16.79	−0.17	16.78	−0.16	121.8	16.71	−0.09	16.69	−0.07
3-CH_3,5-OCH_3	18.26	147.8	18.18	0.08	18.16	0.10	140.0	18.19	0.07	18.14	0.12
3-NO_2,5-NO_2	14.87	112.5	15.33	−0.46	15.34	−0.47	103.7	15.25	−0.38	15.25	−0.38
3-OCH_3,5-OCH_3	18.02	145.2	17.97	0.05	17.95	0.07	138.5	18.06	−0.04	18.01	0.01
3-OCH_3,5-Br	17.03	136.2	17.25	−0.22	17.23	−0.20	127.0	17.13	−0.10	17.10	−0.07
3-OCH_3,5-NO_2	16.31	129.1	16.67	−0.36	16.67	−0.36	121.2	16.66	−0.35	16.64	−0.33
Phenols											
o-Nitro	25.20	142.5	28.64	−3.44	–	–	134.6	28.76	−3.56	–	–
2,4-Dinitro	18.60	95.8	18.69	−0.09	–	–	85.9	19.08	−0.48	–	–
2,5-Dinitro	22.70	113.1	22.37	0.33	–	–	102.2	22.32	0.38	–	–
2,5-Dinitro	25.60	121.8	24.24	1.36	–	–	117.2	25.31	0.29	–	–
2,4,6-Trinitro	9.20	58.8	10.81	−1.61	–	–	46.3	11.23	−2.03	–	–
2,6-Dichloro	26.80	133.0	26.61	0.19	–	–	123.2	26.50	0.30	–	–
2,4,6-Trichloro	25.40	122.1	24.30	1.10	–	–	109.8	23.83	1.57	–	–
2,4,6-Tribromo	24.00	117.0	23.21	0.79	–	–	102.4	22.37	1.63	–	–
Pentachloro	21.20	101.1	19.83	1.37	–	–	87.0	19.30	1.90	–	–
Benzenesulfonamides											
Parent	24.50	173.8	24.40	0.10	24.32	0.18	166.7	24.58	−0.08	24.50	−0.00
p-Fluoro	24.30	171.5	24.18	0.12	24.14	0.16	163.1	24.25	0.05	24.21	0.09
p-Chloro	23.80	170.0	24.03	−0.23	24.02	−0.22	160.5	24.02	−0.22	24.01	−0.21
p-Bromo	24.00	170.9	24.12	−0.12	24.09	−0.09	160.8	24.04	−0.04	24.03	−0.03
m-Chloro	23.70	165.1	23.55	0.15	23.63	0.07	156.0	23.60	0.10	23.65	0.05
3.4-Dichloro	23.10	162.5	23.30	−0.20	23.42	−0.32	151.8	23.21	−0.11	23.31	−0.21
m-Cyano	23.10	161.7	23.23	−0.13	23.36	−0.26	153.3	23.35	−0.25	23.43	−0.33
m-Nitro	22.90	158.2	22.88	0.02	23.07	−0.17	149.6	23.01	−0.11	23.14	−0.24
p-Nitro	22.80	154.6	22.54	0.26	22.79	0.01	144.6	22.56	0.24	22.74	0.06
p-Trifluoromethyl	23.30	163.6	23.42	−0.12	23.51	−0.21	151.2	23.16	0.14	23.27	0.03
p-Methoxy	24.90	179.4	24.95	−0.05	24.77	0.13	171.5	25.02	−0.12	24.88	0.02
m-Methoxy	24.40	173.9	24.41	−0.01	24.33	0.07	164.9	24.42	−0.02	24.36	0.04
p-Methyl	24.80	176.8	24.70	0.10	24.56	0.24	167.6	24.67	0.13	24.57	0.23
m-Methyl	24.70	177.5	24.76	−0.06	24.62	0.08	167.6	24.66	0.04	24.57	0.13
o-Xylyl	25.00	178.1	24.82	0.18	24.66	0.34	168.3	24.73	0.27	24.63	0.37

Note: Experimental values have been taken from References Ludwig (1986)[30]; Pytela (1994)[31]; Kulhánek (1997)[32]; Pytela (1994)[33,34]; Barabanov (1973)[56]; Nowak (1981)[57]; Ludwig (1986)[58] and were shifted as described in Section 2.2. The values $pK_a^{(calc.)}$ are based on LFER parameters that are either specifically fitted for each compound class, referred to as class-specific, or obtained from a generalized fit for a combined set, referred to as generalized, (see text for details). The errors ΔpK_a, that is, the difference between $pK_a^{(exptl.)}$ and $pK_a^{(calc.)}$ are also included.

