doi:10.2533/chimia.2006.691 CHIMIA 2006, 60, No. 10

Chimia 60 (2006) 691–698
© Schweizerische Chemische Gesellschaft
ISSN 0009–4293

Prediction of Physicochemical Properties of Organic Compounds from 2D Molecular Structure – Fragment Methods vs. LFER Models

Gerrit Schüürmann*[a,b], Ralf-Uwe Ebert[a], and Ralph Kühne[a]

Abstract: A large number of models is available to predict physicochemical properties directly from the two-dimensional molecular structure. An alternative to conventional fragment methods is given by linear free-energy relationships (LFERs) employing Abraham parameters. The latter have a solid mechanistic background, but a drawback in practice is the limited availability of Abraham parameters for substances of interest. As a consequence, more complex compounds typically require the estimation of Abraham parameters from the chemical structure. Comparative analysis of prediction methods for Henry's law constant and sorption to soil organic matter shows that at present, fragment methods are superior to the LFER approach when employing calculated Abraham parameters. For the subset of typically more simple compounds with experimental Abraham parameters, the respective LFERs are competitive to general-purpose fragment models. The discussion includes analyses of compound subsets without and with hydrogen bond functionalities, and of the impact of structural complexity on the model performance.

Keywords: Abraham model · Fragment method · Henry's law constant · Hydrogen bonding · REACH · Soil sorption coefficient

Introduction

In the context of the upcoming European REACH regulation of industrial chemicals, physicochemical properties are required for two major purposes. First, they directly drive environmental partitioning and degradation, and thus are key parameters to conduct environmental fate and risk modelling. Second, physicochemical parameters affect the bioavailability of toxicants, which in turn is important for *in vitro* assays and other laboratory testing, and also for evaluating the risk associated with contaminants in the field.

While there is already a large number of methods to predict physicochemical properties [1], models targeted for REACH applications have to be evaluated thoroughly according to the OECD principles for structure–activity relationships [2]. In the present investigation, we focus on two prominent classes of methods to predict partition properties of organic compounds. For the latter, we take Henry's law constant and sorption to soil organic matter as examples, both of which have been subject to recent reviews on state-of-the-art estimation models [3][4].

One model type is the well-established fragment methods where the target value is generated through analysis of the two-dimensional (2D) molecular structure. Here, empirically derived increment tables are applied to calculate the value of interest. An alternative approach is given by the Abraham equation [5][6], which is a special form of a linear free energy relationship (LFER) derived from theoretical considerations. In this case, the model includes compound- and matrix-specific information about fundamental types of intermolecular interaction, and as such forms a mechanistically sound approach to predict the partition behaviour of organic compounds from only few input parameters. In principle, the Abraham equation relates experimental bulk properties to experimental molecule properties. To achieve predictions from structural information only, the Abraham descriptors in turn need to be calculated from the chemical structure. In this study, the performance of both approaches is comparatively analyzed with large data sets, with a particular focus on the impact of molecular characteristics such as polarity and hydrogen bonding. The results show that for more complex structures where experimental Abraham parameters are not available, the prediction capability of the Abraham equation is presently inferior to the one of general-purpose fragment methods.

*Correspondence: Prof. Dr. G. Schüürmann[a,b]
Tel.: +49 341 235 2309
Fax: +49 341 235 2401
E-Mail: gerrit.schuurmann@ufz.de
[a]Department of Ecological Chemistry
UFZ Centre for Environmental Research
Permoserstrasse 15
D-04318 Leipzig
[b]Institute for Organic Chemistry
Technical University Bergakademie Freiberg
Leipziger Strasse 29
D-09596 Freiberg

Theoretical Background

Henry's Law Constant

Henry's law constant H characterizes the equilibrium partitioning of compounds

between air and water. In its dimensionless form (K_{aw}), it can be defined as the ratio of the compound concentrations in air, c_a, and water, c_w:

$$K_{aw} = \frac{c_a}{c_w} \quad (1)$$

For the vapour phase in equilibrium with the aqueous solution, application of the ideal gas law to the compound of interest

$$P = \frac{n}{V} RT = c_a RT \quad (2)$$

(with P = partial pressure, n = number of moles, V = volume, R = gas constant, T = absolute temperature) leads to

$$H = K_{aw} RT = \frac{P}{c_w} \quad (3)$$

Eqn. (3) shows that under equilibrium conditions, the truly dissolved amount of a compound in aqueous solution (c_w) is directly related to its concentration in the gas phase above the solution, the latter of which is expressed as partial pressure (P). In particular, Henry's law states that the ratio P/c_w is constant for varying compound concentrations c_w.

