Predicting the pK_a of Small Molecules

Matthias Rupp[*,1,2], Robert Körner[1] and Igor V. Tetko[1]

[1]*Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany*

[2]*Present Address: Machine Learning Group, Technische Universität Berlin, FR 6-9, Franklinstr. 28/29, D-10587 Berlin, Germany, and Institute for Pure and Applied Mathematics, University of California at Los Angeles, 460 Portola Plaza, Los Angeles, CA 90095-7121, USA*

Abstract: The biopharmaceutical profile of a compound depends directly on the dissociation constants of its acidic and basic groups, commonly expressed as the negative decadic logarithm pK_a of the acid dissociation constant (K_a). We survey the literature on computational methods to predict the pK_a of small molecules. In this, we address data availability (used data sets, data quality, proprietary versus public data), molecular representations (quantum mechanics, descriptors, structured representations), prediction methods (approaches, implementations), as well as pK_a-specific issues such as mono- and multiprotic compounds. We discuss advantages, problems, recent progress, and challenges in the field.

Keywords: pK_a, acid dissociation constant, QSPR, quantitative structure-property relationships.

1. INTRODUCTION

The acid dissociation constant (also protonation or ionization constant) K_a is an equilibrium constant defined as the ratio of the protonated and the deprotonated form of a compound; it is usually stated as p$K_a = -\log_{10} K_a$. The pK_a value of a compound strongly influences its pharmacokinetic and biochemical properties. Its accurate estimation is therefore of great interest in areas such as biochemistry, medicinal chemistry, pharmaceutical chemistry, and drug development. Aside from the pharmaceutical industry, it also has relevance in environmental ecotoxicology, as well as the agrochemicals and specialty chemicals industries. In this work, we survey approaches to the computational estimation of pK_a values of small compounds in an aqueous environment. For related aspects like the prediction of pK_a values of proteins, the prediction of pK_a values in solvents other than water, or, the experimental determination of pK_a values, we refer to the literature (Table **1**).

1.1. History

1.1.1. Quantitative Structure-Property Relationships

The empirical estimation (as opposed to *ab initio* calculations) of pK_a values belongs to the field of quantitative structure-property relationships (QSPR). The basic postulate in QSPR modeling (and the closely related field of quantitative structure-activity relationships, QSAR) is that a compound's physico-chemical properties are a function of its structure as described by (computable) features. The idea that physiological activity of a compound is a (mathematical) function of the chemical composition and constitution of the compound dates back at least to the work by Brown and Fraser [9] in 1868. Major break-throughs include the work by Louis Hammett, who established free energy relationships for equilibrium constants of meta- and para-substituted benzoic acids, $\log K / K_0 = \sigma\rho$, where K_0, K are the equilibrium constants of the substituted and unsubstituted compound, and, σ and ρ are constants depending only on the substituent and the reaction, respectively [10, 11]. Robert Taft modified this equation by separating steric from polar and resonance effects [12]. Later, Corwin Hansch and Toshio Fujita introduced an additional parameter $\pi = \log P - \log P_0$ for the substituent effect on hydrophobicity, where P, P_0 are the octanol-water partition ratios of the substituted and unsubstituted compound [13, 14]. In the same year, Spencer Free and James Wilson [15, 16] published a closely related approach, later improved by Toshio Fujita and Takashi Ban [17], with structural features (presence and absence of substituents) instead of experimentally determined properties.

Table 1. Reviews of pK_a Prediction and Related Topics, Sorted by Year and First Author Name. ADME = Absorption, Distribution, Metabolism, and Excretion

Ref.	Author (Year)	Coverage/Emphasis
[1]	Ho and Coote (2010)	Continuum solvent pK_a calculations
[2]	Cruciani *et al.* (2009)	pK_a prediction and ADME profiling
[3]	Lee and Crippen (2009)	Prediction of pK_a values (proteins and small molecules)
[4]	Manallack (2007)	Distribution of pK_a values in drugs
[5]	Fraczkiewicz (2006)	In silico prediction of ionization (theory and software)
[6]	Wan and Ulander (2006)	High-throughput pK_a screening, and pK_a prediction
[7]	Tomasi (2005)	Quantum mechanical continuum solvation models
[8]	Selassie (2003)	History of quantitative structure-property relationships

*Address correspondence to this author at Technische Universität Berlin, FR 6-9, Franklinstr. 28/29, 10587 Berlin, Germany; Tel: ++49-30-31424927; Fax: ++49-30-31478622; E-mail: mrupp@mrupp.info

1386-2073/11 $58.00+.00

1.1.2. pK_a Estimation

QSPR studies involving pK_a values were published in the early 1940s [18, 19]. Since then, a vast number of books, book chapters, conference contributions, and journal articles have been published on the topic (Section 3).

1.2. Definition

1.2.1. pK_a-Values

According to the Brønsted-Lowry theory of acids and bases, an acid HA is a proton (hydrogen cation) donor, $HA \rightleftharpoons H^+ + A^-$, and a base B is a proton acceptor, $B + H^+ \rightleftharpoons BH^+$. For a weak acid in aqueous solution, the dissociation $HA + H_2O \rightleftharpoons A^- + H_3O^+$ is reversible. In the forward reaction, the acid HA and water, acting as a base, yield the conjugate base A^- and oxonium H_3O^+ (protonated water) as conjugate acid. In the backward reaction, oxonium acts as acid and A^- as base. The corresponding equilibrium constant [20], known as the *acid dissociation constant* K_a, is the ratio of the activities of products and reagents,

$$K_a = \frac{a(A^-)a(H_3O^+)}{a(HA)a(H_2O)}, \tag{1}$$

where $a(\cdot)$ is the activity of a species under the given conditions. The form of Equation 1 follows from the law of mass action for elementary (one-step) reactions like the considered proton transfer reaction. Activity is a measure of "effective concentration", a unitless quantity defined in terms of chemical potential [21, 22], and can be expressed relative to a standard concentration:

$$a(x) = \exp\left(\frac{\mu(x) - \mu^{\ominus}(x)}{RT}\right) = \gamma(x)\frac{c(x)}{c^{\ominus}}, \tag{2}$$

where $\mu(\cdot)$ is the chemical potential of a species under the given conditions (partial molar Gibbs energy[1]), $\mu^{\ominus}(\cdot)$ is the chemical potential of the species in a standard state (molar Gibbs energy), $R = 8.314472(15)\,\mathrm{JK^{-1}mol^{-1}}$ is the gas constant, T is the temperature in kelvin, $\gamma(\cdot)$ is a dimensionless activity coefficient, $c(\cdot)$ is the molar (or molal) concentration of a species, and, $c^{\ominus} = 1\,\mathrm{mol/L}$ (or $1\,\mathrm{mol/kg}$) is a standard concentration. Values of $\gamma(\cdot) \neq 1$ indicate deviations from ideality. Note that the activity of an acid can depend on its concentration [24]. In an ideal solution $\gamma(\cdot) = 1$, and effective concentrations equal analytical ones. With the assumptions $\gamma(\cdot) = 1$ and $c(H_2O) = c^{\ominus} = 1$ mol/L, inserting Equation 2 into Equation 1 yields an approximation valid for low concentrations of HA in water:

$$K_a \approx \frac{c(A^-)c(H_3O^+)}{c(HA)c^{\ominus}}. \tag{3}$$

Taking the negative decadic logarithm $pK_a = -\log_{10}(K_a)$ yields the Henderson-Hasselbalch [25] equation

$$pK_a \approx pH + \log_{10}\frac{c(HA)}{c(A^-)}, \tag{4}$$

where $\mathrm{pH} = -\log_{10}a(H_3O^+) \approx -\log_{10}(c(H_3O^+)/c^{\ominus})$. In an ideal solution, the pK_a of a monoprotic weak acid is therefore the pH at which 50% of the substance is in deprotonated form, and Equation 4 is an approximation of the mass action law applicable to low-concentration aqueous solutions of a single monoprotic compound [26, 27].

1.2.2. pK_b-Values

The protonation of a base $B + H_2O \rightleftharpoons HB^+ + HO^-$ can be described in the same terms as the deprotonation of an acid, leading to the *base association constant* $K_b = a(HB^+)a(HO^-)/(a(B)a(H_2O))$. Adding the reaction equations for the deprotonation of HA and the protonation of its conjugate base A^- gives $2H_2O \rightleftharpoons H_3O^+ + OH^-$, with equilibrium constant $K_w = a(H_3O^+)a(OH^-)/a^2(H_2O)$. It follows that $K_w = K_a K_b$, and therefore $pK_b = pK_w - pK_a \approx 14 - pK_a$, where $pK_w \approx 14$ from $c(H_3O^+) = c(HO^-) \approx 10^{-7}$ mol/L at $T = 298.15\,\mathrm{K}$ and under the same assumptions as for Equation 3. Since pK_a and pK_b use the same scale, pK_a-values are used for both acids and bases; however, data in older references is sometimes given as pK_b-values. For prediction, one should not mix pK_a and pK_b values.

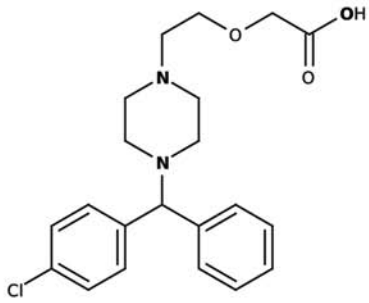
1.2.3. Multiprotic Compounds

A *multiprotic* (also *polyprotic*) compound has more than one ionizable center, i.e., it can donate or accept more than one proton. For n protonation sites, there are 2^n microspecies (each site is either protonated or not, yielding 2^n combinations) and $n2^{n-1}$ micro- pK_as, i.e., equilibrium constants between two microspecies (for each of the 2^n microspecies, each of the n protonation sites can change its state; division by 2 corrects for counting each transition twice). All microspecies with the same number of bound protons form one of the $n+1$ possible macrostates ($0, 1, \ldots, n$ protons bound). Fig. (**1**) presents cetirizine as an example. For $n > 3$, micro- pK_as cannot be derived from titration curves without additional information or assumptions, such as from symmetry considerations [28, 29].

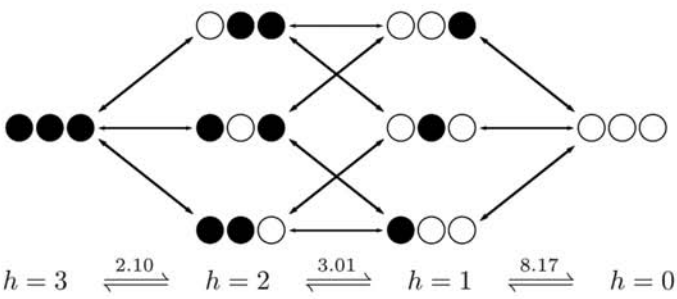
1.2.4. Remarks

Compounds are called *amphiprotic* if they can act as both acid and base, e.g., water, or are multiprotic compounds with both acidic and basic groups. Neutral compounds with formal unit charges of opposite sign are called *zwitterions*;

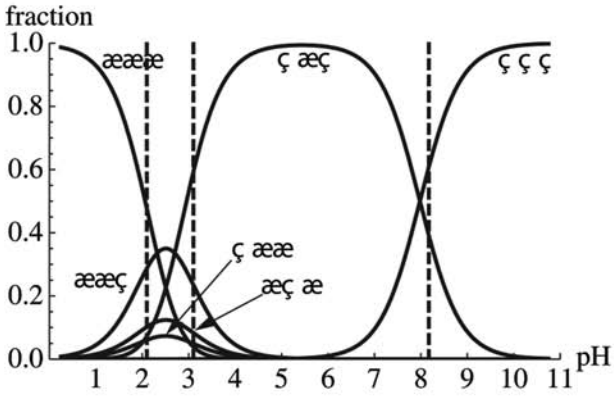
[1](Partial) molar Gibbs energy is also called (partial) molar free enthalpy [23].



(a) Cetirizine. Protonation sites (OH, N, N) in bold face.



(b) Protonation scheme. Cetirizine has $n = 3$ protonation sites, and thus $2^3 = 8$ microstates and $3 \cdot 2^2 = 12$ microequilibria. The $3 + 1 = 4$ macrostates are shown below, with h = number of bound hydrogens.



(c) Distribution of microspecies as a function of pH. The microspecies ● ○ ○ and ○ ○ ● are very close to the baseline.

Microequilibrium	pK_a
● ● ● ⇌ ○ ● ●	2.59
● ● ● ⇌ ● ○ ●	4.03
● ● ● ⇌ ● ● ○	2.28
○ ● ● ⇌ ○ ○ ●	4.58
○ ● ● ⇌ ○ ● ○	2.52
● ○ ● ⇌ ○ ○ ●	3.14

Microequilibrium	pK_a
● ○ ● ⇌ ● ○ ○	5.87
● ● ○ ⇌ ○ ● ○	2.83
● ● ○ ⇌ ● ○ ○	7.62
○ ○ ● ⇌ ○ ○ ○	6.11
○ ● ○ ⇌ ○ ○ ○	8.17
● ○ ○ ⇌ ○ ○ ○	3.38

(d) Microconstants. All values are experimentally determined [30].

Fig. (1). Microspecies and -constants using the example of cetirizine. Microspecies are represented as triplets, where the first position refers to the oxygen of the carboxylic acid group, the second one refers to the middle nitrogen, and the third position refers to the nitrogen farthest from the carboxylic group; e.g., ○ ● ○ represents the zwitterionic form with one proton bound to the middle nitrogen, the dominant neutral form of cetirizine.

the dominant neutral form of cetirizine (Fig. **1**) is an example.

1.3. Factors Influencing pK_a

1.3.1. Environmental Influence

The environment of a compound, in particular temperature, solvent and ionic strength of the surrounding medium, influences its protonation state. For predictive purposes, these are normally assumed constant. Experimental measurements are often done at around 25℃ (whereas a temperature around 37℃ would be physiologically more relevant for drug development) in aqueous solution.

1.3.2. Solvation Effects

Dissociation in aqueous solution is a complex process. Intermolecular solute-solvent interactions have been conventionally divided into two types [31]. The first type is associated to non-specific effects, which are related to the bulk of the solvent, e.g., solvent dielectric polarization in the field of the solute molecule, isotropic dispersion interactions, and solute cavity formation. The second type is associated to specific effects like hydrogen bonding, and other anisotropic solute-solvent interactions.

Note that when modeling a chemical series, e.g., aromatic anilines, a common (aromatic) scaffold can cause similar solute-solvent effects across the series, effectively rendering these effects constant. In such a case, it is not necessary to model them explicitly.