TABLE 2 Class-specific pK_a LFER parameters c_1 and c_0 with their standard deviations, squared correlation coefficient r^2, and root-mean-square deviation (RMSD) as obtained from linear regressions based on Equation (1) for each compound class at the two levels of theory

	BP-TZVP				BP-TZVPD-FINE			
Class	c_0	c_1	r^2	RMSD	c_0	c_1	r^2	RMSD
Acrylic acids	7.708 ± 1.101	0.393 ± 0.046	0.913	0.357	8.154 ± 1.003	0.390 ± 0.044	0.920	0.343
Benzoic acids	6.279 ± 0.302	0.460 ± 0.013	0.940	0.288	6.853 ± 0.296	0.462 ± 0.013	0.935	0.298
Phenols	−1.708 ± 2.788	1.216 ± 0.140	0.916	1.495	2.046 ± 2.732	1.133 ± 0.150	0.890	1.703
Benzenesulfonamides	7.511 ± 0.891	0.555 ± 0.030	0.963	0.142	9.329 ± 0.843	0.522 ± 0.030	0.959	0.151

TABLE 3 Fit parameters c_1, $c_0^{(A)}$, and $c_0^{(B)}$ together with their error estimates as well as effective squared correlation coefficient r^2 and root-mean-square deviation (RMSD) as obtained from a generalized LFER fit according to Equation (4) for the combined data set consisting of the acrylic acids, benzoic acids, and sulfonamides at the two levels of theory

Level	c_1	$c_0^{(A)}$	$c_0^{(B)}$	r^2	RMSD
BP-TZVP	0.455 ± 0.012	6.369 ± 0.289	10.460 ± 0.363	0.960	0.291
BP-TZVPD-FINE	0.454 ± 0.013	7.009 ± 0.290	11.249 ± 0.364	0.953	0.308

set and anions are known to require more diffusive functions. Similar observations were already made during the prediction of total pK_a values for substituted phenols in water using the CBS-QB3 method in combination with the C-PCM solvation model.[62] There is indeed an indication for this in our data as well since we observed a class-averaged difference in the Gibbs free energy in solution between the anion and the neutral acid for BP-TZVP and BP-TZVPD of about −4 kJ mol^{-1} for the phenols and sulfonamides, but a distinct, almost vanishing difference for the carboxylic acids. This is also in perfect agreement with previous study on the absolute pK_a values of carboxylic acids in water,[63] indicating that the size of the basis set has a much smaller affect for these compounds compared to the phenols.

Apart from that, the linear regressions still yield very similar results in terms of the LFER parameters and correlation coefficients. The slopes are generally slightly smaller at the BP-TZVPD-FINE level, with a decrease that is smallest for the acrylic acids (<1%) and largest for the sulfonamides and phenols (~6%), but overall virtually identical to those at the BP-TZVP level within the error bars. A one-to-one comparison of the intercepts is not meaningful though since this parameter is directly affected by changing the slope. However, the difference in the intercepts between the two levels can be indeed used to quantify the aforementioned horizontal shift in the ΔG_{diss} values by assuming identical slopes, which is a reasonable assumption in this case. Doing so, the shift can be estimated to be on average 6.8 kJ mol^{-1} for the carboxylic acids and 18.3 kJ mol^{-1} for the sulfonamides and phenols. Another important difference between the results for these two levels should also be noted: While the correlation slightly increases for the acrylic acids when moving to the BP-TZVPD-FINE level, it slightly decreases for the benzoic acids and sulfonamides, and noticeably for the phenols. This trend is also reflected in the fit quality as expressed by the root-mean-square deviation (RMSD). The RMSD decreases for the acrylic acids at the higher level, but increases for all other classes and the most for the phenols. In analogy to the BP-TZVP results, if the outliers observed for the linear regression of the benzoic acids are excluded again, we obtain a highly corrected data set ($r^2 = 0.979$) with regression coefficients $c_0 = 6.334$ and $c_1 = 0.484$ at an RMSD of 0.075.