If c_w has reached the water solubility S_w, further addition of the compound would form a separate phase, the pure compound phase. Because the dissolved state would be in equilibrium with both the gas phase and the pure compound phase, the latter would also be in equilibrium with the gas phase. Consequently, in the case of $c_w = S_w$ the partial pressure in the gas phase, P, equals the vapour pressure P_v of the pure compound. It follows that Henry's law constant can also be written as

$$H = \frac{P_v}{S_w} \quad (4)$$

A more detailed analysis reveals that Eqn. (4) is in fact an approximate relationship, based on the assumptions that the vapour phase behaves according to the ideal gas law, that the solute concentration is sufficiently small for the solution to be ideal (which may be in conflict with $c_w = S_w$), and that the separate solute phase (that would be formed for compound amounts in solution above S_w) is in fact a pure compound phase that does not contain water. Nevertheless, Eqn. (4) is often used to estimate H if experimental values are unavailable (but preferably with experimental values for P_v and S_w). Models to predict Henry's law constant from molecular structure usually use the decadic logarithm of its dimensionless form, log K_{aw}, as target property; experimental H values can be converted into K_{aw} through the left part of Eqn. (3).

Soil/Water Partition Coefficient Normalized to Organic Carbon Content

The association of waterborne compounds with a solid phase (the sorbent) is called sorption. It covers both adsorption at the surface and absorption into the volume of the solid material. Due to the higher order achieved through fixation of the compound (the sorbate), entropy decreases with increasing sorption except if the molecule–sorbent association is accompanied by dissolution of sorbent components or by sorbate dissociation. Accordingly, spontaneous sorption is driven by a negative (exothermic) sorption enthalpy.

Because the sorbent is usually quantified through its mass (without specification of its molecular weight), the distribution coefficient K_d that quantifies the amount of sorption is expressed in linear form as

$$K_d = \frac{X_s}{c_w} \quad (5)$$

where X_s = number of sorbate moles per kg sorbent, and c_w = compound concentration in aqueous solution. Consequently, K_d has the unit L/kg. Normalization to the organic carbon content of the sorbent, f_{oc} (typically around 0.01–0.03 for soils and sediments), leads to

$$K_{oc} = \frac{1}{f_{oc}} K_d \quad (6)$$

Eqn. (6) assumes that sorption takes place only at the organic carbon fraction of the sorbate, which is a reasonable approximation for non-ionic organic compounds. Note further that according to a more elaborate analysis, sorption would be treated as a two-phase process with a non-linear component that is neglected in Eqns. (5) and (6) [7]. Regression models to predict the sorption of organic compounds into soil organic matter are usually calibrated to log K_{oc}, assuming that the numerical value of the unit L/kg can be taken as unity (and thus neglected).

Fragment Models

Fragment methods in principle work by counting the number of occurrences of molecular substructures (either atom or bond groups), and by adding up increment values for these fragments. An important rule is that such schemes can be applied only to those compounds that can be decomposed completely in terms of the model fragments F_i, and that each atom of the molecule must belong to exactly one fragment.

While simple fragment schemes cannot account properly for steric and electronic interactions between different functional groups, more complex methods include so-called correction factors C_j for this purpose. In contrast to fragments as primary model parameters, several correction factors may be applied to the same atom (if the method contains respective features). Consequently, a given atom or atom group could belong to one fragment and at the same time to one or several correction factors.

Frequently, fragment methods also contain indicator variables I_k that are to be used once per molecule in particular structural constellations. Taking the subgroup of nonpolar and weakly polar compounds discussed below as an example the presence or absence of the respective structural condition can be accomplished by the numerical values 0 and 1 of a corresponding indicator variable.