1.3.3. Thermodynamics

Thermodynamic cycles (Fig. **2**) can be used to predict pK_a values [32, 34, 35]. Let

$$\Delta G = -\mu(\mathrm{HA}) - \mu(\mathrm{H_2O}) + \mu(\mathrm{A^-}) + \mu(\mathrm{H_3O^+})$$

and $\Delta G^\Theta = -\mu^\Theta(\mathrm{HA}) - \mu^\Theta(\mathrm{H_2O}) + \mu^\Theta(\mathrm{A^-}) + \mu^\Theta(\mathrm{H_3O^+})$

denote the free reaction enthalpy and the molar free standard reaction enthalpy [36]. From Equation 2,

$\mu(x) = \mu^\Theta(x) + RT \ln a(x)$. Together,

$$\Delta G = \Delta G^\Theta + RT \ln \frac{a(\mathrm{A^-})a(\mathrm{H_3O^+})}{a(\mathrm{HA})a(\mathrm{H_2O})}. \quad (5)$$

At equilibrium, $\Delta G = 0$ and the last term equals K_a, yielding

$$-\Delta G^{\ominus} = RT \ln K_a \Leftrightarrow pK_a = \frac{\Delta G^{\ominus}}{RT \ln 10} \approx \frac{\Delta G^{\ominus}}{2.303RT}. \quad (6)$$

At $T = 298.15\text{K}$, we get $\text{p}K_a \approx \Delta G^{\ominus} / (5708.02\text{Jmol}^{-1})$. A difference of 5.71kJmol^{-1} in $\Delta G^{\ominus}$ thus corresponds to a unit difference in pK_a value. To calculate $\Delta G^{\ominus}$, the quantities $\Delta G^{\circ}_{\text{solv}}(\text{HA})$, $\Delta G^{\circ}_{\text{solv}}(\text{H}_2\text{O})$, ΔG°_{g}, $\Delta G^{\circ}_{\text{solv}}(\text{A}^-)$, and $\Delta G^{\circ}_{\text{solv}}(\text{H}_3\text{O}^+)$ have to be determined. Of these, $\Delta G^{\circ}_{\text{solv}}(\text{H}_2\text{O})$ and $\Delta G^{\circ}_{\text{solv}}(\text{H}_3\text{O}^+) = -110.2\text{kcalmol}^{-1}$ [37] do not depend on HA and can be experimentally determined. The remaining terms may be calculated, e.g., using *ab initio* methods. Approaches differ mainly in the used solvation model. Major categories include explicit solvent models, where individual solvent molecules are simulated [38-41], and, implicit solvent models [7, 42, 43], where the solvent effect on the solute is calculated using, e. g., the Poisson-Boltzmann equation, the generalized Born equation [44, 45], or, integral equation theory [46-49]. Reported accuracies are on the order of 2.5-3.5 kcalmol^{-1} [50-52], which by Equation 6 corresponds to a difference of 1.83-2.57 pK_a units.

1.3.4. Electronic Effects

These can be divided into electrostatic ("through space", Coulomb's law), inductive ("through bonds"), and mesomeric (resonance) effects. To remove a proton from a compound (acids) or the solvent (bases) requires electrical work to be done, the amount of which is influenced by dipoles and charges. Electrostatic interactions between a charged ionizable center and nearby charges can stabilize or destabilize the protonation of the center, depending on whether the prevailing charges are attractive or repulsive. Inductive effects fall off rapidly with distance in saturated hydrocarbons, but less so in unsaturated ones [53]. Mesomeric (or resonance) effects stem from delocalized electron systems, e.g., conjugated systems such as aromatic and heteroaromatic systems with ortho and para substituents [53]. From Equation 6, a unit change in pK_a value corresponds (at $T = 298.15K$) to a change in free energy of 5.7 kJ/mol. Free energy differences of several kJ/mol can occur from charge delocalization [53].

1.3.5. Steric Effects

Compound stereochemistry can influence the distance between ionizable centers of multiprotic compounds. In the case of dicarboxylic acids like butenedioic acid (Fig. **3**), the closer positioning of the two ionizable centers may cause overlapping of the hydration shells, electrostatic repulsion, or internal hydrogen bonding [53]. Steric hindrance and steric shielding may also influence pK_a values.

1.3.6. Internal Hydrogen Bonding

Fig. (**4**) presents an example where the change in pK_a induced by the same substituent differs by one log -unit for two parent structures due to the formation of an internal hydrogen bond in one case, but not in the other.

1.3.7. Tautomeric Effects

The difference in pK_a between two tautomers determines the observed tautomeric ratio between the two species. If the microconstants are known, they can be used to approximate the tautomeric ratio (Fig. **5**) as [2, 54]

$$K_T = \frac{c(\text{T2})}{c(\text{T1})} \approx \frac{K_{a1}}{K_{a2}} \quad \Leftrightarrow \quad pK_T \approx pK_{a2} - pK_{a1}. \quad (7)$$

1.4. Importance

1.4.1. Drug Development

The ionization state of a compound across the physiological pH range affects, among others, physicochemical parameters such as lipophilicity, and, solubility, but also the compounds ability to diffuse across membranes, to pass the blood-brain barrier, and to bind to proteins. These properties in turn influence the absorption, distribution, metabolism, excretion, and, toxicity (ADMET) characteristics of the compound. As an example, pK_a strongly influences the octanol/water distribution coefficient $\log D$ (which measures the distribution of neutral and charged species). It can be directly estimated from the octanol/water partition coefficient $\log P$ (which measures the distribution of the neutral species alone) as [55]

$$\log D \approx \log P - \log(1 + 10^{(\text{pH} - \text{p}K_a)\delta}), \quad (8)$$

where δ is $+1$ for acids and -1 for bases, assuming that only the neutral form partitions into the organic phase. For multiprotic compounds, the equation should be modified to incorporate correction terms for all ionizable groups. For the

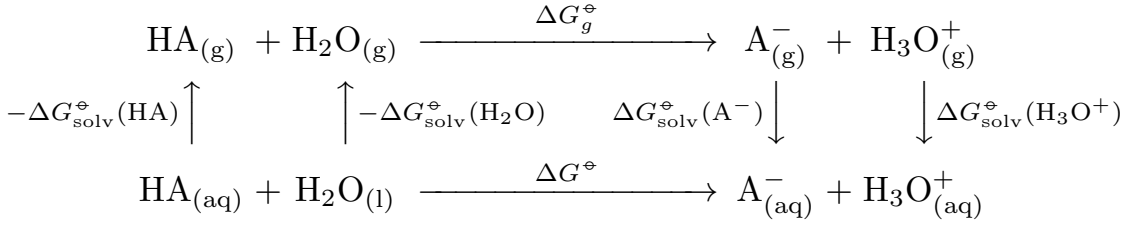


Fig. (2). A thermodynamic cycle [32] (sometimes called Born-Haber cycle [33]) used in pK_a prediction. The cycle describes the change in Gibbs energy upon the dissociation of the acid HA in water. The change in Gibbs energy must be the same for both paths. (g) = gas phase, (aq) = aqueous solution, (l) = liquid phase, solv = solvation.

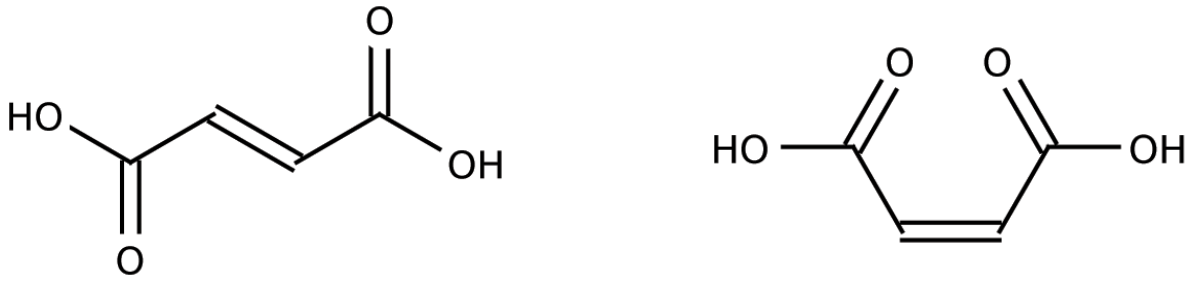


Fig. (3). Example of the influence of steric effects on pK_a. cis/trans-isomerism in butenedioic acid causes marked changes in pK_a values.

(a) Phenylamine (left, pK_a=4.84) and 2-aminophenol (right, pK_a=4.60), a difference of Δp$K_a = 0.24$.

(b) Benzoic acid (left, pK_a=4.20) and 2-hydroxybenzoic acid (right, pK_a=2.98), a difference of Δp$K_a = 1.22$.

Fig. (4). Influence of internal hydrogen bonding on pK_a [2]. The difference in ΔpK_a between (**a**) and (**b**) is due to the different strength of the internal hydrogen bonding.

importance of $\log D$ and $\log P$ in drug discovery, see the literature [56, 57]. pK_a has been considered one of the five most important physico-chemical profiling screens for early ADMET characterization [58]. The protonation state of a compound in aqueous solution is thus directly relevant to many aspects of drug development (Table **2**). When considering these aspects, it is important to take the pH of a particular environment into account, since it determines microspecies composition.

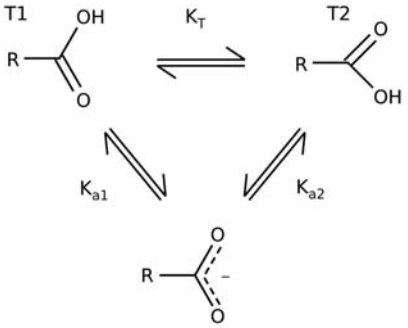


Fig. (5). Approximation of tautomeric ratio by microconstants. Shown are the enol (top left), keto (top right), and anionic (bottom) form of a carboxylic acid.

1.4.2. The Ionizability of Drugs

Most drugs are weak acids and/or bases (Table **3**). The percentage of drugs with at least one group that is ionizable in the physiological pH range from 2 to 12 has been estimated at 63% [70] and 95% [71]. pK_a-values are therefore relevant for (the pharmacodynamic and -kinetic characteristics of) the majority of drugs.

1.4.3. Passive Membrane Diffusion

The ability of a compound to passively diffuse across a biomembrane (lipid layer) depends on its *partition ratio* [73] (also distribution constant, partition coefficient), i.e., the ratio of its concentration $c_{\mathrm{li}}(\cdot)$ in a lipid phase and its concentration $c_{\mathrm{aq}}(\cdot)$ in an aqueous phase at equilibrium, $\mathrm{K_D}(\cdot) = c_{\mathrm{li}}(\cdot)/c_{\mathrm{aq}}(\cdot)$. As a rule of thumb, neutral compounds are more easily absorbed by membranes than ionized species. When one neglects the permeation of ions into the lipid phase, the apparent partition ratio is given by [74]

$$\mathrm{K_D^{app}} = \frac{c_{\mathrm{li}}(\mathrm{HA})}{c_{\mathrm{aq}}(\mathrm{HA}) + c_{\mathrm{aq}}(\mathrm{A^-})}. \quad (9)$$

Combining Equations 1 and 9 with the definition of pH and $\mathrm{K_D}$ and taking logarithms yields

$$\log_{10}\mathrm{K_D^{app}} \approx \log_{10}\mathrm{K_D(HA)} - \mathrm{pH} - \log_{10}(\mathrm{H} + K_a). \quad (10)$$

If $\mathrm{pH}=\mathrm{p}K_a$, then

$$\log_{10}\mathrm{K_D^{app}} = \log_{10}\mathrm{K_D(HA)} - \log_{10}2 \approx \log_{10}\mathrm{K_D(HA)} - 0.301.$$

For $\mathrm{pH} \ll \mathrm{p}K_a$, Equation 10 can be approximated by $\log_{10}\mathrm{K_D^{app}} = \log_{10}\mathrm{K_D(HA)}$, and for $\mathrm{pH} \gg \mathrm{p}K_a$ by $\log_{10}\mathrm{K_D^{app}} = \log_{10}\mathrm{K_D(HA)} - \mathrm{pH} + \mathrm{p}K_a$ [74]. See the literature [74] for equations including the permeation of ions into the lipid phase. By rearranging Equation 10, one can relate the pK_a and pH of a compound to its $K_{\mathrm{D}}^{\mathrm{app}}$ and $\mathrm{K_D(AH)}$ as

$$\log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}} - 1\right) = \mathrm{pH\text{-}p}K_a. \quad (11)$$

Table 2. Relevance of pK_a in Drug Development. BBB = Blood-Brain Barrier

Aspect	Comment
Physico-Chemical	
Lipophilicity	Neutral species are more lipophilic than ionized ones since less energy is required to remove the hydration layer
Solubility	Water is a polar solvent, and pK_a thus directly influences solubility
Fundamental	
pH homeostasis	Organisms maintain a constant pH in blood by using biological buffers. Disturbances in human acid-base balance are directly relevant in medicine [59]
Function	Many biochemical reactions depend on, or directly involve, protonation state, e.g., reactions catalyzed by an enzyme are often initiated by proton transfer or hydrogen bonding [60]. Heterolytic cleavage of C-H bonds starts many enzyme-catalyzed processes [61-65]
Pharmaceutical	
Absorption	Lipophilic species are absorbed better, e.g., intestinal uptake
BBB permeation	It has been suggested that protonation state influences BBB permeability [66]
Formulation	Choice of excipient and counter-ion
Metabolism	pK_a can influence rate and site of metabolization [2, 67]
Signaling	Many neurotransmitters are ionizable amine compounds [68]
Pharmacodynamics	pH in the human body varies between 2 and 12, with the microspecies population of a compound, and thus its behavior, varying accordingly [69]

Table 3. Percentage of Acids and Bases in the Data Set by Williams (Subset of *n*=582) and the World Drug Index (Version of 1999, *n*=51596; Thomson Reuters, www.thomsonreuters.com), as given by Manallack [4]

Data Set	1 Acid	1 Base	2 Acids	2 Bases	1 Acid & 1 Base	Others
Williams	24.4%	45.4%	3.8%	10.5%	11.2%	4.8%
World drug index	11.6%	42.9%	3.0%	24.6%	7.5%	10.4%

1.4.4. Role in Drug Development

The development of high-throughput methods of experimental pK_a determination [6] is in itself an indicator of the importance of pK_a values in drug development. pK_a is often used as a preliminary measure to select prospective compounds [75] due to its close relation with many ADMET properties (Table **2**). Since drug failures get more costly the later they occur during drug development, accurate estimation of pK_a-values can help to reduce costs and development time by acting as an early indicator of ADMET-related problems. The pK_a of a compound is also relevant in the design of combinatorial libraries or the purchase of third party library subsets. Computational methods are a valuable addition to experimental methods. They have the advantage that they can be applied to virtual molecules, e.g., in *de novo* design, or when virtually screening large libraries. Compared to experimental methods, they are fast and cost-effective. However, one should bear in mind that the accuracy of predictions is rather limited, and that the result is only an estimate-for the actual value, experimental determination is required.