3.2 | Combined data set

The class-specific considerations are useful for additional insights (vide infra), but only of limited use for pK_a prediction for new compounds which should be based on a broad and chemically diverse molecular data set. Therefore, we aimed at a meaningful way to combine the four subsets. Regarding the carboxylic acids, this was easy since the acrylic and benzoic acids have comparable class-specific LFER parameters. Since we already noticed that the benzenesulfonamides mainly differ from the carboxylic acids by a constant offset (cf. Figure 1), which is reflected in a similar slope, including this subset was easily achieved by allowing for a shift in the pK_a scale. However, due to the distinct qualitative difference between these three subsets and the phenols, for example, in terms of the slope c_1, we decided to not include the latter in the determination of generalized LFER parameters.

Thus, for combining the carboxylic acids and the sulfonamides, we performed a constraint linear regression that couples the two individual linear regressions for each class

$$pK_a^{(l)} = c_1^{(l)} \frac{\Delta G_{diss}}{RT\ln(10)} + c_0^{(l)}, \quad (4)$$

where $I = A$ for carboxylic acids and $I = B$ for sulfonamides, by requiring identical slopes, that is, $c_1^{(A)} = c_1^{(B)} = c_1$. Similar to the standard linear regression, this problem can be solved analytically and the solution for the slope c_1 and intercepts $c_0^{(I)}$ reads:

$$c_1 = \frac{N_A\left(\langle X_A Y_A\rangle - \langle X_A\rangle\langle Y_A\rangle\right) + N_B\left(\langle X_B Y_B\rangle - \langle X_B\rangle\langle Y_B\rangle\right)}{N_A\left(\langle X_A^2\rangle - \langle X_A\rangle^2\right) + N_B\left(\langle X_B^2\rangle - \langle X_B\rangle^2\right)}, \tag{5}$$

$$c_0^{(I)} = \langle Y_I\rangle - \langle X_I\rangle c_1, \tag{6}$$

where N_I denotes the number of data points in the respective subset I, X_I the subset of calculated $\frac{\Delta G_{diss}}{RT\ln(10)}$ values, and Y_I the subset of experimental pK_a values. Angle brackets $\langle Q\rangle$ represent the expectation value of a given quantity Q. The corresponding variances were calculated according to:

$$\Delta c_1 = \sqrt{\frac{N_A\sigma_A^2 s_{X_A}^2 + N_B\sigma_B^2 s_{X_B}^2}{\left(N_A s_{X_A}^2 + N_B s_{X_B}^2\right)^2}}, \tag{7}$$

$$\Delta c_0^{(I)} = \sqrt{\frac{\sigma_I^2}{N_I} + \langle X_I\rangle^2 (\Delta c_1)^2} \tag{8}$$

with the standard deviation $s_{X_I} = \sqrt{\langle X_I^2\rangle - \langle X_I\rangle^2}$ of X_I and the estimated variance of the error $\sigma_I^2 = \sqrt{\frac{1}{N_I - 3}\sum_{i=0}^{N_I}\left(y_i^{(I)} - \left(c_0^{(I)} + c_1 x_i^{(I)}\right)\right)^2}$. $x_i^{(I)}$ and $y_i^{(I)}$ are the elements of the sets X_I and Y_I, respectively.

The results for the generalized LFER parameters based on Equations (5–8) are given in Table 3, the fits are plotted in Figure 1 (black lines), and the corresponding calculated pK_a values are listed in columns six and 11 together with their errors in columns seven and 12 of Table 1. Slopes of 0.45 that are identical within the error bars are obtained at both levels of theory, which fit well to that values already obtained in the class-specific LFER fits (cf. Table 2) and match our expectations. Due to the energetic shift in the ΔG_{diss} values between BP-TZVP and BP-TZVPD-FINE, the value for the intercepts $c_0^{(A)}$ and $c_0^{(B)}$ are larger for the latter level. As for the class-specific fits and due to almost identical slopes, the difference in the intercepts can be used to quantify this shift: 8.1 kJ mol^{-1} for the carboxylic acids (using $c_0^{(A)}$) and 10.0 kJ mol^{-1} for the sulfonamides (using $c_0^{(B)}$). These values lie in between the ones from the previous fits. The combined data set still yields an excellent overall correlation at both levels ($r^2 > 0.95$) with sufficiently small RMSDs of approximately only one third of a logarithmic unit on the pK_a scale. The difference of the intercepts $c_0^{(A)} - c_0^{(B)}$ is 4.091 for BP-TZVP and 4.240 for BP-TZVPD-FINE, which are roughly the same considering the RMSDs or statistical errors of $c_0^{(A)}$ and $c_0^{(B)}$ of around 0.3. Thus, we can indeed use a common LFER parameter for the slope c_1 for the combined data set.

