Predicting the \( pK_a \) of Small Molecules

Matthias Rupp*1,2, Robert Körner1 and Igor V. Tetko1

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

1Present 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, QSAR, 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 \( pK_a = -\log_{10} K_a \). 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 \textit{ab initio} calculations) of \( pK_a \) values belongs to the field of quantitative structure-property relationships (QSAR). The basic postulate in QSAR 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, \( \sigma \) and \( \rho \) 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](image-url)
1.1.2. pKₐ Estimation

QSAR studies involving pKₐ 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. pKₐ Values

According to the Brønsted-Lowry theory of acids and bases, an acid HA is a proton (hydrogen cation) donor, HA ⇌ H⁺ + A⁻, and a base B is a proton acceptor, B + H⁺ ⇌ BH⁺. For a weak acid in aqueous solution, the dissociation HA + H₂O ⇌ A⁻ + H₃O⁺ is reversible. In the forward reaction, the acid HA and water, acting as a base, yield the conjugate base A⁻ and oxonium H₃O⁺ (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 Kₐ, is the ratio of the activities of products and reagents,

\[
K_a = \frac{a(A^-)a(H_3O^+)}{a(HA)a(H_2O)},
\]  

(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:

\[
a(x) = \exp\left(\frac{\mu(x) - \mu^\oplus(x)}{RT}\right) = \gamma(x)\frac{c(x)}{c^\oplus},
\]  

(2)

where \(\mu(\cdot)\) is the chemical potential of a species under the given conditions (partial molar Gibbs energy\(^1\)), \(\mu^\oplus(\cdot)\) is the chemical potential of the species in a standard state (molar Gibbs energy), \(R = 8.314472(15)\) 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, \(c^\oplus = 1\) mol/L (or 1 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 \(c(H_2O) = c^\oplus = 1\) mol/L, inserting Equation 2 into Equation 1 yields an approximation valid for low concentrations of HA in water:

\[
K_a = \frac{c(A^-)c(H_3O^+)}{c(HA)c^\oplus}.
\]  

(3)

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

\[
pK_a = pH + \log_{10}\left(\frac{c(HA)}{c(A^-)}\right),
\]  

(4)

where \(pH = -\log_{10}(c(H_3O^+)) = -\log_{10}(c(H_2O)\cdot c^\oplus)\). In an ideal solution, the pKₐ 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. pKₐ Values

The protonation of a base B + H₂O ⇌ HB⁺ + HO⁻ can be described in the same terms as the deprotonation of an acid, leading to the base association constant \(K_b = a(BB^+)a(HO^-)/a(B)a(H_2O)\). Adding the reaction equations for the deprotonation of HA and the protonation of its conjugate base A⁻ gives 2H₂O ⇌ H₃O⁺ + OH⁻, with equilibrium constant \(K_w = a(H_3O^+)a(OH^-)/a^2(H_2O)\). It follows that \(K_w = K_aK_b\), and therefore \(pK_w = pK_a - pK_b = 14 - pK_a\), where \(pK_w = 14\) from \(c(H_2O) = c(HO^-) = 10^{-7}\) mol/L at \(T = 298.15K\) and under the same assumptions as for Equation 3. Since pKₐ and pKₜ use the same scale, pKₐ-values are used for both acids and bases; however, data in older references is sometimes given as pKₜ-values. For prediction, one should not mix pKₐ and pKₜ 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- pKₐ-s, 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,...,\(n\) protons bound). Fig. (1) presents cetirizine as an example. For \(n > 3\), micro- pKₐ-s 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;
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°C (whereas a temperature around 37°C 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

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

and

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

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,

$$\Delta G = \Delta G^\Theta + RT \ln \frac{a(A^-)a(H_3O^+)}{a(HA)a(H_2O)}.$$  (5)
At equilibrium, $\Delta G = 0$ and the last term equals $K_a$, yielding

$$-\Delta G^\circ = RT \ln K_a \Leftrightarrow pK_a = \frac{\Delta G^\circ}{RT} \ln 10 = \frac{\Delta G^\circ}{2.303RT}.$$  \hspace{1cm} (6)

