Calculating pH
A literature survey

Mayank Modi

Breakdown of approaches

1. Thermodynamic approach using first principle
2. Empirical approaches that relate certain metrics of the system to pKa

Note: All the approaches use pKa or pKb to calculate pH

First Principle Approaches

1. By dissociation energy calculation

Prediction of the pKa of Carboxylic Acids Using the ab Initio Continuum-Solvation Model PCM-UAHF (1998) **by** Schuurmann,Cossi, Barone, and Tomasi [1]

- Free energy of solvation of gas phase molecular structures calculated using PCM-UAHF in Gaussian 94 software **(eq 7)** $$\Delta\Delta G_s = \Delta G_s(A^-) - \Delta G_s(AH) + \Delta G_s(H_3O^+) - \Delta G_s(H_2O) \quad (7)$$
- Proton transfer energy calculated at SCF and MP2 levels in gas phase and solution phase **(eq 11 and 12)**

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

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

Results

- pKa range (0.89-5.05)
- The precision of calculated free energies of dissociation is **not sufficient** for the prediction of **absolute pKa values**.
- The results with 16 aliphatic carboxylic acids suggest that, within chemical classes, **experimental trends** of pKa can be **well reproduced** when using PCM-UAHF for the solvation contribution to compound acidity
- The gas-phase portion of the proton-transfer energy is apparently better described with 6-31G** and 6-31+G** than with the considerably greater basis sets 6-311G(2d,2p) and 6-311+G(2d,2p), respectively, and better at the SCF and SCFfree energy level than with MP2.

- The average ΔG_{aq}^{SCF} values are 68 (dsp), 45 (dsp+), 93 (ts2p), and 88 kJ/mol (ts2p+) and thus exceed the experimental average by **factors of 1.6-3.3**, and the calculated solution-phase acidity variations of 65 (dsp), 55 (dsp+), 97 (ts2p), and 83 kJ/mol (ts2p+) are greater by **factors of 2.3-4.1** as compared to the experimental range.
- The **failure to predict the correct magnitudes** of solution-phase dissociation energies is mainly caused by apparent deficiencies in quantifying the gas-phase portion properly.

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

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

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

Alternate approach - linear regression

With gas-phase energies as regression parameters, the following general observations can be noted: The results with dsp and dsp+ are clearly superior to the ones with ts2p and ts2p+,the SCF level yields better predictions of pKa than the MP2 level, and ΔGSCF and ΔESCF show very similar performances.

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

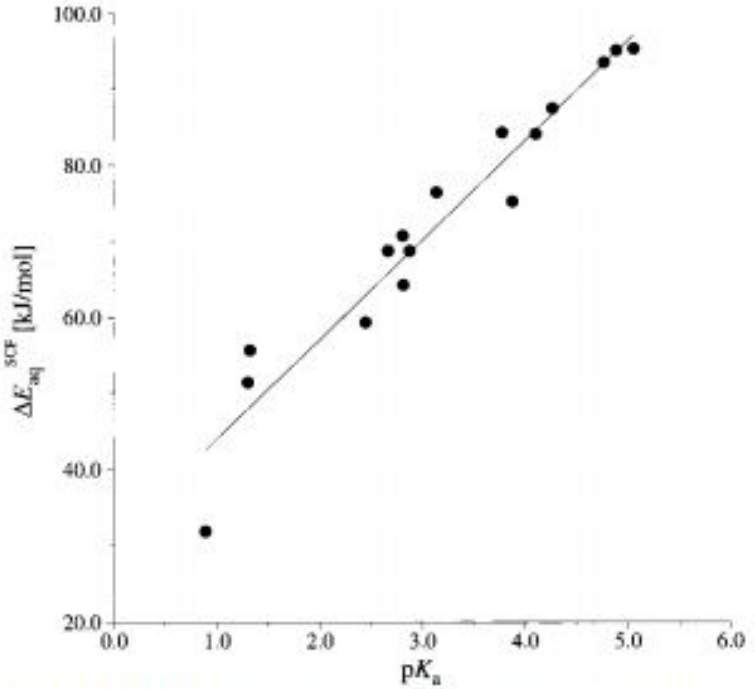


Figure 1. Calculated dissociation energy in aqueous solution ($\Delta E_{aq}{}^{SCF}$) versus experimental pK_a using PCM-UAHF//dsp (see Materials and Methods) together with the linear regression line according to eq 15.

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

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

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

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

With eq 15, the greatest overestimation and underestimation of pKa are found for nitroacetic acid (+0.65 pKa units) and trichloroacetic acid (-0.61 pKa units),

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

the greatest errors with eq 17 are observed with R-chloropropionic acid (+0.42 pKa units) and cyanoacetic acid (-0.33 pKa units).

Empirical Approaches

1. By calculating Q properties

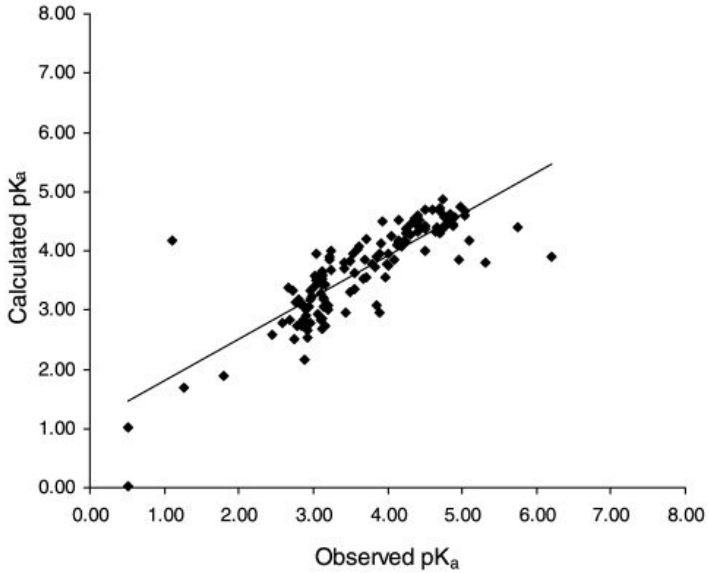
Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. (2002) **by** Tehan,Lloyd, Wong, and team [2]