The shift in the intercept has two important possible interpretations. On the one hand, this deviation could simply have an experimental origin and be caused by a systematic error, for example, a different calibration with respect to the used pK_a standard that leads to a constant offset on the pK_a scale. However, as already stated before, this should not be the case here since the values for the benzoic acids and sulfonamides were determined by the same authors.[30–34,58] On the other hand, the difference in the intercepts can also reflect errors in the calculated values for ΔG_{diss} that depend mainly on the substance class, as discussed below.

Including the phenols in an even more general LFER parameter fit is not straightforward due to uncertainties for this subset. The main criterion for achieving a good overall correlation is a consistent slope throughout all subsets. In particular, the data points for 2-nitrophenol (p$K_a = 25.20$) and 2,4,6-trinitrophenol (p$K_a = 9.2$), which also roughly represent the two extremes on the phenol pK_a scale, deviate substantially more from the class-specific linear regression line (cf. Table 1 and Figure 1). These values include the overall shift of 2.9 as explained in the Section 2.2. The non-shifted value of 6.3 for 2,4,6-trinitrophenol served as a reference point in the phenol studies and was originally measured conductometrically,[64] while the value of 22.3 for 2-nitrophenol was determined by potentiometry (glass electrode) as the other phenols. To test if the fit for the phenols is flawed by errors from these two outliers, we did for the BP-TZVP values three additional linear regressions: (i) one where 2-nitrophenol was excluded, (ii) one where we excluded 2,4,6-trinitrophenol, and (iii) one where both were excluded. The results from these restricted fits are plotted in Figure S1 and the obtained parameters are listed in Table S1 of the Supporting Information. Excluding 2,4,6-trinitrophenol from the fit decreases the slope from 1.216 to 0.887, which is closer to the value obtained for the carboxylic acids and benzenesulfonamides, but reduces the correlation parameter significantly from $r^2 = 0.916$ to 0.774. Excluding 2-nitrophenol, on the other hand, slightly improves the correlation ($r^2 = 0.987$) but further increases the slope to 1.398, which deviates even more from the slope for the other subsets. If, however, both data points are excluded, the results do not significantly differ from those for the unaltered fit that includes both points, while maintaining a good correlation (0.958). Hence, there is no clear indication that one of these two data points is more inaccurate than the remaining ones and we decided to continue with both included.

3.3 | Analysis of the LFER parameters

For a better comparison, we only refer to our values at the BP-TZVP level. This should also be a reasonable simplification since the results from the generalized fit of the combined data set were found almost identical for both levels of theory except for a constant shift in ΔG_{diss}. We obtained a value of 6.369 ± 0.289 for the intercept, which is to some extent in qualitative agreement with the theoretical value of $\lg[S] = -1.134$. However, one should note that any method-dependent error leading to a constant shift in G_{diss} will alter the intercept, hence complicating possible interpretations. Interestingly, we obtain for the LFER parameter for the slope, for the carboxylic acids

and sulfonamides, a value smaller than one ($c_1 \simeq 0.45$), whereas it is larger than and closer to one for the phenols ($c_1 \simeq 1.22$). In previous studies that used COSMO-RS, slopes of 0.58 (for acids)[24] and 0.75 (for bases)[25] were reported for aqueous solutions, while slopes of 0.70 (DMSO)[24] and 0.92 (acetonitrile)[26] were found in organic solvents. This deviation from the ideal slope, $c_1 = 1$, corresponds to an apparent, systematic overestimation of ΔG_{diss}. This phenomenon is not unique to COSMO-RS, but occurs in all first-principles pK_a predictions using simple DCSMs unless significant, but often physically questionable effort is taken to adjust the DCSM parameters (e.g., as pointed out by Chipman[65]). However, it has been argued that including a few explicit solvent molecules does not necessarily lead to better results in terms of increasing the slope toward one since this creates another bias from the exact configuration of these micro-solvated structures.[24] Such hybrid-solvation approaches typically lead to ambiguities and problems in practice: (i) especially for organic solvents, no obvious number of explicit solvent molecules in the solvation shell and (ii) optimization of the solute-solvent structure.[26] Also, the exact correspondence of the experimental pK_a scales to the free energy of dissociation at infinite dilution had been questioned.[24,25]