1.4.5. Other Areas

The degree of ionization influences toxicity and fate of weak organic acids in natural waters [76]. Specific modes of toxic action, e.g., the uncoupling of the oxidative phosphorylation, depend directly on lipophilicity and acidity [77-79].

Protonation and deprotonation processes of compounds in organic solvents are relevant to many chemical reactions, syntheses, and analytical procedures, e.g., acid-base titrations, solvent extraction, complex formation, and ion transport [80]. In this work, we restrict ourselves to the prediction of pK_a in aqueous solution; for organic solvents, we refer to the literature [80-82].

2. DATA

2.1. Sources and Availability

A considerable number of experimentally determined pK_a values have been published in the primary literature. Most are available either in electronic collections or in book form (Table **4**). The two biggest problems with these sources are availability (most databases are commercial) and data quality.

2.2. Data Quality

The reliability and accuracy of publicly available experimentally determined pK_a values is often dubious [3]. Apart from the problems associated with the actual experimental determination, a number of errors occur in data sets:

Table 4. pK_a Data Sets. HSDB = Hazardous Substances Data Bank, NIST = National Institute of Standards and Technology (www.nist.gov)

(a) Databases containing experimental pK_a values. Some databases are electronic versions of books. The number of measurements varies widely, from a few hundred up to ca. $1.5 \cdot 10^5$ (Beilstein). The pKaData data sets contain pK_a measurements that were sponsored by the International Union of Pure and Applied Chemistry (IUPAC) and published in book form [83-86].

Name	Vendor	Values
ACD/pKa DB	Advanced Chemistry Development Inc., Toronto, Canada. www.acdlabs.com	>31 000
ADME index	Lighthouse Data Solutions LLC. www.bio-rad.com	
Beilstein/Gmelin	Elsevier Information Systems GmbH, Frankfurt, Germany. www.elsevier.com	148 880
BioLoom	BioByte Corp., Claremont, California, USA. www.biobyte.com	14 000
ChEMBL	European Bioinformatics Institute, Cambridge, UK. www.ebi.ac.uk/chembldb/	4 650
CRC handbook	Taylor and Francis Group LLC, New York, New York, USA. www.hbcpnetbase.com	
HSDB	National Institutes of Health, www.toxnet.nlm.nih.gov	959
Lange's handbook	Knovel Corp., New York, New York, USA. www.knovel.com	
LOGKOW	Sangster Research Laboratories, Montréal, Québec, Canada. www.logkow.cisti.nrc.ca	
Merck index	Cambridgesoft Corp., Cambridge, Massachusetts, USA. www.cambridgesoft.com	
MolSuite DB	ChemSW, FairField, California, USA. www.chemsw.com	
NIST std. ref. DB 46	National Institute of Standards and Technology, USA. www.nist.gov	
OCHEM	Helmholtz Research Center for Environmental Health, Munich, Germany. www.ochem.eu, www.qspr.eu	>5 000
Pallas pKalc	CompuDrug Ltd., Sedona, Arizona, USA. www.compudrug.com	
PhysProp	Syracuse Research Corp., North Syracuse, USA. www.syrres.com	
pK database	University of Tartu, Estonia. www.mega.chem.ut.ee/tktool/teadus/pkdb/	>20 000
pKaData	pKaData Ltd. www.pkadata.com	
SPARC	University of Georgia, USA. www.ibmlFc2.chem.uga.edu/sparc/	

(b) Books containing experimental pK_a values of compounds in aqueous solution, sorted by year and author name.

Ref.	Author (Year)	Comment	Values
[84]	Kortüm *et al.* (1961)	Organic acids	2 893
[87]	Albert (1963)	Heterocyclic substances	
[88]	Sillén and Martell (1964)	Metal-ion complexes	
[89]	Perrin (1965)	Organic bases	
[90]	Izatt and Christensen (1968)	Book chapter [91]	
[92]	Jencks and Regenstein (1968)	Book chapter [91]	
[93]	Perrin (1969)	Inorganic acids and bases	8 766
[94]	Sillén and Martell (1971)	Metal-ion complexes	
[83]	Perrin (1972)	Weak bases	~4 300
[95]	Martell and Smith (1974)	NIST std. ref. database 46	6 166
[96]	Perrin (1976)	Organic bases	
[85]	Serjeant and Dempsey (1979)	Organic acids	~4 520
[53]	Perrin *et al.* (1981)	Hammett-Taft equations	
[97]	Perrin (1982)	Inorganic acids and bases	
[98]	Albert and Serjeant (1984)	Laboratory manual	
[99]	Drayton (1990)	Pharmaceutical substances	
[100]	Avdeef (2003)	"Gold Standard" data set	
[101]	Speight (2004)	Lange's handbook	
[102]	Lide (2006)	CRC handbook	
[103]	O'Neill (2006)	Merck Index	796
[104]	Prankerd (2007)	Pharmaceutical substances	
[72]	Williams (2008)	Williams data set	

- wrong associations of value with structure, e.g., due to ambiguous or non-standard compound names, or typographical errors in compound names or other identifiers.
- wrong numerical values, e.g., typographical errors in pK_a value, K_a instead of pK_a, $-\log_{10}(\mathrm{p}K_a)$, or, p$K_b$ value instead of pK_a value.
- wrong associations of values with multiple ionizable centers of the same compound.
- duplicate entries; even if the pK_a values are identical, duplicates can upweight the importance of compounds in the training set of statistical methods, or compromise retrospective validation by occurring in training and validation set.
- predicted instead of experimental values.
- wrong specification of experimental conditions, e.g., temperature or solvent.
- wrong or inaccurate published values; e.g., experimental values for dichlorphenamide have been stated both as $\mathrm{p}K_{a1} = 8.24$, $\mathrm{p}K_{a2} = 9.50$ [105], and as $\mathrm{p}K_{a1} = 7.4$, $\mathrm{p}K_{a2} = 8.6$ [106].

The error in experimental determination of pK_a values has been stated as being on the order of 0.5 pK_a units [107], although lower errors have been reported as well [105, 108]. Another factor that influences pK_a prediction is that compounds are often clustered around over-represented compound classes, e. g., phenols, or, carboxylic acids.

Preprocessing, e.g., by filtering according to experimental conditions, statistical comparison of values from different sources, investigation of pK_a differences within series of analogues, investigation of model outliers, manual inspection, and verification of the original references, can, to a limited extent, aid in data curation.

3. PREDICTION

"pK_a does not lend itself to simple calculation" [4].

A wide variety of approaches have been used to establish quantitative structure-property relationships for the pK_a of small molecules in aqueous solution. Table **5** presents a non-comprehensive list of publications on the topic. Different categorizations are possible, e.g., by basic method type (first principles versus empirical), by the dimensionality of the used molecular representation (1D, 2D, 3D), by the used molecular representation, by the investigated compound classes, etc. We decided to separate the publications into those using first principles-based calculations and those using empirical/statistical approaches.

It is not clear how to judge absolute errors in pK_a predictions. Most authors seem to agree that deviations by no more than 1 log -unit are acceptable [4]. Liao & Nicklaus [109] classify predictions based on the absolute deviation a as excellent ($a \leq 0.1$), well ($0.1 < a \leq 0.5$), poor ($1.0 < a \leq 2$), or awful ($2 < a$) (with $0.5 < a \leq 1$ unspecified, we suggest "fair" for this range). We have deliberately refrained from listing performance statistics in Table **5** because these can not be meaningfully compared. There are several reasons for this:

- different performance statistics (R^2, RMSE, MAE, F, SEE, …),
- different retrospective evaluation methods, e.g., different types of cross-validation,
- different data sets (compare Table **7**),
- different pK_a ranges: An error of 0.5 means something else if the data set pK_a values span 12 orders of magnitude rather than two.

These problems could be solved by agreeing on a common set of performance statistics, evaluation methods, and standard benchmark data sets, but such a standard procedure is not in sight.

3.1. Challenges

Challenges specific to the prediction of pK_a values include:

- conformational flexibility. Due to steric effects (Fig. **3**), the conformation of a compound can strongly influence its pK_a internal hydrogen bonding. The formation of internal hydrogen bonds, as well as their strength, influence pK_a (Fig. **4**); an example of this can be found in the work of Tehan *et al.* [155, 156], where separate modeling of phenols that form internal hydrogen bonds and those that do not improved model accuracy.
- multiprotic compounds. The presence of more than one ionizable center complicates modeling due to the necessity to consider microstates.

An important challenge not specific to pK_a is the number of available examples to train the model. Building individual models for each chemical series, as in LFERs, aggravates this problem further. While some types of compounds like phenols, or carboxylic acids, have been extensively investigated, and many pK_a values are available, for other types there is little or no data. Often, the compounds for which predictions are most interesting are new (e.g., not covered by patents), and thus often outside the domain of applicability of empirical models, requiring initial experimental determinations.

3.2. Methods

Different methodological approaches, ranging from simple regression analysis to neural networks and kernel methods, were used to predict pK_a values of small molecules. Since a review of all used methods is not feasible, we limit ourselves to selected major methodological categories and studies on pK_a prediction that were used to predict more than 500 molecules.

Table 5. Published pK_a Models

Ref.	*n*	S	MP	Method	Remarks
ab initio					
[129]	5	yes	no	*ab initio* (MP3/MP2/6-31+G(d)//6-31+G(d)	organic acids
[130]	16	yes	no	SCF (6-31G**//6-31+G**//6-311G (2d,2p)//6-311+G(2d,2p)); PCM-UAHF	aliphatic carboxylic acids
[131]	12	yes	no	*ab initio* (HF/6-31+G**; PCM)	carboxylic acids
[132]	15	yes	no	*ab initio* (HF/6-31+G**; PCM)	aliphatic alcohols, thiols, halogenated carboxylic acids
[133]	8	no	yes	*ab initio* (MP2, G2MP2, DFT B3LYP/6-311++G(d,p); PCM)	pK_a up to 50
[134]	36	yes	no	MEP-V_{min}, MEP-$V_{S,min}$, $I_{S,min}$, Hammet (sigma); *ab initio* (HF/6-311G(d,p)	anilines
[135]	6	yes	no	*ab initio* (CBS-QB3, CBS-APNO; CPCM)	carboxylic acids
[136]	20	yes	no	*ab initio* (HF/CPCM with 8 different solvation models)	phenols
[137]	17	no	no	*ab initio* (MP2/6-311+G(2df,2p))	
[138]	26	yes	no	*ab initio* (DFT B3LYP/6-31G** & cc-pvqz, Becke(1/2); two-step)	carboxylic acids, phenols, imides, heterocycles
[139]	12	yes	no	*ab initio* (CBS-QB3, MP2/6-311+G(d,p), HF/6-31+G(d,p); CPCM)	pK_a up to 34
[140]	36	yes	no	MEP-$V_{S,min}$, MEP-$V_{S,max}$, $I_{S,min}$, $I_{S,max}$ and $V_{S,max}$; *ab initio* (HF/STO-5G(d)//B3LYP/6-311G8d,p))	phenols and benzoic acids
[141]	13	yes	no	*ab initio* (B3LYP/6-31+G(d,p)-PCM(opt)	13 different Methods, Basis Sets, Solvent Models
[142]	66	yes	no	*ab initio* (DFT B3LYP/6-31+G**; PCM)	carboxylic acids
[143]	63	yes	no	MEP-$V_{S,min}$, MEP-$V_{S,max}$, $I_{S,min}$; *ab initio* (HF/6-31G*)	
[68]	24	yes	yes	*ab initio* (B3LYP/6-31+G*, MP2/6-311++G**)	alcohols, amines, anilines, carboxylic acids, imines, pyridines, pyrimidines
[144]	12	yes	yes	*ab initio* (CBS-QB3, HF/6-31G(d); CPCM)	hydroxamic acids
[145]	4	yes	no	PB continuum solvation; *ab initio* (B3LYP/6-311++G(d,p)	different tautomers are considered
[146]	228	yes	no	MEP, DFT B3LYP/6-311+G(d,p); PCM	amines, anilines, carboxylic acids, alcohols, sulfonic acids, thiols
[1]	55	no	no	direct, proton exchange, hybrid cluster-continuum, implicit-explicit; CPCM-UAKS/UAHF, IPCM, SM6, COSMO-RS	neutral organic and inorganic acids
[75]	34	yes	no	*ab initio* (OLYP/3-21G(d)//6-311+G**; COSMO)	anilines, amines, pyridines, phenols, alcohols, carboxylic acids
[147]	370	yes	yes	*ab initio* (OLYP/3-21G(d)//6-311+G**; COSMO)	carboxylic acids, phosphonic acids, phenols, alcohols, thiols, hydroxamic acids, oximes
[148]	85	yes	yes	*ab initio* (DFT B3LYP/6-31+G(d,p); PCM)	carboxylic acids, direct approach
Empirical					
[149]	135	yes	no	LFER	monoprotic oxy acids
[150]	3995	yes	no	LFER; fragmental and QM descriptors[2]; MCASE	first pK_a value; carboxylic acids, phenols, others
[151]	3685	yes	yes	LFER; SAR, PMO; SPARC	
[152]	123	no	yes	LFER; SAR, PMO; SPARC	hydration and tautomerization is considered

[2]Including fragments, partition coefficient, water solubility, molecular weight, Hückel molecular orbital charge densities, HOMO, LUMO, absolute electronegativity, and hardness.

(Table 5) contd…..