At $T = 298.15 \text{K}$, we get $pK_a = \Delta G^\circ / (5708.02 \text{Jmol}^{-1})$. A difference of 5.71kJmol$^{-1}$ in $\Delta G^\circ$ thus corresponds to a unit difference in $pK_a$ value. To calculate $\Delta G^\circ$, the quantities $\Delta G_{\text{solv}}(\text{HA})$, $\Delta G_{\text{solv}}(\text{H}_2\text{O})$, $\Delta G^\circ$, and $\Delta G_{\text{solv}}(\text{H}_3\text{O}^+)$ have to be determined. Of these, $\Delta G_{\text{solv}}(\text{H}_2\text{O})$ and $\Delta G_{\text{solv}}(\text{H}_3\text{O}^+)$ have not to 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 $T = 298.15 \text{K}$) 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 \[ K_T = \frac{c(T2)}{c(T1)} = \frac{K_{a1}}{K_{a2}} \Leftrightarrow pK_T = pK_{a1} - pK_{a2}. \] \hspace{1cm} (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 \[ \log D = \log P - \log(1 + 10^{\delta - pK_a}), \] \hspace{1cm} (8)

where $\delta$ 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.](image-url)
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. (3). Example of the influence of steric effects on $pK_a$. cis/trans-isomerism in butenedioic acid causes marked changes in $pK_a$ values.](image)

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

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

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

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 $c_i$ in a lipid phase and its concentration $c_a$ in an aqueous phase at equilibrium, $K_D = c_i / c_a$. 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]

$$K_D^{app} = \frac{c_i(HA)}{c_a(HA) + c_a(A^-)}.$$  \hspace{1cm} (9)

Combining Equations 1 and 9 with the definition of pH and $K_D$ and taking logarithms yields

$$\log_{10} K_D^{app} = \log_{10} K_D(HA) - pH - \log_{10}(H + K_a).$$ \hspace{1cm} (10)

If pH=$pK_a$, then

$$\log_{10} K_D^{app} = \log_{10} K_D(HA) - 0.301.$$  \hspace{1cm} (11)

For pH$\ll$pK$_a$, Equation 10 can be approximated by

$$\log_{10} K_D^{app} = \log_{10} K_D(HA),$$  \hspace{1cm} (12)

and for pH$\gg$pK$_a$ by

$$\log_{10} K_D^{app} = \log_{10} K_D(HA) - pH + pK_a.$$ \hspace{1cm} (13)

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 $K_D^{app}$ and $K_D(AH)$ as

$$\log_{10} \left( \frac{K_D(HA)}{K_D^{app}} - 1 \right) = pH - pK_a.$$ \hspace{1cm} (11)
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 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]

<table>
<thead>
<tr>
<th>Data Set</th>
<th>1 Acid</th>
<th>1 Base</th>
<th>2 Acids</th>
<th>2 Bases</th>
<th>1 Acid &amp; 1 Base</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williams</td>
<td>24.4%</td>
<td>45.4%</td>
<td>3.8%</td>
<td>10.5%</td>
<td>11.2%</td>
<td>4.8%</td>
</tr>
<tr>
<td>World drug index</td>
<td>11.6%</td>
<td>42.9%</td>
<td>3.0%</td>
<td>24.6%</td>
<td>7.5%</td>
<td>10.4%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physico-Chemical</td>
<td></td>
</tr>
<tr>
<td>Lipophilicity</td>
<td>Neutral species are more lipophilic than ionized ones since less energy is required to remove the hydration layer</td>
</tr>
<tr>
<td>Solubility</td>
<td>Water is a polar solvent, and $pK_a$ thus directly influences solubility</td>
</tr>
<tr>
<td>Fundamental</td>
<td></td>
</tr>
<tr>
<td>pH homeostasis</td>
<td>Organisms maintain a constant pH in blood by using biological buffers. Disturbances in human acid-base balance are directly relevant in medicine [59]</td>
</tr>
<tr>
<td>Function</td>
<td>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]</td>
</tr>
<tr>
<td>Pharmaceutical</td>
<td></td>
</tr>
<tr>
<td>Absorption</td>
<td>Lipophilic species are absorbed better, e.g., intestinal uptake</td>
</tr>
<tr>
<td>BBB permeation</td>
<td>It has been suggested that protonation state influences BBB permeability [66]</td>
</tr>
<tr>
<td>Formulation</td>
<td>Choice of excipient and counter-ion</td>
</tr>
<tr>
<td>Metabolism</td>
<td>$pK_a$ can influence rate and site of metabolization [2, 67]</td>
</tr>
<tr>
<td>Signaling</td>
<td>Many neurotransmitters are ionizable amine compounds [68]</td>
</tr>
<tr>
<td>Pharmacodynamics</td>
<td>$pH$ in the human body varies between 2 and 12, with the microspecies population of a compound, and thus its behavior, varying accordingly [69]</td>
</tr>
</tbody>
</table>
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. \( 1.5 \times 10^5 \) (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].