Reference [26] discussed that the degree of deviation from the ideal behavior is largely solvent-dependent and has its main origin in the solvent's capability of solvating anions, which is also a relevant factor regarding the partitioning of data sets (vide infra). In water, which has the highest anion-solvating ability among the solvents studied with COSMO-RS, the slope has the lowest value (i.e., deviates most from the ideal case), it then increases for DMSO (with moderate stabilization), and finally reaches the largest and closest value for acetonitrile (with lowest stabilization). Precisely quantifying this anion-stabilization ability is difficult since it is a consequence of the overall solvent polarity, a concept that is vividly discussed in the literature.[66] However, we infer that the hydrogen-bond donation ability, for example, as described by the Kamlet-Taft solvent parameter[67] α and the relative dielectric constant ϵ_r, entering the energy expression for Coulomb interactions, should have major effects. Values of α and ϵ_r are in the following taken from Reference [68]. Whereas water is clearly the most polar solvent in terms of both descriptors ($\alpha = 1.17$, $\epsilon_r = 78.3$), the two contributions show opposite trends for DMSO ($\alpha = 0.00$, $\epsilon_r = 46.5$) and acetonitrile ($\alpha = 0.19$, $\epsilon_r = 35.9$). The order of the solvents (water–DMSO–acetonitrile) as claimed by the authors would thus hint to the dielectric constant having the dominant effect as the more important property, which is reasonable considering the stabilization energy of ions.

Relating to this work, the current results seem to fall even below the water regime (slope of 0.58 for acids) based on our slope of 0.45 from the generalized fit for the combined data set that includes the carboxylic acids and benzenesulfonamides. In terms of the two polarity descriptors, acetone ($\alpha = 0.08$, $\epsilon_r = 20.56$) is expected to have an anion stabilization in between that of DMSO and acetonitrile. Assuming that ϵ_r is the dominant contribution, the stabilizing effect of acetone should be even smaller compared to that of acetonitrile. However, our results do not fit into this order, which indicates that also other factors seem to affect the slope.

In the current work, we observe a dissection of the molecular data set that at first may appear to be comparable to the observations made for acetonitrile in Reference [26]. The latter study also reported differences between LFER parameters for different molecular classes that could be dealt with by fixing the slope and adjusting a constant offset, and the physical origin was assigned to the degree of charge (de-)localization in the anions. More specifically, the data set reported in Reference [26] is dissected into compounds that newly form an aromatic system after deprotonation (i.e., CH acids) and those with ordinary acidic functional groups where the charge will remain more or less localized (e.g., phenols, carboxylic acids, sulfonic acids, etc.). This idea can also not fully explain the discrepancy observed in our study though since we identified the phenols, that is, strongly delocalized systems as the main exception here. In particular, the phenols are expected to behave very similarly compared to the benzoic acids in terms of charge delocalization via resonance stabilization. However, this idea might actually give us an explanation for the sulfonamides whose deviation from the carboxylic acids was indeed successfully accounted for by a constant shift in the pK_a scale while retaining a common slope (vide supra). The sulfonamides are indeed structurally different from the carboxylic acids and the phenols since both the protonated and the deprotonated structures were found to be non-planar. Therefore, the negative charge of the deprotonated species cannot be stabilized via resonance and remains localized in a lone electron pair at the nitrogen atom of the functional group. Regarding the phenols, we also infer that the zero-point vibrational energy might have a larger contribution compared to the other classes, assuming that the degree of conjugation is much higher for the phenols and hence the vibrations would depend stronger on structural differences that are induced by the deprotonation.

3.4 | Special behavior of the phenol subset

For the phenols, it is unclear if the large deviation from the other subsets is due to a systematic error in the experimental data or has an actual physical origin. It does seem that they have a special role in this data set though. For them, the slope is closest to the theoretical value of one, but their results are also affected the most by changing the computational level. In this context, we note that this subset is also the one that covers the largest range of pK_a values. Hence, the absolute differences are naturally much larger when comparing it to the other subsets with relatively smaller pK_a ranges.