Ref.	*n*	S	MP	Method	Remarks
[153]	28	yes	no	regression; RM1, B3LYP/6-31G*, SM5.4/A, charges, energy differences	aliphatic amines
[154]	19	yes	no	regression; HF 6-311G**, natural charges	anilines
[155]	417	yes	no	regression; semi-empirical (frontier electron theory)	phenols, carboxylic acids; electrophilic superdelocalisability
[156]	282	yes	no	regression; semi-empirical (frontier electron theory)	amines, anilines, heterocycles; electrophilic superdelocalisability
[157]	271	yes	no	regression; empirical coefficient	
[158]	28	yes	no	regression; semi-empirical RM, DFT B3LYP/6-31G*	aliphatic amines
[159]	36	yes	no	regression; semi-empirical (PM3 electron density at phosphorus atom)	alkylphosphonic acids
[160]	38	yes	no	regression; B3LYP/6-311G(d,p); atomic charges	anilines, phenols; 7 different charge schemes
[161]	150	no	no	regression; MERA atomic charges	full equalization of orbital electronegativities
[162]	32	yes	no	regression (multivariate)	carbon acids
[163]	74	yes	no	MLR	aromatic acids
[164]	190	yes	no	MLR; semi-empirical QM-descriptors	phenols, aromatic & aliphatic carboxylic acids
[165]	215	yes	no	MLR; semi-empirical descriptors (MO)	phenols, alcohols, carboxylic and benzoic acids
[166]	1410	yes	yes	MLR; topol., steric, inductive descriptors	1122 carboxylic acids and 288 alcohols
[167]	15	yes	no	MLR; quantum chemical descriptors	atomic charge on N3 proton and frontier orbital energy
[168]	676	yes	no	MLR / ANN; quantum-chemical and fragmental descriptors	phenols, carboxylic acids, nitrogen-containing compounds
[169]	74	no	no	PLS, RBFNN; constitutional, topol., geom., electrostatic, QM-descriptors	
[170]	26013	yes	yes	PLS; MIFs (3D Molecular Interaction Fields)	
[69]	3169	yes	yes	PLS; semi-empirical and information-based descriptors	alcohols, amines, anilines, carboxylic acids, imines, pyridines, pyrimidines
[171]	95	yes	no	PLS; quantum topological molecular similarity QTMS	carboxylic acids, anilines, phenols
[172]	1037	yes	yes	PLS; tree-structured fingerprints	625 acids and 412 bases
[173]	228	yes	no	PLS, SVM, RBFNN; QTMS	carboxylic acids
[174]	62	yes	no	PLS; QTMS	phenols
[175]	49	yes	no	PLS (CoMSA); MEP (molecular electrostatic potential)	PLS with iterative variable elimination (IVE-PLS)
[176]	49	yes	no	PLS (CoMFA)	benzoic acids
[177]	39	yes	no	PLS (CoMFA)	imidazoles and imidazolines
[178]	25	yes	no	PLS / PCA (CoMFA)	nucleic acid components (nitrogenated bases, nucleosides, nucleotides)
[179]	242	no	no	ANN; semi-empirical, theoretical descriptors	benzoic acids and phenols
[180]	94	yes	no	ANN; AM1/CODESSA[3]	phenols; water and 9 organic solvents
[80]	136	yes	no	ANN; AM1/CODESSA[4]	benzoic acids; water and 8 organic solvents

[3]ALFA polarizability (DIP), max. partial charge for H, LUMO+1 energy, max. e-n attraction for C-O, relative positive charged surface area.

[4]Max. partial charge for H, fractional partial positive surface area, max. e-e repulsion for O, min. resonance energy for O-H, min. valency of C.

3.2.1. Ab Initio Calculations

Traditionally, thermodynamic cycles (Fig. **2**) are used for *ab initio* pK_a predictions because deprotonation energy is easier to calculate in the gas phase. Such approaches differ

(Table 5) contd…..

Ref.	*n*	S	MP	Method	Remarks
[181]	282	yes	no	PC-MLR, PC-ANN and GA	nitrogen containing compounds; anilines, amines, pyridines, pyrimidines, imidazoles, benzimidazoles, quinolines
[182]	107	no	no	SVM, LS-SVM, CART; AM1/Dragon	pH indicators
[183]	28	no	no	SVM, PCR, PLS, MLR; AM1/Dragon, B3LYP/6-31+G**//Gaussian98	
[184]	64	no	no	COSMO-RS	organic and inorganic acids
[185]	43	no	no	COSMO-RS	bases (amidines, anilines, benzodiazepines, guanidines, heterocyclics, pyrroles, indoles)
[186]	1881	no	no	decision tree; SMARTS pattern	substructure-based
[187]	31	yes	no	anti-connectivity topological index	
[188]	7	yes	yes	MD continuum solvation	diprotic acids and bases
[189]	4700	no	yes	structural fingerprints, database lookup	

n = number of structures (the number of pK_a values can be higher if multiprotic compounds were included), S = compounds organized into series or restricted to one series, MP = multiprocity (whether microconstants were treated), B3LYP = hybrid-exchange correlation functional of Becke, Lee, Yang, Parr [110, 111], CART = classification and regression trees [112], CBS = complete basis set, CODESSA = comprehensive descriptors for structural and statistical analysis [113], COSMO = conductor-like screening model [114], CPCM = conductor-like polarizable continuum model, DFT = density functional theory, Dragon = descriptors by Dragon [115, 116], Gaussian98 = descriptors by Gaussian98 [117], HF = Hartree-Fock, HOMO = highest occupied molecular orbital, LFER = linear free energy relationship, LUMO = lowest unoccupied molecular orbital, MEP = molecular electrostatic potential, MP2 = second order Møller-Plesset perturbation theory, OLYP = OPTX + LYP exchange functional [118], PCM = polarizable continuum model, PLS = partial least squares, PMO = perturbed molecular orbital theory [119], QTMS = quantum topological molecular similarity [120, 121], RBFNN = radial basis function neural network [122, 123], RM1 = Recife model 1 [124], SM5.4/A = solvation model 5.4 using AM1 [125], SMARTS = SMILES arbitrary target specification, SMILES = simplified molecular input line entry specification [126], SPARC = SPARC performs automated reasoning in chemistry, SVM = support vector machine [127, 128].

mainly in the solvation model employed. First principles calculations of pK_a values in the gas phase require computationally demanding levels of theory, i.e., large basis sets and a high level of electron correlation [139], but can achieve accuracy comparable to experimental determination. It has recently been argued [75] that with the level of theory computationally feasible today, the detour via the gas phase is counter-productive, as the gain from improved accuracy in the gas phase is outweighed by errors due to conformational differences between gas and aqueous phase. Others [1] advocate proton exchange schemes based on the cluster continuum model over direct methods because the latter are mainly limited to structures similar to those used in the original parameterization of the chosen solvation model. Optimization of the structure is necessary for accurate estimation [75]. Conformational flexibility is a problem, as it is not always possible to identify the global energy minimum; in such cases, multiple low energy conformations should be used as starting points [147]. Although efforts have been made to increase the scale of quantum chemical pK_a estimations, present applications are for computational reasons still limited to smaller data sets containing structurally closely related compounds. Another factor that hinders more widespread use of quantum mechanical methods is the expertise that is needed to set up, conduct, and evaluate the results of these methods.

3.2.2. Statistical and Machine Learning Methods

In QSPR modeling of pK_a, structural or experimentally determined properties of compounds are statistically related to their pK_a values. Structural properties can be symbolic representations of a molecule, such as strings (e.g., SMILES [126] notation), graphs (e.g., structure graph, reduced graph), or densities (e.g., electron density). Most of the time, they are calculated values, called chemical *descriptors*, "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" [116]. Descriptors encode specific properties of molecules that are related to the property under investigation, here pK_a values. Owing to the variety of chemical phenomena and structures, a large number of molecular descriptors have been developed: The handbook of molecular descriptors [116] lists more than 1600 of them. These descriptors are used to train statistical or machine learning methods to predict or to model the pK_a values of new substances. The predictive power of such a model depends on its ability to detect linear or non-linear relationships between the chemical descriptors and the property pK_a. Many combinations of descriptors and methods have been published so far (Table **5**).

Linear free energy relationships. In a linear free energy relationship (LFER), "a linear correlation between the logarithm of a rate constant or equilibrium constant for one series of reactions and the logarithm of the rate constant or equilibrium constant for a related series of reactions" [190] is established, e.g., for pK_a prediction [149-151]. pK_a values are linearly related to changes in Gibbs free energy (molar free standard reaction enthalpy; Equation 6). If these changes are not too big, the contributions of substituents are approximately additive, leading to the Hammett-Taft equation [11]

$$\log_{10}\frac{K_a}{K_a^0} = \rho\sum_{i=1}^{m}\sigma_i \Leftrightarrow \mathrm{p}K_a = \mathrm{p}K_a^0 - \rho\sum_{i=1}^{m}\sigma_i, \qquad (12)$$

where $\mathrm{p}K_a^0$ is the dissociation constant of the parent (unsubstituted) molecule, ρ is a constant specific for the modeled class of molecules, m is the number of substituent positions, and the σ_i are constants expressing the substituent effect on the dissociation constant. A disadvantage of this approach is that the σ constants have to be known (experimentally determined) for all involved substituents

[191]. For details on pK_a LFERs, see the book by Perrin *et al.* [53]. The determination of Hammett-Taft σ constants is still content of current examinations [82]. LFERs were predominantly used in the early days of pK_a prediction [53, 192, 193], but remain useful in successful prediction software tools (Section 3.4) and research [151].

Regression. Many variants of regression exist, e.g., simple linear and multi-linear regression, ridge regression [194], principle components regression, or symbolic regression [195]. Ordinary regression is a good method for exploring simple relationships between structural descriptors and pK_a. A variant that is popular in QSPR is partial least squares (PLS) [196], which is similar to ordinary regression on principal components, but includes the experimental measurements in the calculation of the components, i.e., it considers not only the variance in the input descriptors, but also their correlation with the pK_a values.

In general, regression methods are easy to interpret, since there is a direct correlation between the descriptors and the property itself. Many QSPR approaches therefore use linear regression, multi-linear regression (MLR), or partial least squares (PLS) [69, 155, 156, 166, 168, 170, 172].

Artificial neural networks. An artificial neural network (ANN) consists of units (neurons) organized into layers and connected via coefficients (weights). Every ANN consists of at least three layers: an input layer, an output layer, and at least one hidden layer between them. ANNs are adaptive systems modeled after biological neural networks. They are used to model non-linear relationships between inputs and outputs. For the training of ANNs, a variety of different computational methods exist, e.g., back-propagation (BPNN) [197], principal component analysis (PCA-ANN) [198], genetic algorithms (GA-BPNN) [199], or radial basis functions (RBFNN) [200]. Due to their success in detecting complex non-linear relationships amongst data, ANNs have become popular [201] in QSPR/QSAR models, including pK_a prediction [179, 180].

Kernel-based machine learning. Kernel methods [202] are systematically derived non-linear versions of linear machine learning algorithms by means of the kernel trick. Prominent algorithms include support vector machines (SVM) [128], kernel principle component analysis [203], and kernel partial least squares [204, 205]. The idea behind the kernel trick is to implicitly calculate similarities between non-linear projections of the input descriptors. An advantage of this approach is the systematic and rigorous treatment of non-linearity (encoded by the used kernel function) that often leads to excellent performance. A disadvantage is that solutions, e.g., weight vectors, refer to training examples, not input dimensions, leading to higher runtimes and reduced interpretability.

Kernel-based learning methods have only recently been used for pK_a prediction [182, 183, 206]. In a recent study, we used kernel ridge regression with a graph kernel [207] designed for the comparison of small molecules to predict the pK_a values of a published set of 698 compounds. The results were similar to those of a previously published semi-empirical approach [155, 156] based on frontier electron theory, but without the need for structure optimization.

3.2.3. Selected Studies

There is a large number of studies on pK_a prediction (Table **5**). We provide a brief overview of studies that were used to predict at least 500 molecules.

Klopman and Fercu [150] used the MULTI-CASE methodology to estimate pK_a values based on 3813 monoprotic acids. This was one of the first studies using such a large and diverse set of compounds. Their data were collected from the book by Kortüm *et al.* [84], as well as from a number of other sources. The MULTI-CASE approach partitions molecules based on subfragments of 2-10 atoms, and uses statistical approaches to identify "biophores", significant fragments with a chance of at most 5% to occur by chance alone. Once a biophore was identified, compounds that contained it were removed, and analysis repeated. For each set of compounds with a common biophore a local QSAR model was constructed based on fragments (modulators) that increase or decrease the activity of molecules due to the biophore. In addition to fragments, other physico-chemical and quantum-chemical molecular parameters like logP, HOMO, LUMO, and absolute electronegativity were used. In this study, all molecules were first classified as active (p$K_a \leq 6.5$), marginal ($6.5 < \text{p}K_a < 7.8$), and inactive (p$K_a \geq 7.8$). Based on this, 22 biophores were identified that were used to predict a test set of 192 organic acids [208] with $R \approx 0.82$ and standard error of 1.58 pK_a units.

The SPARC (SPARC performs automated reasoning in chemistry) approach [151, 209] uses linear free energy relationships and perturbed molecular orbital theory [119] to describe resonance, solvation, electrostatic, and quantum effects. For example, its resonance models were developed using light absorption spectra. Data on physico-chemical properties were used to derive solvation models, and electrostatic models were developed for pK_a data. The system uses parameters derived from different properties and can perform mechanistic modeling resulting in interpretable models. For pK_a prediction, 13 ionizable centers (c) were identified and their pK_a values $(\text{p}K_a)_c$ were tabulated. Any molecular structure p appended to the center was considered a perturber. The pK_a of the center was calculated as $\text{p}K_a = (\text{p}K_a)_c + \delta_p(\text{p}K_a)_c$, where $\delta_p(\text{p}K_a)_c$ is the change in ionization behavior caused by p. The perturbation was subdivided into resonance, electrostatic, solvation and H-bonding of p with the protonated and unprotonated forms of the ionizable center. This allowed SPARC to estimate pK_a microionization constants that in turn could be used to derive macro constants and other related characteristics, e.g., titration curves. This approach is limited in its scope by the number of parameterized substituents and reactive centers, for which characteristics need to be derived from experiments. The method was applied to calculate the pK_a of 3685 compounds, including multiprotic compounds with up to six centers and a range of over 30 units, with a RMSE of 0.37.

While SPARC tries to account for all effects explicitly, including the distance from the ionizable center, several studies explicitly accounted for the distance by using descriptors centered on the ionizable center. These descriptors dissect local

structural information in expanding concentric levels of bond distance from the ionizable site. By specifying the number of levels, one can control for molecular description details depending on the analyzed ionisable center. One of the first studies using this approach was done by Xing *et al.* [172]. The authors counted the number of Sybyl atom types of different distances from the ionizable center in a vector as a representation of the ionizable group. In addition to the 22 atom types, 11 chemical groups (nitro, nitroso, cyano, carbonyl, carboxylate, sulfone, sulfonate, sulfoxide, sulfinate, hydroxyl, and sulfhydryl) that are explicitly involved in π-electron systems were also considered. A maximum of five distance levels was used, resulting in 165 descriptors. Atom and group types not found in the neighborhood of an ionizable group were excluded. Partial least squares was used for regression. Separate models were created for four classes of acids (aromatic acids; aliphatic acids and alcohols; phenols and thiophenols; acidic carbons and acidic nitrogens) and bases (pyridines, anilines, imidazoles, and alkylamines). The approach was validated using 25 acids and bases from Perrin's book that did not participate in model development, resulting in a RMSE of 0.40. For four compounds, no appropriate model could be found due to missing atom types in the respective training sets.