<table>
<thead>
<tr>
<th>Name</th>
<th>Vendor</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACD/pKa DB</td>
<td>Advanced Chemistry Development Inc., Toronto, Canada. <a href="http://www.acdlabs.com">www.acdlabs.com</a></td>
<td>&gt;31 000</td>
</tr>
<tr>
<td>ADME index</td>
<td>Lighthouse Data Solutions LLC. <a href="http://www.bio-rad.com">www.bio-rad.com</a></td>
<td></td>
</tr>
<tr>
<td>BioLoom</td>
<td>BioByte Corp., Claremont, California, USA. <a href="http://www.biobyte.com">www.biobyte.com</a></td>
<td>14 000</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>European Bioinformatics Institute, Cambridge, UK. <a href="http://www.ebi.ac.uk/chembldb/">www.ebi.ac.uk/chembldb/</a></td>
<td>4 650</td>
</tr>
<tr>
<td>CRC handbook</td>
<td>Taylor and Francis Group LLC., New York, New York, USA. <a href="http://www.hbcpnetbase.com">www.hbcpnetbase.com</a></td>
<td></td>
</tr>
<tr>
<td>LOGKOW</td>
<td>Sangster Research Laboratories, Montréal, Québec, Canada. <a href="http://www.logkow.cisti.nrc.ca">www.logkow.cisti.nrc.ca</a></td>
<td></td>
</tr>
<tr>
<td>Merck index</td>
<td>Cambridgesoft Corp., Cambridge, Massachusetts, USA. <a href="http://www.cambridgesoft.com">www.cambridgesoft.com</a></td>
<td></td>
</tr>
<tr>
<td>MolSuit DB</td>
<td>ChemSW, FairField, California, USA. <a href="http://www.chemsw.com">www.chemsw.com</a></td>
<td></td>
</tr>
<tr>
<td>NIST std. ref. DB 46</td>
<td>National Institute of Standards and Technology, USA. <a href="http://www.nist.gov">www.nist.gov</a></td>
<td></td>
</tr>
<tr>
<td>OCHEM</td>
<td>Helmholtz Research Center for Environmental Health, Munich, Germany. <a href="http://www.ochem.eu">www.ochem.eu</a>, <a href="http://www.qspr.eu">www.qspr.eu</a></td>
<td>&gt;5 000</td>
</tr>
<tr>
<td>Pallas pKalc</td>
<td>CompuDrug Ltd., Sedona, Arizona, USA. <a href="http://www.compudrug.com">www.compudrug.com</a></td>
<td></td>
</tr>
<tr>
<td>PhysProp</td>
<td>Syracuse Research Corp., North Syracuse, USA. <a href="http://www.syres.com">www.syres.com</a></td>
<td></td>
</tr>
<tr>
<td>( pK ) database</td>
<td>University of Tartu, Estonia. <a href="http://www.mega.chem.ut.ee/tktool/teadus/pkdb/">www.mega.chem.ut.ee/tktool/teadus/pkdb/</a></td>
<td>&gt;20 000</td>
</tr>
<tr>
<td>pKaData</td>
<td>pKaData Ltd. <a href="http://www.pkadata.com">www.pkadata.com</a></td>
<td></td>
</tr>
<tr>
<td>SPARC</td>
<td>University of Georgia, USA. <a href="http://www.ibmlFc2.chem.uga.edu/sparc/">www.ibmlFc2.chem.uga.edu/sparc/</a></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Author (Year)</th>
<th>Comment</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>[84]</td>
<td>Kortüm et al. (1961)</td>
<td>Organic acids</td>
<td>2 893</td>
</tr>
<tr>
<td>[87]</td>
<td>Albert (1963)</td>
<td>Heterocyclic substances</td>
<td></td>
</tr>
<tr>
<td>[88]</td>
<td>Sillén and Martell (1964)</td>
<td>Metal-ion complexes</td>
<td></td>
</tr>
<tr>
<td>[89]</td>
<td>Perrin (1965)</td>
<td>Organic bases</td>
<td></td>
</tr>
<tr>
<td>[90]</td>
<td>Izatt and Christensen (1968)</td>
<td>Book chapter [91]</td>
<td></td>
</tr>
<tr>
<td>[92]</td>
<td>Jencks and Regenstein (1968)</td>
<td>Book chapter [91]</td>
<td></td>
</tr>
<tr>
<td>[93]</td>
<td>Perrin (1969)</td>
<td>Inorganic acids and bases</td>
<td>8 766</td>
</tr>
<tr>
<td>[94]</td>
<td>Sillén and Martell (1971)</td>
<td>Metal-ion complexes</td>
<td></td>
</tr>
<tr>
<td>[83]</td>
<td>Perrin (1972)</td>
<td>Weak bases</td>
<td>~4 300</td>
</tr>
<tr>
<td>[95]</td>
<td>Martell and Smith (1974)</td>
<td>NIST std. ref. database 46</td>
<td>6 166</td>
</tr>
<tr>
<td>[85]</td>
<td>Serjeant and Dempsey (1979)</td>
<td>Organic acids</td>
<td>~4 520</td>
</tr>
<tr>
<td>[53]</td>
<td>Perrin et al. (1981)</td>
<td>Hammet-Taft equations</td>
<td></td>
</tr>
<tr>
<td>[102]</td>
<td>Lide (2006)</td>
<td>CRC handbook</td>
<td></td>
</tr>
</tbody>
</table>
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 $pK_a$, $-\log_{10}(pK_a)$ or, $pK_b$ 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 $pK_{a_1} = 8.24$, $pK_{a_2} = 9.50$ [105], and as $pK_{a_1} = 7.4$, $pK_{a_2} = 8.6$ [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 $\alpha$ as excellent ($\alpha \leq 0.1$), well ($0.1 < \alpha \leq 0.5$), poor ($1.0 < \alpha \leq 2$), or awful ($2 < \alpha$) (with $0.5 < \alpha \leq 1$ 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, $F^2$, 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<sub>a</sub> Models