In a previous COSMO-RS study on aqueous pK_a, the underlying computational methods of the BP-TZVP level, that was also invoked in this work, were already suspected to be insufficient for describing the structural changes of the related compound 4-nitrophenol upon reorganization after deprotonation.[24] For that compound it was found that the deprotonated species adapts a chinoidal form as indicated by a pronounced shortening of the C—O bond lengths, which is accompanied by an overall stronger ionization. A chinoidal form is based on simple resonance considerations equally possible for the

2-nitrophenol and hence our observations might have indeed the same physical origin.

4 | CONCLUSIONS

For the 120 organic acids considered in the current work in the solvent acetone, class-specific LFER fits yield very good linear correlations between ΔG_{diss} computed with COSMO-RS and DFT and the experimental pK_a values within each compound class. The acrylic and benzoic acids could be easily combined with the benzenesulfonamide subset by allowing for an additional offset in the experimental pK_a scale. The resulting combined data set is still highly linearly correlated ($r^2 > 0.95$) and has the same slope of 0.45 at the two levels of theory BP-TZVP and BP-TZVPD-FINE, respectively. That this slope is below the ideal value of 1 is in agreement with previous observations made in the literature[24–26] for water, DMSO, and acetonitrile, although the values do not fit to an ordering of the solvents that would be expected if the deviations from the ideal value of 1 were solely determined by the dielectric constant or the hydrogen-bond donation ability. The constant shift in the pK_a scale between the two subsets reaches at both levels ~4 within the standard deviation. Using these LFER parameters, the mean square deviation with respect to the experimental pK_a values is only ~0.3.

In contrast to the carboxylic acids and benzenesulfonamides, the phenols revealed to be distinctly different. Especially the slope appeared to be larger than one and thus we did not find a natural way to combine this subset with the other ones. Moreover, there seems to be a higher uncertainty in this subset as the data points on the two extremes of the pK_a scale (2-nitrophenol and 2,4,6-trinitrophenol) rather appear to have outlier character. Consequently, these points can affect the LFER parameters significantly if one or both of them are excluded from the fitting process. Having both included, the RMSD of pK_a prediction amounts to 1.7–1.9 for the phenols (with $r^2 \sim 0.9$), which is still satisfying considering the broad range from 9.2 to 26.8. A substantial increase up to almost perfect linear correlation ($r^2 = 0.987$) indicates that excluding 2-nitrophenol from the data set might improve the fit quality. The bad performance of 2-nitrophenol seems to be comparable to problems with the related compound 4-nitrophenol in a preceding study[24] of pK_a in water that was assigned to insufficient accuracy of the underlying electronic structure methods.

ACKNOWLEDGMENTS

We are grateful to an unknown reviewer for pointing out to us References [30, 31, and 33]. We acknowlegde financial support by the Deutsche Forschungsgemeinschaft through project Ha 2588/8-1. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2033 - 390677874 - RESOLV. The authors declare the following competing financial interest: Arnim Hellweg is employee of Dassault Systèmes. Dassault Systèmes commercially distributes TURBOMOLE, COSMOtherm, and COSMOconf within the BIOVIA brand.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.5676011, reference number 5676011.

ORCID

Niklas Sülzner https://orcid.org/0000-0002-8734-5807
Julia Haberhauer https://orcid.org/0000-0002-0100-1959
Christof Hättig https://orcid.org/0000-0002-5752-2710
Arnim Hellweg https://orcid.org/0000-0002-9508-4889