The MoKa program developed by Cruciani *et al.* [170] can be considered an extension of this approach. There, the same idea of layers surrounding the ionizable center is used, combined with the idea to convert atoms to energies calculated with 3D molecular interaction fields. Since calculation of 3D conformations can be computationally demanding, the authors represented each atom in a molecule as a pre-computed fragment for which minimum energies with a pre-selected set of ten probes were calculated. These energies are binned for each layer and summed to calculate a vector representing the ionizable group in each layer. The number of layers was varied from 7 to 13 while the energies were binned using 25 levels. 33 pK_a prediction models were developed to cover different ionizable groups. While it would have been possible to build a single model using all groups, creating more fine-grained models allowed the authors to balance accuracy of prediction with robustness of the models. In a recent validation [210] on a database of 5581 molecules of F. Hoffmann-La Roche AG, this approach resulted in a RMSE of 1.09. This result was further improved to 0.49 after retraining with an additional 6226 pK_a values from in-house compounds.

The two previous approaches required a relatively large number of descriptors (several hundreds). However, compared to SPARC, these descriptors to some extent were only indirectly related to the ionization potential. pK_a is directly related to electronic properties of the ionizable center, and it is thus possible to develop models using a much smaller number of selected descriptors. Tehan *et al.* [155, 156] used quantum chemical descriptors based on frontier electron theory to describe the ionizable center and its neighbors. This corresponds to the use of level 1 and 2 neighborhoods in the notation of the previously discussed studies. Electrophilic superdelocalisability was found to be highly correlated with pK_a. The authors constructed equations using one to three descriptors for 15 data sets containing between 14 and 143 molecules, with an average RMSE of around 0.5 pK_a units. Larger errors were observed for bases, e.g., RMSEs of 1.85 and 1.4 were obtained for ortho pyridines and for pyrimidines. Such large errors may indicate that complex heterocyclic compounds, especially with nitrogen in an aromatic ring, are not adequately represented with the local neighborhood only.

Zhang [166] investigated whether more efficient descriptors could be proposed for the prediction of acids and alcohols. They introduced a new inductive descriptor $Q_{\sigma,i}$ that provides a weighted (by squared topological distance) sum of atomic partial charges. This descriptor had good correlation with Taft's constants ($R^2 = 0.85$) as well as with pK_a values of 1410 compounds ($R = -0.91$). Four other descriptors, describing accessibility of the central atom in 2D space, accessibility and polarizability of the acidic oxygen atom in an acid, π-electronegativity of the R-carbon atom in an acid, and, an indicator variable for α-amino acids, were used. The final model resulted in a $RMSE = 0.42$ and $R^2 = 0.81$ for 1122 aliphatic carboxylic acids. An analysis of 288 alcohols gave a similar $R^2 = 0.82$ with only four variables ($Q_{\sigma,o}$, σ-electronegativity of the oxygen atom in the acidic hydroxyl group, and two indicator variables). It is interesting that the correlation calculated using just the inductive descriptor ($R^2 = -0.91^2 = 0.83$) is higher than the reported individual correlations ($R^2 = 0.81$ and 0.82) for both subseries. This might be explained by the higher range of pK_a values (1-16) compared to the individual ranges of 1-6 for aliphatic carboxylic acids and of 4-16 for alcohols.

These analyses show that there is a good correlation between pK_a and simple and physically meaningful descriptors, and that this property can be predicted with reasonable accuracy. The studies of Tehan and Zhang used only monoprotic compounds, or compounds where the macro pK_a value could be unambiguously assigned to one ionizable group. Jelfs *et al.* [69] extended this approach by combining descriptors proposed in their work as well as in work by Xing *et al.* [172] for the prediction of multiprotic molecules. The authors attempted to identify a main path of ionization, starting from a neutral molecule and finding the "most basic group". Once such a group was found, it was ionized and the process was repeated. However, when several groups have very similar predicted pK_a values and thus compete with one another, the authors used a more accurate ranking. They first ionized each group and once again predicted pK_a values for the remaining neutral groups. The group with the higher basic pK_a was selected as ionized for the given round.

Studies by Kogej and Muresan [189] as well as by Lee *et al.* [186] show that even simpler methods, such as look-up in a database and/or SMART pattern search can be sufficient to develop reasonable models for pK_a prediction. In recent work [206], we employed kernel methods and graph kernels to predict pK_a with similar accuracy as the semi-empirical models of Tehan *et al.* [155, 156] on the same data.

3.3. Multiprotic Compounds

Most algorithms developed for pK_a prediction deal with monoprotic compounds or/and multiprotic compounds in which

the macro pK_a can be unambiguously related to the micro pK_a values. Several difficulties are associated with the prediction of microconstants for complex molecules with several ionizable centers. One is that considerably less data is available for microconstants, as it is more difficult to determine them experimentally. Therefore, a number of approaches [2, 69, 170] try to determine a main path of ionization, and thus treat macro pK_a as micro pK_a. However, there may be no unambiguous pathway of dissociation (which is why the software ACD/pK_a reports two microconstants for the same nitrogen of 3-[4-(dimethylamino)phenyl]acrylic acid [211]: it is simply not possible to report an unambiguous pathway and the microconstants closest to thermodynamic (averaged) pK_a).

3.4. Software

"There is immense interest in developing new and better software for pK_a prediction" [212].

A variety of mostly commercial programs exists for the prediction of pK_a values (Table **6**). The majority uses statistical approaches, in particular linear free energy relationships. Several comparative studies [109, 173, 212-215] have investigated the performance of some of these programs on different data sets. The focus of these studies was on predictive accuracy, but one should bear in mind that, in particular in industrial settings, other aspects such as documentation, usability, reliability, automation, batch processing, improvement of models via inclusion of in-house libraries, as well as long-term commitment and maintenance, are also important.

Two problems of comparative studies are that absolute performance statistics are not comparable, both due to the use of different performance statistics and different data sets (Table **7**), and, that performance might be artificially high for statistical approaches due to the use of literature data that are likely included in the training data sets of all major software suits [109]. These problems are reflected in the reported performance values. Moreover, the overlap in the sets of benchmarked programs between the studies is small; the only programs that were tested in all studies were Marvin and ACD/pK_a. A laudable exception is the study by Manchester *et al.* [215], who experimentally determined the pK_a values of 211 drug-like compounds not found in the literature and used these as their benchmarking set. In their study, errors were higher than in another study [109] based on literature data.

The programs ADME Boxes, ACD/ pK_a, and Marvin often occupy top ranks in the studies. This is somewhat attenuated by the possibility to train programs with own experimental data (i.e., extend their domain of applicability towards the in-house data). A quantitative estimate of the reliability of the prediction [232], e.g., an estimate of the prediction error, would be a useful feature here.

All programs that exclusively use statistics-based approaches (LFER, QSPR) are fast and can be applied to large compound libraries. SPARC is somewhat slower than these due to its inclusion of perturbed molecular orbital theory. Jaguar, the only quantum mechanical approach, is by far the slowest program. As an example, ADMET predictor estimated the pK_a values of 197 compounds in less than 1 s, whereas Jaguar took more than two days to predict the pK_a of the tertiary amine site of one of these compounds, hexobendine [109]. In this study, Jaguar was also rated worse in terms of prediction accuracy than its empirical competitors. The authors explain this by a lack of parameters for infrequent sites and close pK_a values for others; also, quantum chemical approaches were introduced more recently, and (commercially available) implementations might not yet have reached the level of maturity of the empirical ones. Interestingly, in the same study most programs performed worse on a subset of compounds with pK_a values in the range 5.4-9.4, but the performance of Jaguar remained the same.

Table 6. Software for pK_a Prediction. All Programs Support Microconstants

Program	Vendor/Organisation	Ref.	Method
ACD/pK_a	Advanced Chemistry Development	[216]	LFER
ADMET Predictor	Simulations Plus	[217]	QSPR
ADME Boxes*	Pharma Algorithms	[218]	QSPR
Epik	Schrödinger	[219, 220]	LFER
Jaguar	Schrödinger	[219, 221]	DFT/SCRF
Marvin	ChemAxon	[222, 223]	QSPR
MoKa	Molecular Discovery	[170, 224]	QSPR
Pallas/pKalc	CompuDrug	[225, 226]	LFER
Pipeline Pilot	Accelrys	[227]	QSPR
SPARC	University of Georgia	[152, 228]	LFER/PMO
OCHEM**	Helmholtz Center Munich	[206, 229]	QSPR

*The restricted version of this algorithm (only first acidic or basic pK_a values), is available through VCCLab [230, 231].
**Support for microconstants is planned, but not implemented so far.

Table 7. RMSE Values of Two Programs in Three Comparative Studies. Meloun and Bordovská (2007) [212] Use Three Separate Data Sets a, b, c. The Variance Between Data Sets is Greater than the Variance Between Programs

RMSE	[212]				
	a	b	c	[173]	[215]
ACD/Pka	0.35	0.22	0.54	0.26	0.8
Marvin	0.48	0.32	0.51	0.39	0.9

All in all, we concur with Liao & Nicklaus in that "the best pK_a predicting programs currently available are useful tools in the arsenal of the drug developer" [109].

4. DISCUSSION AND OUTLOOK

Over the last decades, many different approaches to the prediction of pK_a values of small molecules have been proposed. They can be roughly categorized into quantum mechanical *ab initio* calculations and empirical models based on statistics. The former can be subdivided into approaches using thermodynamic cycles (gas phase pK_a and direct approaches), the latter into linear free energy relationships (LFER) and descriptor-based statistical models. Approaches based on first principles offer the highest potential for general predictions. In practice, however, accuracy is often poor (absolute deviations of

about two log units [1]), limited mainly by the solvation models. Excessive computational demands are another problem. The LFER approach is the oldest one, introduced over 70 years ago [10, 11], and also the most mature one. One of the best-ranked programs (ACD/ pK_a by Advanced Chemistry Development) is based on LFERs. Later on, statistical approaches based on neural networks, and recently on kernel-based machine learning, were introduced for pK_a prediction. These purely empirical approaches usually deliver fair performance (absolute deviations of less than one log unit) and are fast enough to process large compound libraries. However, due to their nature they are limited to compounds similar to the ones used to parameterize the method.

In our opinion, improvements in prediction accuracy are most likely to be seen with *ab initio* calculations and statistical models, in particular those using kernel learning. However, one should keep in mind that statistical models have other disadvantages, e.g., they do not provide a succinct, explicit analytical formula in terms of descriptors, making interpretation of the model in physico-chemical terms difficult.

The single most important aspect in pK_a prediction are the data. Although a lot of measurements have been published and are publicly available, they are not easily accessible in electronic form, and data quality is a big problem. The best data can probably be found in the companies that offer commercial software for pK_a prediction. Since methodological innovation tends to come more from academia, this poses a problem. Increased cooperation between industrial and academic partners might be a solution here.

A problem in the assessment of both, programs as well as proposed methods for pK_a prediction, is the lack of a standard for evaluation, i.e., there is no common set of performance measures, retrospective validation procedure, and benchmark data sets. Although in most publications statistical measures, like correlation coefficient (r), determination coefficient (r^2), standard error (s) or Fisher's F-test, are given, a fair comparison of the methods is still not possible, due to a missing "golden standard" collection of test sets. If the training data set of a program based on an empirical method is not known, a fair comparison is impossible since predictive power might simply be a look-up of known data.

An aspect of pK_a prediction that is currently not considered enough is the domain of applicability [233]. Proposed methods should offer quantitative guidance on the reliability of each prediction, and an investigation of the reliability of these error estimates should be part of each study. Until now, such guidance is mostly available only in a very rough qualitative way, e.g., implicitly by the chemical series and substituents studied or used to construct models.

With respect to further method development, it has been argued that "a combination of first-principles based methods with QSPR-like descriptors appears ideal" [147], but it is not clear how such a combination could look like. Descriptors based on quantum mechanics have been used so far with good results [155, 156, 173]. Another possibility is to look for new developments in kernel-based learning, such as graph kernels [206, 234]; Gaussian process regression [235] provides built-in domain of applicability; multi-task learning might be used to predict pK_a in different solvents simultaneously.

ACKNOWLEDGEMENTS

This work was partially supported by the GO-Bio BMBF grant 0313883 "Development of ADME/T methods using associative neural networks: a novel self-learning software for confident ADME/T predictions". We thank Wolfram Teetz for helpful discussions, and an anonymous reviewer for detailed feedback.

APPENDIX

Derivation of Equation 10:

$$
\begin{aligned}
\log_{10}(\mathrm{K}_\mathrm{D}^{\mathrm{app}}) &= \log_{10}\frac{c_{\mathrm{li}}(\mathrm{HA})}{c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-)} = \log_{10}\frac{c_{\mathrm{li}}(\mathrm{HA})c_{\mathrm{aq}}(\mathrm{HA})}{c_{\mathrm{aq}}(\mathrm{HA})(c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-))} \\
&= \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) + \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{HA})}{c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-)} = \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-)}{c_{\mathrm{aq}}(\mathrm{HA})} \\
&= \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \mathrm{pH} + pH - \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-)}{c_{\mathrm{aq}}(\mathrm{HA})} \\
&\approx \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \mathrm{pH} - \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})}{c^{\ominus}} - \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{HA})+c_{\mathrm{aq}}(\mathrm{A}^-)}{c_{\mathrm{aq}}(\mathrm{HA})} \\
&= \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \mathrm{pH} - \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})c_{\mathrm{aq}}(\mathrm{HA}) + c_{\mathrm{aq}}(\mathrm{H_3O^+})c_{aq}(\mathrm{A}^-)}{c^{\ominus}c_{\mathrm{aq}}(\mathrm{HA})} \\
&= \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \mathrm{pH} - \log_{10}\left(\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})}{c^{\ominus}} + \frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})c_{\mathrm{aq}}(\mathrm{A}^-)}{c^{\ominus}c_{\mathrm{aq}}(\mathrm{HA})}\right) \\
&\approx \log_{10}\mathrm{K}_\mathrm{D}(\mathrm{HA}) - \mathrm{pH} - \log_{10}(\mathrm{H} + K_a).
\end{aligned}
\tag{13}
$$

(APPENDIX) contd…..