For a property to be modelled by a fragment method, the general model equation thus can be written as follows:

$$\log \text{Property} = \sum_i a_i F_i + \sum_j b_j C_j + \sum_k c_k I_k + d \quad (7)$$

where a_i, b_j and c_k are the regression coefficients associated with the fragments F_i, correction factors C_j and indicator variables I_k, and d is the intercept. While F_i and C_j represent the number of occurrences of the respective fragment or correction (which may be zero if the relevant substructure is missing in the compound of interest), I_k is never greater than 1 (even if the structural condition occurs more than once in a given molecule).

Abraham Equation

According to Abraham *et al.* [5][6], the basic LFER to model the decadic logarithm of any partition coefficient K is

$$\log K = e E + s S + a A + b B + v V + c \quad (8)$$

with compound descriptors E, S, A, B, V, respective phase descriptors e, s, a, b, v, and the regression constant c. V is a characteristic volume term defined by Abraham and McGowan [8], and can be calculated by a simple fragment scheme presented there. Theoretically, the use of V is only valid for liquid/liquid systems, and should be replaced by a solubility term L in case of gas/liquid partitioning such as for air/water partition coefficients. In practice, both versions exist for K_{aw}, and the L version does not perform better.

The other compound descriptors are experimental values. A und B denote hydrogen bond acidity (donor strength) and basicity (acceptor strength), E is the excess molar refraction, and S accounts for dipolarity and polarisability. For B, there is a slight difference between B^H for dry systems (*e.g.* pure octanol in case of octanol/air partitioning) and B^O for wet systems (*e.g.*

water-saturated octanol in case of octanol/water partitioning). To turn the Abraham approach into a real prediction model for compounds not (yet) available physically, these descriptors have to be calculated from the chemical structure. At this stage, the only general-purpose model available is a fragment method published by Platts *et al.* [9], now offered as a module in commercial software [10].

Materials and Methods

Logarithmic Partition Coefficient Air/Water log K_{aw} at 25 °C

A validated data set of log K_{aw} values for 2070 organic compounds covering many important organic compound classes and a data range from –16.5 to 3.1 (average: –3.7) has been taken from our in-house database [11]. The data originate from several literature sources, and will be published in a forthcoming study. Part of the experimental values resulted from direct measurements, and others were obtained indirectly from ratios of measured vapour pressures and water solubilities. In a few cases, K_{aw} was calculated from other experimental partition coefficients (*e.g.* $K_{aw} \approx K_{ow}/K_{oa}$, with K_{ow} and K_{oa} being the octanol/water and octanol/air partition coefficient, respectively). Moreover, some data have been interpolated to 25 °C from temperature series within the range of 20–40 °C.

The chemical domain of the data set can be characterized as follows. There are 551 compounds consisting only of C, H and partly halogen, 598 compounds with oxygen as only additional heteroatom, and 921 substances with other heteroatoms. 822 compounds have no or only one functional group, 384 compounds have a multiple occurrence of a single functional group, 582 chemicals contain two different functional groups, and 339 substances have more complex chemical structures.

With regard to the intermolecular interaction profile, 564 compounds (hydrocarbons, halogenated hydrocarbons, organic Si compounds without or with oxygen, monofunctional aromatic alcohols and amines without other heteroatoms) can be characterized as nonpolar or weakly polar. The underlying rationale is as follows: In an early study on modelling log K_{oc}, 72 compounds consisting of hydrocarbons and respective monofunctional compounds (except for the Si compounds) could be combined into one simple linear regression equation based on molecular connectivity [12], although one might argue that the compound classes mentioned differ in their ability for hydrogen bonding. Among the remaining 1431 more polar compounds there are 542 substances with strong hydrogen bond donor sites (OH, NH_2, NH, SH), and 889 chemicals with strong hydrogen bond acceptor sites (O, –S–, =S, N except as ≡N).

Logarithmic Soil Sorption Coefficient log K_{oc}

Experimental log K_{oc} data for 571 compounds have been taken primarily from several studies compiled by Huuskonen [13], augmented by some updates from Nguyen *et al.* [14]. The log K_{oc} data range from 0.0–6.5, with an average of 2.7. A detailed description of the data set and the validation procedure is given in [15].