REFERENCES

[1] S. Hammes-Schiffer, *J. Phys. Chem. B* **2021**, *125*, 3725.
[2] Y.-Y. Ren, S.-F. Zhu, Q.-L. Zhou, *Org. Biomol. Chem.* **2018**, *16*, 3087.
[3] P. A. Ash, R. Hidalgo, K. A. Vincent, *ACS Catal.* **2017**, *7*, 2471.
[4] M. Shabani, H. Younesi, M. Ponti, A. Rahimpour, M. Rahimnejad, A. A. Zinatizadeh, *J. Clean. Prod.* **2020**, *264*, 121446.
[5] E. F. Caldin, V. Gold, *Proton-Transfer Reactions*, Springer, Boston, MA **2013**.
[6] J. Reijenga, A. van Hoof, A. van Loon, B. Teunissen, *Anal. Chem. Insights* **2013**, *8*, ACI.S12304.
[7] B. Pathare, V. Tambe, V. Patil, *Int. J. Pharm. Pharm. Sci.* **2014**, *6*, 26.
[8] L. Barcza, Á. Buvári-Barcza, *J. Chem. Educ.* **2003**, *80*, 822.
[9] Á. Buvári-Barcza, L. Barcza, *Pharmazie* **2005**, *60*, 243.
[10] L. Soovali, I. Kaljurand, A. Kütt, I. Leito, *Anal. Chim. Acta* **2006**, *566*, 290.
[11] K. S. Alongi, G. C. Shields, in *Annual Reports in Computational Chemistry*, Vol. 6 (Ed: R. A. Wheeler), Elsevier, **2010**, p. 113.
[12] P. G. Seybold, G. C. Shields, *WIREs Comput. Mol. Sci.* **2015**, *5*, 290.
[13] B. Mennucci, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 386.
[14] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999.
[15] A. Klamt, G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.
[16] A. Schäfer, A. Klamt, D. Sattel, J. C. Lohrenz, F. Eckert, *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187.
[17] A. Klamt, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 699.
[18] J. R. Pliego, *Chem. Phys. Lett.* **2003**, *367*, 145.
[19] R. Casasnovas, J. Ortega-Castro, J. Frau, J. Donoso, F. Muoz, *Int. J. Quantum Chem.* **2014**, *114*, 1350.
[20] A. Klamt, *J. Phys. Chem.* **1995**, *99*, 2224.
[21] A. Klamt, V. Jonas, T. Bürger, J. C. Lohrenz, *Chem. A Eur. J.* **1998**, *102*, 5074.
[22] F. Eckert, A. Klamt, *AIChE J* **2002**, *48*, 369.
[23] A. Klamt, F. Eckert, W. Arlt, *Annu. Rev. Chem. Biomol. Eng.* **2010**, *1*, 101.
[24] A. Klamt, F. Eckert, M. Diedenhofen, M. E. Beck, *Chem. A Eur. J.* **2003**, *107*, 9380.
[25] F. Eckert, A. Klamt, *J. Comput. Chem.* **2006**, *27*, 11.
[26] F. Eckert, I. Leito, I. Kaljurand, A. Kütt, A. Klamt, M. Diedenhofen, *J. Comput. Chem.* **2009**, *30*, 799.
[27] BIOVIA COSMOtherm, Release **2020**; Dassault Systèmes. http://www.3ds.com.
[28] K. Izutsu, *Acid-base dissociation constants in dipolar aprotic solvents, 1st edn.* Blackwell Scientific Publications, Oxford. **1990**.
[29] A. M. Slater, *J. Comput. Aided Mol. Des.* **2014**, *28*, 1031.
[30] M. Ludwig, V. Baron, K. Kalfus, O. Pytela, M. Večeřa, *Collect. Czechoslov. Chem. Commun.* **1986**, *51*, 2135.
[31] O. Pytela, J. Liška, *Collect. Czechoslov. Chem. Commun.* **1994**, *59*, 2005.
[32] J. Kulhánek, O. Pytela, *Collect. Czechoslov. Chem. Commun.* **1997**, *62*, 913.
[33] O. Pytela, J. Kulhánek, M. Ludwig, V. Říha, *Collect. Czechoslov. Chem. Commun.* **1994**, *59*, 627.
[34] O. Pytela, J. Kulhánek, M. Ludwig, *Collect. Czechoslov. Chem. Commun.* **1994**, *59*, 1637.