Derivation of Equation 11:

$$\log_{10}(\mathrm{K_D^{app}}) = \log_{10}\mathrm{K_D(HA)} - \mathrm{pH} - \log_{10}(\mathrm{H} + K_a)$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}}\right) = \mathrm{pH} + \log_{10}(\mathrm{H} + K_a)$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}}\right) = -\log_{10}\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})}{c^{\ominus}} + \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+}) + K_a\, c^{\ominus}}{c^{\ominus}}$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}}\right) = \log_{10}\frac{(c_{\mathrm{aq}}(\mathrm{H_3O^+}) + K_a\, c^{\ominus})c^{\ominus}}{c^{\ominus} c_{\mathrm{aq}}(\mathrm{H_3O^+})}$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}}\right) = \log_{10}\left(1 + \frac{K_a\, c^{\ominus}}{c_{\mathrm{aq}}(\mathrm{H_3O^+})}\right)$$

$$\Leftrightarrow \quad \frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}} = 1 + \frac{K_a\, c^{\ominus}}{c_{\mathrm{aq}}(\mathrm{H_3O^+})} \qquad (14)$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}} - 1\right) = \log_{10}\frac{K_a\, c^{\ominus}}{c_{\mathrm{aq}}(\mathrm{H_3O^+})}$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}} - 1\right) = \log_{10}K_{a-} \log_{10}\frac{c_{\mathrm{aq}}(\mathrm{H_3O^+})}{c^{\ominus}}$$

$$\Leftrightarrow \quad \log_{10}\left(\frac{\mathrm{K_D(HA)}}{\mathrm{K_D^{app}}} - 1\right) = \mathrm{pH\text{-}p}K_a.$$

REFERENCES

[1] Ho, J.; Coote, M. A universal approach for continuum solvent pK_a calculations: Are we there yet? *Theor. Chim. Acta*, **2010**, *125*(1-2), 3-21.

[2] Cruciani, G.; Milletti, F.; Storchi, L.; Sforna, G.; Goracci, L. *In silico* pKa prediction and ADME profiling. *Chem. Biodivers.*, **2009**, *6*(11), 1812-1821.

[3] Lee, A.; Crippen, G. Predicting pKa. *J. Chem. Inf. Model.*, **2009**, *49*(9), 2013-2033.

[4] Manallack, D. The pKa distribution of drugs: Application to drug discovery. *Perspect. Med. Chem.*, *1*, 25-38, **2007**.

[5] Fraczkiewicz, R. In silico prediction of ionization. In: B. Testa; H. van de Waterbeemd, eds., Comprehensive Medicinal Chemistry II, Elsevier, Oxford, England, **2006**, vol. *5*, pp. 603-626.

[6] Wan, H.; Ulander, J. High-throughput pKa screening and prediction amenable for ADME profiling. *Expert Opin. Drug Metab. Toxicol.*, **2006**, *2*(1), 139-155.

[7] Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.*, **2005**, 105(8), 2999-3094.

[8] Selassie, C. History of quantitative structure-activity relationships. In: D. Abrahams, ed., Burger's Medicinal Chemistry and Drug Discovery,Wiley, 2003, vol. *1*, chap. 1, pp. 1-48. 6th ed.

[9] Brown, A.C.; Fraser, T. On the connection between chemical constitution and physiological action. *Trans. R. Soc. Edinburgh*, **1868**, *25*, 1-53.

[10] Hammett, L. Some relations between reaction rates and equilibrium constants. *Chem. Rev.*, **1935**, *17*(1), 125-136.

[11] Hammett, L. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.*, **1937**, *59*(1), 96-103.

[12] Taft, R. Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J. Am. Chem. Soc*., **1952**, *74*(12), 3120-3128.

[13] Hansch, C.; Maloney, P.; Fujita, T. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, **1962**, *194*(4824), 178-180.

[14] Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.*, **1964**, *86*(23), 5175-5180.

[15] Free, S.; Wilson, J. A mathematical contribution to structure-activity studies. *J. Med. Chem.*, 1964, 7(4), 395-399.

[16] Kubinyi, H. Free Wilson analysis. Theory, applications and its relationship to Hansch analysis. *Quant. Struct. Act. Rel.*, **1988**, *7*(3), 121-133.

[17] Fujita, T.; Ban, T. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem.*, **1971**, *14*(2), 148-152.

[18] Bell, P.; Roblin, R.O., Jr. Studies in chemotherapy. VII. A theory of the relation of structure to activity of sulfanilamide type compounds. *J. Am. Chem. Soc.*, **1942**, *64*(12), 2905-2917.

[19] Albert, A.; Rubbo, S.; Goldacre, R.; Darcy, M.; Stove, J. The influence of chemical constitution on antibacterial activity II. A general survey of the acridine series. *Br. J. Exp. Pathol.*, **1945**, *26*, 160-192.

[20] Rossotti, F.; Rossotti, H. The Determination of Stability Constants and Other Equilibrium Constants in Solution. McGraw-Hill, **1961**.

[21] Kaplan, T. The chemical potential. *J. Stat. Phys.*, **2006**, *122*(6), 1237-1260.

[22] Job, G.; Herrmann, F. Chemical potential — a quantity in search of recognition. *Eur. J. Phys.*, **2006**, *27*(2), 353-371.

[23] McNaught, A.D.; Wilkinson, A. *Compendium of Chemical Terminology*, 2nd ed. *IUPAC Recommendations*. Blackwell Science, Oxford, England, **1997**.

[24] Hamer, W.; Wu, Y.C. Osmotic coefficients and mean activity coefficients of uni-univalent electrolytes in water at 25 C. *J. Phys. Chem. Ref. Data*, **1972**, *1*(4), 1047-1100.

[25] Hasselbalch, K.A. Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebundenen Kohlensäure desselben, und die

Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl. *Biochemische Zeitschrift*, **1916**, *78*, 112-144.

[26] Po, H.; Senozan, N. The Henderson-Hasselbalch equation: Its history and limitations. *J. Chem. Educ.*, **2001**, *78*(11), 1499-1503.

[27] de Levie, R. The Henderson approximation and the mass action law of Guldberg and Waage. *Chem. Educat.*, **2002**, *7*(3), 132-135.

[28] Szakács, Z.; Noszál, B. Protonation microequilibrium treatment of polybasic compounds with any possible symmetry. *J. Math. Chem.*, **1999**, *26*, 139-155.

[29] Ullmann, M. Relations between protonation constants and titration curves in polyprotic acids: A critical view. *J. Phys. Chem. B*, **2003**,*107*(5), 1263-1271.

[30] Marosi, A.; Kovács, Z.; Béni, S.; Kökösi, J.; Noszál, B. Triprotic acid-base microequilibria and pharmacokinetic sequelae of cetirizine. *Eur. J. Pharmaceut. Sci.*, **2009**, *37*(3-4), 321-328.

[31] Reichardt, C.; Welton, T. *Solvents and Solvent Effects in Organic Chemistry*, 4th ed. Wiley-VCH, Weinheim, **2010**.

[32] Pliego J., Jr. Thermodynamic cycles and the calculation of pK_a. *Chem. Phys. Lett.*, **2003**, *367*(1-2), 145-149.

[33] Cramer, C. Essentials of Computational Chemistry. Theory and Models.Wiley, West Sussex, 2004, 2nd ed.

[34] da Silva, C.; da Silva, E.; Nascimento, M. Comment on thermodynamic cycles and the calculation of pK_a *Chem. Phys. Lett.*, **2003**, *381*(1-2), 244-245.

[35] Pliego J., Jr. Reply to comment on: Thermodynamic cycles and the calculation of pK_a. *Chem. Phys. Lett.*, **2003**, *381*(1-2), 246-247.

[36] Atkins, P.; de Paula, J. Physical Chemistry. Oxford University Press, Oxford, England, 2010, 9th ed.

[37] Pliego J., Jr.; Riveros, J. Gibbs energy of solvation of organic ions in aqueous and dimethyl sulfoxide solutions. *Phys. Chem. Chem. Phys.*, **2002**, *4*(9), 1622-1627.

[38] Reynolds, C.A.; King, P.M.; Richards, W.G. Free energy calculations in molecular biophysics. *Mol. Phys.*, **1992**, *76*(2), 251-275.

[39] Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, **1993**, 93(7), 2395-2417.

[40] Jorgensen, W.L.; Tirado-Rives, J. Free energies of hydration for organic molecules from Monte Carlo simulations. *Perspect. Drug Discov. Des.*, **1995**, *3*(1), 123-138.

[41] Knight, J.L.; Brooks, C.L., III. λ-dynamics free energy simulation methods. *J. Comput. Chem.*, **2009**, *30*(11), 1692-1700.

[42] Tomasi, J.; Persico, M. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.*, **1994**, *94*(7), 2027-2094.

[43] Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.*, **1999**, *78*(1-2), 1-20.

[44] Bashford, D.; Case, D.A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **2000**, *51*, 129-152.

[45] Ulmschneider, M.B.; Ulmschneider, J.P.; Sansom, M.S.; Nola, A.D. A generalized Born implicit-membrane representation compared to experimental insertion free energies. *Biophys. J.*, **2007**, *92*(7), 2338-2349.

[46] Manson, P.A.; Morriss, G.P. Recent progress in the statistical mechanics of interaction site fluids. In: I. Prigogine; S.R. Rice, Eds., Advances in Chemical Physics, Wiley, New York, 1990, vol. *77*, pp. 451-550.

[47] Hirata, F., Ed. *Molecular Theory of Solvation*. Kluwer, Dordrecht, Netherlands, **2003**.

[48] Hansen, J.P.; McDonald, I.R. *Theory of Simple Liquids*, 3rd ed. Academic Press, Amsterdam, Netherlands, **2006**.

[49] Palmer, D.S.; Sergiievskyi, V.P.; Jensen, F.; Fedorov, M.V. Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *J. Chem. Phys.*, **2010**, *133*(4), 044104.

[50] Marenich, A.V.; Cramer, C.J.; Truhlar, D.G. Performance of SM6, SM8, and SMD on the SAMPL1 test set for the prediction of small-molecule solvation free energies. *J. Phys. Chem. B*, **2009**, *113*(14), 4538-4543.

[51] Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of the free energy of hydration of a challenging set of pesticide-like compounds. *J. Phys. Chem. B*, **2009**, *113*(14), 4508-4510.

[52] Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E.O. Prediction of SAMPL-1 hydration free energies using a continuum electrostatics-dispersion model. *J. Phys. Chem. B*, **2009**, *113*(14), 4511-4520.

[53] Perrin, D.; Dempsey, B.; Serjeant, E.P. *pKa Prediction for Organic Acids and Bases*. Chapman and Hall / CRC Press, Boca Raton, **1981**.

[54] Szegezdi, J.; Csizmadia, F. Tautomer generation. pK_a *based dominance conditions for generating dominant tautomers*. In: 234th National Meeting of the American Chemical Society, Boston, Massachusetts, USA, **2007**.

[55] Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley, New York, **1979**.

[56] Lombardo, F.; Faller, B.; Shalaeva, M.; Tetko, I.; Tilton, S. The good, the bad and the ugly of distribution coefficients: Current status, views and outlook. In: R. Mannhold, ed., Molecular Drug Properties. Measurement and Prediction, Wiley, Weinheim, Germany, 2008, chap. *16*, pp. 407-437.

[57] Mannhold, R.; Poda, G.I.; Ostermann, C.; Tetko, I.V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log p methods on more than 96,000 compounds. *J. Pharm. Sci.*, **2009**, *98*(3), 861-893.

[58] Kerns, E.H.; Di, L. Physicochemical profiling: Overview of the screens. *Drug Discov. Today Tech.*, **2004**, *1*(4), 343-348.

[59] Story, D.A. Bench-to-bedside review: A brief history of clinical acidbase. Crit. Care, **2004**, 8(4), 253-258, 2004.

[60] Nielsen, J.E.; McCammon, J.A. Calculating pK_a values in enzyme active sites. *Protein Sci.*, **2003**, *12*(9), 1894-1901.

[61] Gerlt, J.A.; Gassman, P.G. Understanding the rates of certain enzyme-catalyzed reactions: Proton abstraction from carbon acids, acyl transfer reactions, and displacement reactions of phosphodiesters. *Biochemistry*, **1993**, *32*(45), 11943-11952.

[62] Fitzpatrick, P.F. Substrate dehydrogenation by flavoproteins. *Accounts Chem. Res.*, **2001**, *34*(4), 299-307.

[63] Richard, J.P.; Amyes, T.L. Proton transfer at carbon. *Curr. Opin. Chem. Biol.*, **2001**, *5*(6), 626-633.

[64] Toney, M.D. Reaction specificity in pyridoxal phosphate enzymes. *Arch. Biochem. Biophys.*, **2005**, *433*(1), 279-287.

[65] Houk, R.J.; Monzingo, A.; Anslyn, E.V. Electrophilic coordination catalysis: A summary of previous thought and a new angle of analysis. *Accounts Chem. Res.*, **2008**, *41*(3), 401-410.

[66] Clark, D.E. In silico prediction of blood-brain barrier permeation. *Drug Discov. Today*, **2003**, *8*(20), 927-933, 2003.

[67] Upthagrove, A.; Nelson, W. Importance of amine pK_a and distribution coefficient in the metabolism of fluorinated propranolol derivatives. Preparation, identification of metabolite regioisomers, and metabolism by CYP2D6. *Drug Metab. Dispos.*, **2001**, *29*(11), 1377-1388.

[68] Lu, H.; Chen, X.; Zhan, C.G.G. First-principles calculation of pK_a for cocaine, nicotine, neurotransmitters, and anilines in aqueous solution. *J. Phys. Chem. B*, 2007, *111*(35), 10599-10605.

[69] Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pKa for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.*, **2007**, *47*(2), 450-459.

[70] Comer, J.; Tam, K. Lipophilicity profiles: Theory and measurement. In: B. Testa; H. van de Waterbeemd; G. Folkers; R. Guy, eds., Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical and Computational Strategies, Wiley, **2001**, pp. 275-304.

[71] Wells, J. *Pharmaceutical Preformulation*. Ellis Horwood, **1988**.

[72] Williams, D. pKa values for some drugs and miscellaneous organic acids and bases, 6th ed. In: T. Lemke; D. Williams; V. Roche; W. Zito, eds., Foye's Principles of Medicinal Chemistry, Lippincott Williams & Wilkins, 2008, pp. 1343-1353.

[73] Rice, N.; Irving, H.; Leonard, M. Nomenclature for liquid-liquid distribution (solvent extraction). *Pure Appl. Chem.*, **1993**, 65(11), 2373-2396.

[74] van der Giesen, W.F.; Janssen, L.H.M. Influence of ionization and ion-pair formation on lipophilicity of some 4-hydroxycoumarin derivatives in the octanol-water system. *Int. J. Pharm.*, **1982**, *12*(2-3), 231-249.

[75] Zhang, S.; Baker, J.; Pulay, P. A reliable and efficient first principles-based method for predicting pK_a values. 1. Methodology. J. *Phys. Chem. A*, **2010**, *114*(1), 425-431.

[76] Schüürmann, G. Ecotoxic modes of action of chemical substances. In: Schüürmann, G.; Markert, B., Eds.; Ecotoxicology. Ecological Fundamentals, Chemical Exposure, and Biological Effects, Wiley/Spektrum, New York/Heidelberg, 1998, chap. *22*, pp. 665-749.

[77] Terada, H. Uncouplers of oxidative phosphorylation. *Environ. Health Perspect.*, **1990**, *87*, 213-218.