<table>
<thead>
<tr>
<th>Ref.</th>
<th>n</th>
<th>S</th>
<th>MP</th>
<th>Method</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ab initio</td>
<td>(MP3/MP2/6-31+G(d)/6-31+G(d))</td>
<td>organic acids</td>
</tr>
<tr>
<td>[129]</td>
<td>5</td>
<td>yes</td>
<td>no</td>
<td>ab initio (MP2, G2MP2, DFT B3LYP/6-311+G(dp); PCM)</td>
<td>pK&lt;sub&gt;a&lt;/sub&gt;, up to 50</td>
</tr>
<tr>
<td>[130]</td>
<td>16</td>
<td>yes</td>
<td>no</td>
<td>SCF (6-31G**//6-31+G**/6-311G (2d,2p)//6-31+G(2d,2p)); PCM-UAHF</td>
<td>aliphatic carboxylic acids</td>
</tr>
<tr>
<td>[131]</td>
<td>12</td>
<td>yes</td>
<td>no</td>
<td>ab initio (HF/6-31+G**, PCM)</td>
<td>carboxylic acids</td>
</tr>
<tr>
<td>[132]</td>
<td>15</td>
<td>yes</td>
<td>no</td>
<td>ab initio (HF/6-31+G**, PCM)</td>
<td>aliphatic alcohols, thiols, halogenated</td>
</tr>
<tr>
<td>[133]</td>
<td>8</td>
<td>no</td>
<td>yes</td>
<td>ab initio (MP2, G2MP2, DFT B3LYP/6-311+G(dp); PCM)</td>
<td>pK&lt;sub&gt;a&lt;/sub&gt;, up to 50</td>
</tr>
<tr>
<td>[134]</td>
<td>36</td>
<td>yes</td>
<td>no</td>
<td>MEP-V&lt;sub&gt;env&lt;/sub&gt;, MEP-V&lt;sub&gt;env&lt;/sub&gt;, I&lt;sub&gt;env&lt;/sub&gt;, Hammet (sigma);</td>
<td>anilines</td>
</tr>
<tr>
<td>[135]</td>
<td>6</td>
<td>no</td>
<td>yes</td>
<td>ab initio (CBS-QB3, CBS-APNO; PCM)</td>
<td>carboxylic acids</td>
</tr>
<tr>
<td>[136]</td>
<td>20</td>
<td>yes</td>
<td>no</td>
<td>ab initio (HF/CPCM with 8 different solvation models)</td>
<td>phenols</td>
</tr>
<tr>
<td>[137]</td>
<td>17</td>
<td>no</td>
<td>no</td>
<td>ab initio (MP2/6-311+G(2df,2p))</td>
<td></td>
</tr>
<tr>
<td>[138]</td>
<td>26</td>
<td>yes</td>
<td>no</td>
<td>ab initio (DFT B3LYP/6-31G** &amp; cc-pvqz, Beckel(1/2); two-step)</td>
<td>carboxylic acids, phenols, heterocycles</td>
</tr>
<tr>
<td>[139]</td>
<td>12</td>
<td>yes</td>
<td>no</td>
<td>ab initio (CBS-QB3, MP2/6-311+G(dp), HF/6-311G(dp)); PCM</td>
<td>pK&lt;sub&gt;a&lt;/sub&gt;, up to 34</td>
</tr>
<tr>
<td>[140]</td>
<td>36</td>