[35] J. Knorr, N. Sülzner, B. Geissler, C. Spies, A. Grandjean, R. J. Kutta, G. Jung, and P. Nuernberger **2022**. In Press.
[36] N. Sülzner, B. Geissler, A. Grandjean, G. Jung, and P. Nuernberger (**2022**), https://doi.org/10.1002/cptc.202200041R2
[37] S. K. Khani, A. M. Khah, C. Hättig, *Phys. Chem. Chem. Phys.* **2018**, *20*, 16354.
[38] S. K. Khani, A. M. Khah, C. Hättig, *J. Chem. Theory Comput.* **2020**, *16*, 4554.
[39] A. Hellweg, F. Eckert, *AIChE J* **2017**, *63*, 3944.
[40] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, R. Ahlrichs, *Chem. Phys. Lett.* **1995**, *240*, 283.
[41] O. Treutler, R. Ahlrichs, *J. Chem. Phys.* **1995**, *102*, 346.
[42] TURBOMOLE V7.3, A development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH 2007, **2018**; http://www.turbomole.com.
[43] BIOVIA COSMOconf, Release 2020; Dassault Systèmes. http://www.3ds.com.25
[44] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.
[45] J. P. Perdew, *Phys. Rev. B* **1986**, *33*, 8822.
[46] A. Schäfer, C. Huber, R. Ahlrichs, *J. Chem. Phys.* **1994**, *100*, 5829.
[47] K. Eichkorn, F. Weigend, O. Treutler, R. Ahlrichs, *Theor. Chem. Accounts* **1997**, *97*, 119.
[48] D. Rappoport, F. Furche, *J. Chem. Phys.* **2010**, *133*, 134105.
[49] A. Klamt, M. Diedenhofen, *J. Comput. Chem.* **2018**, *39*, 1648.
[50] N. Sülzner, J. Haberhauer, C. Hättig, and A. Hellweg, Raw Data to "Prediction of acid pKa values in the solvent acetone based on COSMO-RS" (**2022**), https://doi.org/10.5281/zenodo.5676011.
[51] J. J. Klicić, R. A. Friesner, S.-Y. Liu, W. C. Guida, *Chem. A Eur. J.* **2002**, *106*, 1327.
[52] F. Himo, L. Noodleman, M. R. A. Blomberg, P. E. M. Siegbahn, *Chem. A Eur. J.* **2002**, *106*, 8757.
[53] L. Rottmannová, P. Škorňa, J. Rimarčík, V. Lukeš, E. Klein, *Acta Chim. Slov.* **2013**, *6*, 60.
[54] J. Rimarčík, V. Lukeš, E. Klein, M. Ilčin, *J. Mol. Struct. THEOCHEM* **2010**, *952*, 25.
[55] I. A. Topol, G. J. Tawa, S. K. Burt, A. A. Rashin, *J. Chem. Phys.* **1999**, *111*, 10998.
[56] V. P. Barabanov, V. M. Tsentovskii, A. Y. Tret'yakova, *Russ. J. Phys. Chem.* **1973**, *47*, 1304.
[57] B. Nowak, Z. Pawlak, *Pol. J. Chem.* **1981**, *55*, 2407.
[58] M. Ludwig, O. Pytela, M. Večeřa, *Collect. Czechoslov. Chem. Commun.* **1986**, *51*, 1948.
[59] M. Foltin, P. Majer, *Collect. Czechoslov. Chem. Commun.* **1978**, *43*, 95.
[60] A. Kütt, I. Leito, I. Kaljurand, L. Sooväli, V. M. Vlasov, L. M. Yagupolskii, I. A. Koppel, *J. Org. Chem.* **2006**, *71*, 2829.
[61] A. Kütt, S. Tshepelevitsh, J. Saame, M. Lõkov, I. Kaljurand, S. Selberg, and I. Leito, *Eur. J. Org. Chem.* **2021**, 1407, *2021*.
[62] M. D. Liptak, K. C. Gross, P. G. Seybold, S. Feldgus, G. C. Shields, *J. Am. Chem. Soc.* **2002**, *124*, 6421.
[63] M. D. Liptak, G. C. Shields, *J. Am. Chem. Soc.* **2001**, *123*, 7314.
[64] C. French, I. Roe, *Trans. Faraday Soc.* **1953**, *49*, 314.
[65] D. M. Chipman, *Chem. A Eur. J.* **2002**, *106*, 7413.
[66] C. Reichardt, *Chem. Rev.* **1994**, *94*, 2319.
[67] M. J. Kamlet, J. L. M. Abboud, M. H. Abraham, R. Taft, *J. Org. Chem.* **1983**, *48*, 2877.
[68] C. Spies, B. Finkler, N. Acar, G. Jung, *Phys. Chem. Chem. Phys.* **2013**, *15*, 19893.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: N. Sülzner, J. Haberhauer, C. Hättig, A. Hellweg, *J. Comput. Chem.* **2022**, 1. https://doi.org/10.1002/jcc.26864