[78] Schüürmann, G.; Somashekar, R.K.; Kristen, U. Structure-activity relationships for chloro- and nitrophenol toxicity in the pollen tube growth test. Environ. *Toxicol. Chem*., **1996**, *15*(10), 1702-1708.

[79] Schüürmann, G.; Segner, H.; Jung, K. Multivariate mode-of-action analysis of acute toxicity of phenols. *Aquat. Toxicol.*, **1997**, *38*(4), 277-296.

[80] Jover, J.; Bosque, R.; Sales, J. QSPR prediction of pK_a for benzoic acids in different solvents. *QSAR Comb. Sci.*, **2008**, 27(5), 563-581.

[81] Bosch, E.; Rived, F.; Rosés, M.; Sales, J. HammettTaft and Drago models in the prediction of acidity constant values of neutral and cationic acids in methanol. *J. Chem. Soc. Perkin Trans.*, **1999**, *2*, 1953-1958.

[82] Píchaa, J.; Cibulka, R.; Liškaa, F.; Paříkb, P.; Pytela, O. Reparametrization and/or determination of Hammett, inductive, mesomeric and AISE substituent constants for five substituents: $N^+(CH_3)_3$,$CH_2N^+(CH_3)_3$, CH_2Py^+,$CH_2SO_2CH_3$ and $PO(OCH_3)_2$. *Collect. Czech. Chem. Comm.*, **2004**, *69*(12), 2239-2252.

[83] Perrin, D.D. *Dissociation Constants of Weak Bases in Aqueous Solution, 1st Ed*. Butterworths, **1972**.

[84] Kortüm, G.; Vogel, W.; Andrussow, K., Eds. Dissociation Constants of Organic Acids in Aqueous Solution. Butterworths, London, **1961**.

[85] Serjeant, E.P.; Dempsey, B. *Ionization Constants of Organic Acids in Aqueous Solution*, vol. 23 of IUPAC Chemical Data Series. Pergamon Press, New York, **1979**.

[86] Izutsu, K. *Acid-Base Dissociation Constants in Dipolar Aprotic Solvents, vol. 35 of IUPAC Chemical* Data Series. Blackwell Science, **1990**.

[87] Albert, A. Ionization constants (of heterocyclic substances). In: A.R. Katritzky, Ed., Physical Methods in Heterocyclic Chemistry, Academic Press, 1963, vol. *1*.

[88] Sillén, L.; Martell, A. *Stability Constants of Metal-Ion Complexes*, vol. 17 of Special Publications. The Chemical Society, **1964**.

[89] Perrin, D., Ed. *Dissociation Constants of Organic Bases in Aqueous Solutions*. Butterworths, London, **1965**.

[90] Izatt, R.; Christensen, J. Heats of proton ionization and related thermodynamic quantitites. In: Sober [91], pp. J-49.

[91] Sober, H.A., Ed. *Handbook of Biochemistry*: Selected Data for Molecular Biology. Chemical Rubber Company, **1968**.

[92] Jencks, W.P.; Regenstein, J.M. *Ionization constants of acids and bases*. In: Sober [91], pp. J-150.

[93] Perrin, D., Ed. *Dissociation Constants of Inorganic Acids and Bases in Aqueous Solution*. Butterworths, London, **1969**.

[94] Sillén, L.; Martell, A. *Stability Constants of Metal-Ion Complexes*, vol. 25 of Special Publication.The Chemical Society, **1971**.

[95] Martell, A.; Smith, R. *Critical Stability Constants*, vol. 1-6. Plenum Press, **1974**.

[96] Perrin, D.D. *Dissociation Constants of Organic Bases in Aqueous Solution, 2nd Ed*. Elsevier, **1976**.

[97] Perrin, D.D. *Ionisation constants of inorganic acids and bases in aqueous solution*, vol. 29 of IUPAC Solubility Data Series. 2nd Ed. Pergamon Press.

[98] Albert, A.; Serjeant, E.P. *The Determination of Ionization Constants, 3rd Ed*. A laboratory manual. Chapman and Hall, **1984**.

[99] Drayton, C., Ed. *Cumulative Subject Index & Drug Compendium*, vol. 6 of Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Applications of Chemical Compounds. Pergamon Press, **1990**.

[100] Avdeef, A. *Absorption and Drug Development. Solubility, Permeability, and Charge State*. Wiley, Hoboken, New Jersey, **2003**.

[101] Speight, J., Ed. *Lange's Handbook of Chemistry*. 16th ed. McGraw-Hill, **2004**.

[102] Lide, D., Ed. *CRC Handbook of Chemistry and Physics*. 87th ed. CRC Press, **2006**.

[103] O'Neill, M., Ed. *The Merck Index*. 14th ed. Merck Publishing, **2006**.

[104] Prankerd, R.J. *Critical Compilation of pKa Values for Pharmaceutical Substances*, vol. 33 of Profiles of Drug Substances, Excipients, and Related Methodology.Elsevier, **2007**.

[105] Shalaeva, M.; Kenseth, J.; Lombardo, F.; Bastin, A. Measurement of dissociation constants (pK_a values) of organic compounds by multiplexed capillary electrophoresis using aqueous and cosolvent buffers. *J. Pharm. Sci.*, **2008**, *97*(7), , 2581-2606.

[106] Martin, A.N.; Swarbrick, J.; Cammarata, A. *Physical Pharmacy: Physical Chemical Principles in the Pharmaceutical Sciences*. 2nd ed. Lea & Febiger, **1969**.

[107] Kazakevich, Y.; LoBrutto, R., Eds. *HPLC for Pharmaceutical Scientists*. Wiley, Hoboken, New Jersey, **2007**.

[108] Takács-Novák, K.; Avdeef, A. Interlaboratory study of log p determination by shake-flask and potentiometric methods. *J. Pharmaceut. Biomed. Anal.*, **1996**, *14*(11), 1405-1413.

[109] Liao, C.; Nicklaus, M. Comparison of nine programs predicting pK_a values of pharmaceutical substances. *J. Chem. Inf. Model.*, **2009**, *49*(12), 2801-2812.

[110] Becke, A. Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys., 1993, 98(7), 5648- 5652.

[111] Lee, C.; Yang, W.; Parr, R. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, **1988**, *37*(2), 785-789.

[112] Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*. CRC Press, Boca Raton, Florida, **1984**.

[113] CODESSA (comprehensive descriptors for structural and statistical analysis), reference manual, Semichem (Shawnee, Kansas, USA; www.semichem.com) and the University of Florida, (accessed 2010-07-22).

[114] Klamt, A.; Schüürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans.*, **1993**, *2*, 799-805.

[115] Dragon software for the calculation 3224 descriptors, Talete srl, Milano, Italy. www.talete.mi.it, (accessed 2010-07-22).

[116] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, 2nd ed. Wiley-VCH, Weinheim, Germany, **2009**.

[117] Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Montgomery, Jr., J.A.; Vreven, T.; Kudin, K.N.; Burant, J.C.; Millam, J.M.; Iyengar, S.S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.E.; Hratchian, H.P.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Ayala, P.Y.; Morokuma, K.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Zakrzewski, V.G.; Dapprich, S.; Daniels, A.D.; Strain, M.C.; Farkas, O.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Ortiz, J.V.; Cui, Q.; Baboul, A.G.; Clifford, S.; Cioslowski, J.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.L.; Fox, D.J.; Keith, T.; Al-Laham, M.A.; Peng, C.Y.; Nanayakkara, A.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Gonzalez, C.; Pople, J.A. Gaussian 98, Gaussian inc., Pittsburgh, Pennsylvania, and Wallingford, Connecticut, USA, 2004. www.gaussian.com, (accessed 2010-07-22).

[118] Hoe, W.M.; Cohen, A.; Handy, N. Assessment of a new local exchange functional OPTX. *Chem. Phys. Lett.*, **2001**, 341(3-4), 319-328.

[119] Dewar, M.J.; Dougherty, R.C. *The PMO Theory of Organic Chemistry*. Plenum Press, **1975**.

[120] Popelier, P.L. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A*, **1999**, *103*(15), 2883-2890.

[121] O'Brien, S.E.; Popelier, P.L. Quantum molecular similarity. 3. QTMS descriptors. *J. Chem. Inform. Comput. Sci.*, **2001**, *41*(3), 764-775.

[122] Haykin, S. *Neural Networks*: A Comprehensive Foundation. Prentice Hall, **1998**.

[123] Yao, X.; Wang, Y.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemometr. Intell. Lab. Syst.*, **2002**, *62*(2), 217-225.

[124] Rocha, G.B.; Freire, R.O.; Simas, A.M.; Stewart, J.J. RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.*, 2006, *27*(10), 1101-1111.

[125] Chambers, C.C.; Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.*, **1996**, *100*(40), 16385-16398.

[126] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.*, **1988**, *28*(1), 31-36.

[127] Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, **2000**.

[128] Ivanciuc, O. Applications of support vector machines in chemistry. In: K. Lipkowitz; T. Cundari, eds., Reviews in Computational Chemistry, Wiley, Hoboken, 2007, vol. *23*, chap. 6, pp. 291-400.

[129] Jorgensen, W.L.; Briggs, J.M. A priori pK_a calculations and the hydration of organic anions. *J. Am. Chem. Soc.*, **1989**, *111*(12), 4190-4197.

[130] Schüürmann, G.; Cossi, M.; Barone, V.; Tomasi, J. Prediction of the pK_a of carboxylic acids using the *ab initio* continuum-solvation model PCM-UAHF. *J. Phys. Chem. A*, **1998**, *102*(33), 6706-6712.

[131] da Silva, C.; da Silva, E.; Nascimento, M. *Ab initio* calculations of absolute pK_a values in aqueous solution I. Carboxylic acids. *J. Phys. Chem. A*, **1999**, *103*(50), 11194-11199.

[132] da Silva, C.; da Silva, E.; Nascimento, M. *Ab initio* calculations of absolute pK_a values in aqueous solution II. Aliphatic alcohols, thiols, and halogenated carboxylic acids. *J. Phys. Chem. A*, **2000**, *104*(11), 2402-2409.

[133] Topol, I.; Tawa, G.; Caldwell, R.; Eissenstat, M.; Burt, S. Acidity of organic molecules in the gas phase and in aqueous solvent. *J. Phys. Chem. A*, **2000**, *104*(42), 9619-9624.

[134] Gross, K.C.; Seybold, P.G.; Peralta-Inga, Z.; Murray, J.S.; Politzer, P. Comparison of quantum chemical parameters and Hammett constants in correlating pKa values of substituted anilines. *J. Org. Chem.*, **2001**, *66*(21), 6919-6925.

[135] Liptak, M.; Shields, G. Accurate pK_a calculations for carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods. *J. Am. Chem. Soc.*, **2001**, *123*(30), 7314-7319.

[136] Liptak, M.D.; Gross, K.C.; Seybold, P.G.; Feldgus, S.; Shields, G.C. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.*, **2002**, *124*(22), 6421-6427.

[137] Pliego, J.R.; Riveros, J.M. Theoretical calculation of pK_a using the cluster-continuum model. *J. Phys. Chem. A*, **2002**, *106*(32), 7434-7439.

[138] am Busch, M.S.; Knapp, E.W. Accurate pK_a determination for a heterogeneous group of organic molecules. *ChemPhysChem*, **2004**, *5*(10), 1513-1522.

[139] Magill, A.; Cavell, K.; Yates, B. Basicity of nucleophilic carbenes in aqueous and nonaqueous solvents— theoretical predictions. *J. Am. Chem. Soc*., **2004**, *126*(28), 8717-8724.

[140] Ma, Y.; Gross, K.C.; Hollingsworth, C.A.; Seybold, P.G.; Murray, J.S. Relationships between aqueous acidities and computed surface-electrostatic potentials and local ionization energies of substituted phenols and benzoic acids. *J. Mol. Model.*, **2004**, *10*(4), 235-239.

[141] Brown, T.N.; Mora-Diez, N. Computational determination of aqueous pK_a values of protonated benzimidazoles (part 1). *J. Phys. Chem. B*, **2006**, *110*(18), 9270-9279.

[142] Namazian, M.; Halvani, S. Calculations of pKa values of carboxylic acids in aqueous solution using density functional theory. *J. Chem. Therm.*, **2006**, *38*(12), 1495-1502.

[143] Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P.K. pK_a prediction using group philicity. *J. Phys. Chem. A*, **2006**, *110*(20), 6540-6544.

[144] Dissanayake, D.; Senthilnithy, R. Thermodynamic cycle for the calculation of *ab initio* pK_a values for hydroxamic acids. *J. Mol. Struct.*, **2009**, *910*(1-3), 93-98.

[145] Jang, Y.H.; Hwang, S.; Chang, S.B.; Ku, J.; Chung, D.S. Acid dissociation constants of melamine derivatives from density functional theory calculations. *J. Phys. Chem. A*, **2009**, *113*(46), 13036-13040.

[146] Liu, S.; Pedersen, L.G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J. Phys. Chem. A*, **2009**, *113*(15), 3648-3655.

[147] Zhang, S.; Baker, J.; Pulay, P. A reliable and efficient first principles-based method for predicting pK_a values. 2. Organic acids. *J. Phys. Chem. A*, **2010**, *114*(1), 432-442.

[148] Zeng, Y.; Qian, H.; Chen, X.; Li, Z.; Yu, S.; Xiao, X. Thermodynamic estimate of pK_a values of the carboxylic acids in aqueous solution with the density functional theory. Chin. J. Chem., **2010**, 28(5), 727-733.

[149] Dixon, S.L.; Jurs, P.C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.*, **1993**, *14*(12), 1460-1467.

[150] Klopman, G.; Fercu, D. Application of the multiple computer automated structure evaluation methodology to a quantitative structure-activity relationship study of acidity. J. Comput. Chem., **1994**, 15(9), 1041-1050, 1994.

[151] Hilal, S.H.; Karickhoff, S.W.; Carreira, L.A. A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pKas. *Quant. Struct. Act. Rel.*, **1995**, *14*(4), 348-355

[152] Lee, P.; Ayyampalayam, S.; Carreira, L.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. In silico prediction of ionization constants of drugs. *Mol. Pharm.*, **2007**, *4*(4), 498-512.

[153] Seybold, P.G. Analysis of the pKas of aliphatic amines using quantum chemical descriptors. *Int. J. Quant. Chem*., **2009**, *108*(15), 2849-2855.

[154] Gross, K.; Seybold, P. Substituent effects on the physical properties and pK_a of aniline. *Int. J. Quant. Chem.*, **2000**, *80*(4), 1107-1115.