<td>yes</td>
<td>no</td>
<td>MEP-V&lt;sub&gt;env&lt;/sub&gt;, MEP-V&lt;sub&gt;env&lt;/sub&gt;, I&lt;sub&gt;env&lt;/sub&gt; and V&lt;sub&gt;env&lt;/sub&gt;;</td>
<td>phenols and benzoic acids</td>
</tr>
<tr>
<td>[141]</td>
<td>13</td>
<td>no</td>
<td>yes</td>
<td>ab initio (B3LYP/6-31+G(dp);PCM(opt)</td>
<td>13 different Methods, Basis Sets, Solvent</td>
</tr>
<tr>
<td>[142]</td>
<td>66</td>
<td>yes</td>
<td>no</td>
<td>ab initio (DFT B3LYP/6-31+G(dp); PCM)</td>
<td>carboxylic acids</td>
</tr>
<tr>
<td>[143]</td>
<td>63</td>
<td>yes</td>
<td>no</td>
<td>MEP-V&lt;sub&gt;env&lt;/sub&gt;, MEP-V&lt;sub&gt;env&lt;/sub&gt;, ab initio (HF/6-31G**)</td>
<td></td>
</tr>
<tr>
<td>[68]</td>
<td>24</td>
<td>yes</td>
<td>yes</td>
<td>ab initio (B3LYP/6-31+G*, MP2/6-311++G**)</td>
<td>alcohols, amines, anilines, carboxylic acids,</td>
</tr>
<tr>
<td>[144]</td>
<td>12</td>
<td>yes</td>
<td>yes</td>
<td>ab initio (CBS-QB3, HF/6-31G(dp); PCM)</td>
<td>hydroxamic acids</td>
</tr>
<tr>
<td>[145]</td>
<td>4</td>
<td>yes</td>
<td>no</td>
<td>PB continuum solvation; ab initio (B3LYP/6-311+G(dp))</td>
<td>different tautomers are considered</td>
</tr>
<tr>
<td>[146]</td>
<td>228</td>
<td>yes</td>
<td>no</td>
<td>MEP, DFT B3LYP/6-311+G(dp); PCM</td>
<td>amines, anilines, carboxylic acids,</td>
</tr>
<tr>
<td>[1]</td>
<td>55</td>
<td>no</td>
<td>no</td>
<td>direct, proton exchange, hybrid cluster-continuum, implicit-explicit;</td>
<td>neutral organic and inorganic acids</td>
</tr>
<tr>
<td>[75]</td>
<td>34</td>
<td>yes</td>
<td>no</td>
<td>(OLYP/3-21G(dp)/6-311+G**); COSMO</td>
<td>anilines, amines, phenols, alcohols,</td>
</tr>
<tr>
<td>[147]</td>
<td>370</td>
<td>yes</td>
<td>yes</td>
<td>ab initio (OLYP/3-21G(dp)/6-311+G**); COSMO</td>
<td>carboxylic acids, phosphonic acids, phenols,</td>
</tr>
<tr>
<td>[148]</td>
<td>85</td>
<td>yes</td>
<td>yes</td>
<td>ab initio (DFT B3LYP/6-31+G(dp); PCM)</td>
<td>carboxylic acids, direct approach</td>
</tr>
</tbody>
</table>