- All the structures were initially extracted as 2D SDfiles and converted into 3D models using Corina
-

- QM descriptors (equations included in paper):
  - Electrophilic frontier electron density (FE)
  - Nucleophilic frontier electron density (FN)
  - Electrophilic superdelocalisability (SE)
  - Nucleophilic superdelocalisability (SN)
  - Radical superdelocalisability (SR)
  - Atom self-polarisability (ALP)
  - Partial atomic charges (AQ)
  - Energies of the HOMO and LUMO

Results

- Results(Aliphatic Carboxylic Acids):
  - (all - 185 molecules)
    - pKa= 4.29*ALP1 - 41.77*AQ2 - 30.04*AQ3+0.71*FE3 +56.06 **(eq 11)**
    - r =0.83, r^2 = 0.69, r_{cv}^2 =0.67, F= 101.28, s= 0.564
  - (excluding amino acids - 143 molecules)
    - pKa =3.24*ALP3 -2.80*SE3+ 19.43 (**eq 12)**
    - r= 0.84, r^2 = 0.70, r_{cv}^2 0.69, F =165.80, s=0.510
  - (excluding amino acids - 141 molecules)
    - pKa =3.53*ALP3 -2.65*SE3+ 25.01 (**eq 13)**
    - r= 0.90, r^2 = 0.70, r_{cv}^2 =0.69, F =165.80, s=0.510



Plot of calculated vs. observed pK_a for the aliphatic acid dataset, using Eq. 12 from Table 7.

1. Using *MULTICASE*[4]

Application of the multiple computer automated structure evaluation methodology to a quantitative structure–activity relationship study of acidity (1994) **by** Klopman, Fercu [3]

- Uses MULTICASE program to analyse the relationship with organic acid structures and their (first) pka value.
- There are no implementation details about the MULTICASE AI algorithm, just an explanation of what it does. Couldn’t locate a source code either. Although the paper say s that they are available from Professor Klopman.

Data

1. 2464 molecules
2. Contains 242 aliphatic and alicyclic carboxylic acids
3. More details*(breakdown of types)* in database section of paper
4. The program can only handle neutral species therefore zwitter ions are entered into their training data in neural form.
5. Only most stable tautomers were used - usually with the weaker acidic proton.

Input

1. Structural formula
2. Activities (pKb in this paper)

Method

1. All molecular structures are fragmented into all possible substructures (descriptors) of 2 to 10 linearly connected heavy atoms, which are labeled as active or inactive depending on whether the parent molecule is acidic.
2. Fragments with highest probability of being responsible for acidity are considered to perform multivariate regression analysis.
3. pKb is calculated using the coefficients obtained in the QSAR (Quantitative Structure Activity Relationship) equation.

Modulator	r_i	n_i	M_i	$r_i(n_iM_i)$
F — C — CO — OH	+8.8	2	1	17.6
Cl — C — CO — OH	+8.3	1	1	8.3
log P	−1.6	1	1.05	−1.7
Hardness, $(\varepsilon_{HOMO} - \varepsilon_{LUMO})/2$	+5.1	1	1.14	−5.8
Total				18.4

Constant = 56.8;
$pK_b = 56.8 + 18.4 = 75.2$ Multi-CASE units = 13.4 pK units;
$pK_a = 14.0 - pK_b = 0.6$;
cf. experimental $pK_a = 0.46$;
error = 0.14 pK units.

SCHEME 1. The multi-CASE prediction of the acidity in the case of chlorodifluoroacetic acid.

Results

TABLE I.
Statistical Parameters of the Multi-CASE Predictions of the Acidity.

Set	Training / Test	Phi^2	OC, %	Sens. %	Spec. %	*r*	SD
1	616 / 1848	0.861	97.0	90.5	99.7	.925	0.907
2	1232 / 1232	0.902	97.9	93.8	99.6	.942	0.831
3	1848 / 616	0.897	97.8	94.3	99.3	.952	0.774

Phi^2 measures the accuracy of the predictions with respect to expectations from acidity randomness and equals 1 for a perfect fit. Observed concordance (OC) is the ratio of the sum of true positives and true negatives divided by the total number of predictions. Sensitivity (sens) represents the probability of an experimentally active compound to be predicted active. Specificity (spec) renders the probability of an experimentally inactive compound to be predicted inactive. The *r* value is the correlation coefficient between the experimental and predicted acidities and is 0 for no correlation and 1 for a perfect one. The SD value is the standard deviation.

References

[1] Prediction of the pKa of Carboxylic Acids Using the ab Initio Continuum-Solvation Model PCM-UAHF

[2] Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids

[3] Application of the multiple computer automated structure evaluation methodology to a quantitative structure–activity relationship study of acidity

[4] MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program

References

- Experiment stands corrected: accurate prediction of the aqueous pKa values of sulfonamide drugs using equilibrium bond lengths
- https://researchoutreach.org/articles/physical-sciences/pka-prediction-from-ab-initio-calculations/
- pKa Prediction from an ab initio bond length: part 2—phenols
- The AIBLHiCoS Method: Predicting Aqueous pKa Values from GasPhase Equilibrium Bond Lengths