[155] Tehan, B.; Lloyd, E.; Wong, M.; Pitt, W.; Montana, J.; Manallack, D.; Gancia, E. Estimation of pKa using semiempirical molecular orbital methods. Part 1: Application to phenols and carboxylic acids. *Quant. Struct. Act. Rel*., **2002**, *21*(5), 457-472.

[156] Tehan, B.; Lloyd, E.; Wong, M.; Pitt, W.; Gancia, E.; Manallack, D. Estimation of pKa using semiempirical molecular orbital methods. Part 2: Application to amines, anilines and various nitrogen containing heterocyclic compounds. *Quant. Struct. Act. Rel.*, **2002**, *21*(5), 473-485.

[157] Zevatskii, Y.; Lysova, S. Empirical procedure for the calculation of ionization constants of organic compounds in water from their molecular volume. Russ. *J. Org. Chem.*, **2009**, *45*(6), 825-834.

[158] Seybold, P.G. Analysis of the pK_as of aliphatic amines using quantum chemical descriptors. *Int. J. Quant. Chem*., **2008**, *108*(15), 2849-2855.

[159] Ohta, K. Prediction of pK_a values of alkylphosphonic acids. *Bull. Chem. Soc. Jpn*., **1992**, *65*(9), 2543-2545.

[160] Gross, K.C.; Seybold, P.G.; Hadad, C.M. Comparison of different atomic charge schemes for predicting pK_a variations in substituted anilines and phenols. *Int. J. Quant. Chem*., **2002**, *90*(1), 445-458.

[161] Potemkin, V.A.; Pogrebnoy, A.A.; Grishina, M.A. Technique for energy decomposition in the study of receptor-ligand complexes. *J. Chem. Inf. Model.*, **2009**, *49*(6), 1389-1406.

[162] Duchowicz, P.R.; Castro, E.A. QSPR study of the acidity of carbon acids in aqueous solutions. *Mendeleev Commun.*, **2002**, *12*(5), 187-189.

[163] Ghasemi, J.; Saaidpour, S.; Brown, S.D. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct.*, **2007**, *805*(1-3), 27-32.

[164] Grüber, C.; Buß, V. Quantum-mechanically calculated properties for the development of quantitative structure-activity relationships (QSAR's). pKa-values of phenols and aromatic and aliphatic carboxylic acids. *Chemosphere*, **1989**, *19*(10-11), 1595-609.

[165] Citra, M.J. Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere*, **1999**, *38*(1), 191-206.

[166] Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.*, **2006**, *46*(6), 2256-2266.

[167] Soriano, E.; Cerdán, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct.*, **2004**, *684*(1-3), 121-128.

[168] Ivanova, A.; Baskin, I.; Palyulin, V.; Zefirov, N. Estimation of ionization constants for different classes of organic compounds with the use of the fragmental approach to the search of structure-property relationships. *Dokl. Chem.*, **2007**, *413*(2), 90-94, 2007.

[169] Luan, F.; Ma, W.; Zhang, H.; Zhang, X.; Liu, M.; Hu, Z.; Fan, B. Prediction of pK_a for neutral and basic drugs based on radial basis function neural networks and the heuristic method. *Pharmaceutical Research*, **2005**, *22*(9), 1454-1460, 2005.

[170] Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.*, **2007**, *47*(6), 2172-2181, 2007.

[171] Chaudry, U.A.; Popelier, P.L.A. Estimation of pK_a using quantum topological molecular similarity descriptors: Application to carboxylic acids, anilines and phenols. *J. Org. Chem*., 2004, *69*(2), 233-241, 2004.

[172] Xing, L.; Glen, R.C.; Clark, R.D. Predicting pKa by molecular tree structured fingerprints and PLS. *J. Chem. Inform. Comput. Sci.*, **2003**, *43*(3), 870-879, 2003.

[173] Harding, A.; Wedge, D.; Popelier, P. pKa prediction from "quantum chemical topology". *J. Chem. Inf. Model.*, **2009**, *49*(8), 1914-1924.

[174] Roy, K.; Popelier, P. Predictive QSPR modeling of the acidic dissociation constant (pKa) of phenols in different solvents. *J. Phys. Org. Chem.*, **2009**, *22*(3), 186-196.

[175] Gieleciak, R.; Polanski, J. Modeling robust QSAR. 2. Iterative variable elimination schemes for COMSA: Application for modeling benzoic acid pK_a values. *J. Chem. Inf. Model.*, **2007**, *47*(2), 547-556.

[176] Kim, K.H.; Martin, Y.C. Direct prediction of linear free energy substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substituted benzoic acids. *JOrgChem*, **1991**, *56*(8), 2723-2729.

[177] Kim, K.H.; Martin, Y.C. Direct prediction of dissociation constants (pK_as) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (COMFA) approach. *J. Med. Chem.*, **1991**, *34*(7), 2056-2060.

[178] Gargallo, R.; Sotriffer, C.A.; Liedl, K.R.; Rode, B.M. Application of multivariate data analysis methods to comparative molecular field analysis (COMFA) data: Proton affinities and pK_a prediction for nucleic acids components. *J. Comp. Aided Mol. Des.*, **1999**, *13*(6), 611-623.

[179] Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. Prediction acidity constant of various benzoic acids and phenols in water using linear and nonlinear QSPR models. *Bull. Kor. Chem. Soc.*, **2005**, *26*(12), 2007-2016.

[180] Jover, J.; Bosque, R.; Sales, J. Neural network based QSPR study for predicting pK_a of phenols in different solvents. *QSAR Comb. Sci.*, **2007**, *26*(3), 385-397.

[181] Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water. *Monatsh. Chem. Chem. Mon.*, **2009**, *140*(1), 15-27, 2009.

[182] Goodarzi, M.; Freitas, M.P.; Wu, C.H.; Duchowicz, P.R. pK_a modeling and prediction of a series of pH indicators through genetic algorithm-least square support vector regression. *Chemometr. Intell. Lab. Syst.*, 2010, *101*(2), 102-109.

[183] Goudarzi, N.; Goodarzi, M. Prediction of the acidic dissociation constant (pK_a) of some organic compounds using linear and nonlinear QSPR methods. *Mol. Phys.*, **2009**, *107*(14), 1495-1503.

[184] Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M.E. First principles calculations of aqueous pK_a values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK_a scale. *J. Phys. Chem. A*, **2003**, *107*(44), 9380-9386.

[185] Eckert, F.; Klamt, A. Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.*, **2005**, 27(1), 11-19.

[186] Lee, A.; Yu, J.-Y.; Crippen, G. pK_a prediction of monoprotic small molecules the SMARTS way. *J. Chem. Inf. Model.*, **2008**, 48(10), 2042-2053.

[187] Pompe, M.; Randić, M. Variable connectivity model for determination of pK_a values for selected organic acids. *Acta Chim. Slov.*, **2007**, *54*, 605-610.

[188] Potter, M.J.; Gilson, M.K.; McCammon, J.A. Small molecule pK_a prediction with continuum electrostatics calculations. *J. Am. Chem. Soc.*, **1994**, *116*(22), 10298-10299.

[189] Kogej, T.; Muresan, S. Database mining for pKa prediction. *Curr. Drug Discov. Tech.*, **2005**, *2*(4), 221-229.

[190] Müller, P. Glossary of terms used in physical organic chemistry. *Pure Appl. Chem.*, **1994**, *66*(5), 1077-1184.

[191] Ertl, P. Simple quantum chemical parameters as an alternative to the Hammett sigma constants in QSAR studies. *Quant. Struct. Act. Rel.*, **1997**, *16*(5), 377-382, 1997.

[192] Jaffé, H.H. A re-examination of the Hammett equation. *Chem. Rev.*, **1953**, *53*(2), 191-261, 1953.

[193] Wells, P.R. *Linear Free Energy Relationships.* Academic Press, **1968**.

[194] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York, **2009**.

[195] Schmidt, M.; Lipson, H. Distilling free-form natural laws from experimental data. *Science*, **2009**, *324*(5923), 81-85.

[196] Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. Chemometr. *Intell. Lab. Syst.*, **2001**, *58*(2), 109-130.

[197] Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural networks applied to quantitative structure-activity relationship analysis. *J. Med. Chem.*, **1990**, *33*(9), 2583-2590.

[198] A New QSAR Algorithm Combining Principal Component Analysis with a Neural Network: Application to Calcium Channel Antagonists; http://www.netsci.org/Science/Compchem/feature07.html) network science, (accessed 2010-08-20).

[199] So, S.S.; Karplus, M. Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. *J. Med. Chem.*, **1996**, *39*(7), 1521-1530.

[200] Lohninger, H. Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters. *J. Chem. Inf. Comput. Sci.*, **1993**, *33*(5), 736-744.

[201] Livingstone, D.J., Ed. *Artificial Neural Networks: Methods and Applications (Methods in Molecular Biology)*. Humana Press, **2008**.

[202] Hofmann, T.; Schölkopf, B.; Smola, A. Kernel methods in machine learning. *Ann. Stat.*, **2008**, *36*(6), 1171-1220.

[203] Schölkopf, B.; Smola, A.; Müller, K.R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural. Comput.*, **1998**, *10*(5), 1299-1319.

[204] Rosipal, R.; Trejo, L. Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, **2001**, *2*, 97-123.

[205] Bennett, K.; Embrechts, M. An optimization perspective on kernel partial least squares regression. In: J. Suykens; G. Horváth; S. Basu; C. Micchelli; J. Vandewalle, eds., Proceedings of the NATO Advanced Study Institute on Learning Theory and Practice, Leuven, Belgium, IOS Press, 2002, chap. *11*, pp. 227-250.

[206] Rupp, M.; Körner, R.; Tetko, I. Estimation of acid dissociation constants using graph kernels. *Mol. Inf.*, **2010**, *29*(10), 731-740.

[207] Rupp, M.; Proschak, E.; Schneider, G. Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.*, **2007**, *47*(6), 2280-2286.

[208] Newton, D.W.; Kluza, R.B. pKa values of drug substances and pH values of tissue fluids. In: Principles of Medicinal Chemistry, Lea & Febiger, Philadelphia, 1989, pp. 861-872.

[209] Carreira, L.A.; Hilal, S.; Karickhoff, S.W. Estimation of chemical reactivity parameters and physical properties of organic molecules using SPARC. In: P. Politzer; J.S. Murray, eds., Quantitative Treatments of Solute/Solvent Interactions, Elsevier, New York, 1994, chap. *9*, pp. 291-309.

[210] Milletti, F.; Storchi, L.; Goracci, L.; Bendels, S.; Wagner, B.; Kansy, M.; Cruciani, G. Extending pKa prediction accuracy: High-throughput pKa measurements to understand pKa modulation of new chemical series. *Eur. J. Med. Chem.*, **2010**, *45*(9), 4270-4279.

[211] Kolovanov, E.; Bhal, S. Why does ACD/pKa DB sometimes report the dissociation of a centre twice? Tech. rep., Advanced Chemistry Development Inc., Toronto, Canada, www.acdlabs.com, 2006.

[212] Meloun, M.; Bordovská, S. Benchmarking and validating algorithms that estimate pKa values of drugs based on their molecular structures. *Anal. Bioanal. Chem*, **2007**, *389*(4), 1618-1642.

[213] Dearden, J.; Cronin, M.T.; Lappin, D. A comparison of commercially available software for the prediction of pK_a values. In: UK-QSAR and ChemoInformatics Group Spring Meeting, AstraZeneca, Alderley Park, UK, **2007**.

[214] Balogh, G.T.; Gyarmati, B.; Nagy, B.; Molnár, L.; Keserü, G.M. Comparative evaluation of in silico pK_a prediction tools on the gold standard dataset. *QSAR Comb. Sci.*, **2009**, *28*(10), 1148-1155.

[215] Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of pK_a estimation methods on 211 druglike compounds. J. *Chem. Inf. Model.*, **2010**, *50*(4), 565-571.

[216] Advanced Chemistry Development Inc., Toronto, Canada. www.acdlabs.com, (accessed 2010-07-22).

[217] Simulations Plus Inc., Lancaster, California, USA. www.simulations-plus.com, (accessed 2010-07-22).
[218] Pharma Algorithms, Toronto, Canada. www.ap-algorithms.com, (accessed 2010-07-22).
[219] Schrödinger LLC, New York, New York, USA. www.schrodinger.com, (accessed 2010-07-22).
[220] Shelley, J.C.; Cholleti, A.; Frye, L.L.; Greenwood, J.R.; Timlin, M.R.; Uchimaya, M. Epik: A software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.*, **2007**, *21*(12), 681-691.
[221] Klicić, J.J.; Friesner, R.A.; Liu, S.Y.; Guida, W.C. Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. *J. Phys. Chem. A*, **2002**, *106*(7), 1327-1335.
[222] ChemAxon Ltd., Budapest, Hungary. www.chemaxon.com/marvin, (accessed 2010-07-22).
[223] Szegezdi, J.; Csizmadia, F. *A method for calculating the* p*K$_a$ values of small and large molecules*. In: 233rd National Meeting of the American Chemical Society, Chicago, Illinois, USA, **2007**.
[224] Molecular Discovery Ltd., Pinner, United Kingdom. www.moldiscovery.com, (accessed 2010-07-22).
[225] CompuDrug Inc., Sedona, Arizona, USA. www.compudrug.com, (accessed 2010-07-22).
[226] Csizmadia, F.; Szegezdi, J.; Darvars, F. Expert system approach for predicting p*K$_a$*. In: C.G. Wermuth, ed., Proceedings of the 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modelling, Strasbourg, France, Escom, 1992, pp. 507-510.
[227] Accelrys inc., San Diego, California, USA. www.accelrys.com, (accessed 2010-07-22).
[228] University of Georgia, Athens, Georgia, USA. www.ibmlc2.chem.uga.edu/sparc, (accessed 2010-07-22).
[229] Helmholtz Center Munich, Neuherberg, Germany. www.ochem.eu, www.qspr.eu, (accessed 2010-07-22).
[230] Virtual Computational Chemistry Laboratory, Germany. www.vcclab.org, (accessed 2010-07-22).
[231] Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; Tanchuk, V.Y.; Prokopenko, V.V. Virtual computational chemistry laboratory—design and description. *J. Comput. Aided Mol. Des*., **2005**, *19*(6), 453-463.
[232] Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.*, **2008**, *48*(9), 1733-1746.
[233] Tetko, I.; Bruneau, P.; Mewes, H.W.; Rohrer, D.; Poda, G. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today*, **2006**, *11*(15-16), 700-707.
[234] Rupp, M.; Schneider, G. Graph kernels for molecular similarity. *Mol. Inf.*, 2010, 2*9*(4), 266-273.
[235] Rasmussen, C.; Williams, C. Gaussian Processes for Machine Learning. MIT Press, Cambridge, **2006**.

Received: September 8, 2010 **Revised: January 29, 2011** **Accepted: March 1, 2011**