**Empirical**

<table>
<thead>
<tr>
<th>Ref.</th>
<th>n</th>
<th>S</th>
<th>Method</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>[149]</td>
<td>135</td>
<td>yes</td>
<td>no</td>
<td>LFER</td>
</tr>
<tr>
<td>[150]</td>
<td>3995</td>
<td>yes</td>
<td>no</td>
<td>LFER; fragmental and QM descriptors&lt;sup&gt;2&lt;/sup&gt;; MCASE</td>
</tr>
<tr>
<td>[151]</td>
<td>3685</td>
<td>yes</td>
<td>yes</td>
<td>LFER; SAR, PMO; SPARC</td>
</tr>
<tr>
<td>[152]</td>
<td>123</td>
<td>no</td>
<td>yes</td>
<td>LFER; SAR, PMO; SPARC</td>
</tr>
</tbody>
</table>

<sup>2</sup>Including fragments, partition coefficient, water solubility, molecular weight, Hückel molecular orbital charge densities, HOMO, LUMO, absolute electronegativity, and hardness.
3.2.1. Ab Initio Calculations

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

$$\log_{10} \frac{K_a'}{K_a} = \rho \sum_{i=1}^{m} \sigma_i \Leftrightarrow pK_a' = pK_a + \rho \sum_{i=1}^{m} \sigma_i,$$

(12)

where $pK_a'$ is the dissociation constant of the parent (unsubstituted) molecule, $\rho$ is a constant specific for the modeled class of molecules, $m$ is the number of substituent positions, and the $\sigma_i$ are constants expressing the substituent effect on the dissociation constant. A disadvantage of this approach is that the $\sigma$ constants have to be known (experimentally determined) for all involved substituents.

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

$n =$ 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, $MP =$ multiproticity (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 Moller-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], RBNN = radial basis function neural network [122, 123], RMI = Recife model [124], SVM, LS-SVM = support vector machine [127, 128].
[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 (pK_a≤6.5), marginal (6.5<pK_a<7.8), and inactive (pK_a≥7.8). Based on this, 22 biophores were identified that were used to predict a test set of 192 organic acids [208] with R = 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 (pK_a) were tabulated. Any molecular structure p appended to the center was considered a perturber. The pK_a of the center was calculated as pK_a = (pK_a) + δp(pK_a), where δp(pK_a) 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 microioniztion 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
Predicting the $pK_a$ of Small Molecules

Combinatorial Chemistry & High Throughput Screening, 2011, Vol. 14, No. 5

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 $\pi$-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 carboxylic acids and acid nitro and nitroso groups) 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 $Q_{\sigma^*}$ 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, $\pi$-electronegativity of the R-carbon atom in an acid, and, an indicator variable for $\alpha$-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 ($Q_{\sigma^*}$, $\sigma$-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 $pK_a$ 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/pKa 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/pKa. 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/p $pK_a$, 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

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

*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

<table>
<thead>
<tr>
<th>RMSE</th>
<th>[212]</th>
<th>[173]</th>
<th>[215]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>ACD/pKα</td>
<td>0.35</td>
<td>0.22</td>
<td>0.54</td>
</tr>
<tr>
<td>Marvin</td>
<td>0.48</td>
<td>0.32</td>
<td>0.51</td>
</tr>
</tbody>
</table>

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/ pKa 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 pKa 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 pKa 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 pKa 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 pKa 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 (r²), 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 pKa 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 pKa 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:

\[
\log_{10}(K_{\text{pp}}) = \log_{10} \frac{c_{\text{aq}}(H^+)c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)}.
\]

\[
= \log_{10} K_D(\text{HA}) + \log_{10} \frac{c_{\text{aq}}(HA) + c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA) + c_{\text{aq}}(A^-)} - \log_{10} \frac{c_{\text{aq}}(HA) + c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)}
\]

\[
= \log_{10} K_D(\text{HA}) - \log_{10} \frac{c_{\text{aq}}(HA) + c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)}.
\]

\[
= \log_{10} K_D(\text{HA}) - pH + \log_{10} \frac{c_{\text{aq}}(H^+)c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)}.
\]

\[
= \log_{10} K_D(\text{HA}) - pH + \log_{10} \left( \frac{c_{\text{aq}}(H^+)c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)} + \frac{c_{\text{aq}}(H^+)c_{\text{aq}}(A^-)}{c_{\text{aq}}(HA)} \right)
\]

\[
= \log_{10} K_D(\text{HA}) - pH + \log_{10} (H + K_a).
\]
Derivation of Equation 11:

\[
\log_{10}(K_{D}^{app}) = \log_{10} K_{D}(HA) - \text{pH} - \log_{10}(H + K_{a})
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}}) = \text{pH} + \log_{10}(H + K_{a})
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}}) = -\log_{10} \frac{c_{aq}(H_{3}O^{+})}{c^{\leftrightarrow}} + \log_{10} \frac{c_{aq}(H_{3}O^{+}) + K_{a}c^{\leftrightarrow}}{c^{\leftrightarrow}c_{aq}(H_{2}O^{+})}
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}}) = \log_{10}(1 + \frac{K_{a}c^{\leftrightarrow}}{c_{aq}(H_{3}O^{+})})
\]

\[
\frac{K_{D}(HA)}{K_{D}^{app}} = 1 + \frac{K_{a}c^{\leftrightarrow}}{c_{aq}(H_{3}O^{+})}
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}} - 1) = \log_{10} \frac{K_{a}c^{\leftrightarrow}}{c_{aq}(H_{3}O^{+})}
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}} - 1) = \log_{10} K_{a} - \log_{10} \frac{c_{aq}(H_{3}O^{+})}{c^{\leftrightarrow}}
\]

\[
\log_{10}(\frac{K_{D}(HA)}{K_{D}^{app}} - 1) = \text{pH} - \text{p}K_{a}
\]

REFERENCES


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).


Predicting the pKa of Small Molecules

Combinatorial Chemistry & High Throughput Screening, 2011, Vol. 14, No. 5


Ertl, P. Simple quantum chemical parameters as an alternative to
Müller, P. Glossary of terms used in physical organic chemistry.
Kim, K.H.; Martin, Y.C. Direct prediction of dissociation constants
Potter, M.J.; Gilson, M.K.; McCammon, J.A. Small molecule p
Roy, K.; Popelier, P. Predictive QSPR modeling of the acidic
dissociation constant (pKa) of phenols in different solvents. J.
Gielecka, R.; Polanski, J. Modeling robust QSAR. 2. Iterative
variable elimination schemes for COMSA: Application for modeling benzoic acid pKa values. J. Chem. Inf. Model., 2007,
47(2), 547-556.
Kim, K.H.; Martin, Y.C. Direct prediction of linear free energy
substituent effects from 3D structures using comparative molecular
Kim, K.H.; Martin, Y.C. Direct prediction of dissociation constants
(pKa) 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.
Gargallo, R.; Sotri,
Jover, J.; Bosque, R.; Sales, J. Neural network based QSPR study for predicting pKa of phenols in different solvents. QSQR Comb.
Bull. 2007(263), 385-397.
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.,
Goodarzi, M.; Freitas, M.P.; Wu, C.H.; Duchowicz, P.R. pK,
modeling and prediction of a series of pH indicators through genetic algorithm-least square support vector regression. Chemomut.
48(10), 2042-2053.
Slov., 2007, 54, 605-610.
Potter, M.J.; Gilson, M.K.; McCammon, J.A. Small molecule pK,
prediction with continuum electrostatics calculations. J. Am. Chem.
Müller, P. Glossary of terms used in physical organic chemistry.
Ertl, P. Simple quantum chemical parameters as an alternative to