With regard to the chemical domain, 48 compounds are simple hydrocarbons or halogen hydrocarbons, 122 compounds contain O (possibly in addition to halogen), and 401 compounds contain other heteroatoms (possibly in addition to halogen or oxygen or both). The structural complexity can be described as follows: There are 156 substances with a maximum of one functional group, 60 with multiple occurrence of a single group, 233 with two different groups, and 122 more complex chemicals. With respect to polarity and hydrogen bonding, there are 81 nonpolar or weakly polar chemicals (defined as described above), and 287 molecules with strong hydrogen bond donor sites. Of the remaining 203 compounds without H donor sites, 201 compounds contain strong hydrogen bond acceptor groups.

Fragment Methods

For log K_{aw}, the bond fragment model of Meylan and Howard [16] in the computerized version [17] is known as one of the most popular and best performing methods [3]. It consists of 124 bond fragments (that correspond to two-atom fragments with each bond to be counted exactly once), 36 correction factors, and 3 indicators. In this study, the original software [17] was used to perform the calculation. For comparison, the ALOGS method of Viswanadhan *et al.* [18] has been included as implemented in our ChemProp software [11]. ALOGS contains 68 fragments but no correction factors, and as such is expected to perform less well with more complex compounds.

For log K_{oc}, the model of Tao *et al.* [19] is the only general-purpose pure fragment model currently available. It consists of 74 fragments, 23 correction factors and 1 indicator variable. Even though the publication contains obvious inconsistencies as outlined below, it was possible to implement a correspondingly modified version in our in-house software system ChemProp [11] to run the calculations automatically. The second method selected was PCKOCWIN [20][21], which is based on the first-order connectivity index, $^1\chi$, and 21 correction factors (C_i in Eqn. (7)) and 6 indicator variables (I_k in Eqn. (7)), and as such is not a pure fragment method. Instead, it is a combination of fragments and molecular connectivity, with all basic fragments F_i being replaced by $^1\chi$.

Abraham Methods

The model equation used for log K_{aw} was developed by Abraham *et al.* [22]

$$\log K_{aw} = -0.577\,E - 2.549\,S - 3.813\,A - 4.841\,B^H + 0.869\,V + 0.994 \quad (9)$$

(as already noted, this model is superior to the alternative equation applying L instead of V, despite the fact that using L to predict K_{aw} is preferred from theoretical considerations). For log K_{oc}, the model from Poole and Poole [23] was included in our comparative analysis,

$$\log K_{oc} = 0.74\,E - 0.31\,A - 2.27\,B^o + 2.09\,V + 0.21 \quad (10)$$

which performed slightly better for the present data set than two more recent models [14].

With both the log K_{aw} and log K_{oc} data sets, experimental Abraham parameters are available for only small portions of the compounds. Consequently, calculated Abraham parameters have been used for the comparative analysis of the full data sets, employing the methods of Abraham and McGowan [8] for V, and of Platts *et al.* [9] for E, S, A and B. Note, however, that the implementation of the Platts methods was not trivial due to some apparent ambiguities, as is discussed in more detail below. In addition, the statistical performance of all methods was also evaluated with the subsets of compounds where experimental Abraham parameters were available.

Statistical Evaluation

The method performances are characterized in terms of the number of formally valid results, the predictive squared correlation coefficient q^2 that accounts also for systematic errors (for an explanation, see *e.g.* [24]), the conventional squared correlation coefficient r^2 (that automatically corrects for systematic errors), the bias that denotes the absolute systematic error (referring to q^2), and the predictive standard error *se*.

Results and Discussion

ChemProp Implementation of Tao Model

The Tao model consists of 74 fragment values and 24 correction factors, and as such is prepared to handle also more complex chemical structures. Surprisingly, the model as published [19] cannot be applied to all of its training set compounds due to missing

fragments. Examples include cyanazine (**1** in the Scheme) where there is no fragment for the C≡N group attached to aliphatic C, and urea (no. **2**) that cannot be decomposed completely when using the method fragments available. Interestingly enough, calculated log K_{oc} values of 2.62 and 1.27 were reported for these compounds, but we could not reproduce these and other calculations as well as the published statistics that, as reported, should refer to their total set of 592 compounds including 7 chemicals contained twice (r^2 = 0.967, average error = 0.316 [19]). Table A1 of the appendix (given as supplementary material) lists a slightly modified version of the model as implemented in ChemProp [11] where identified ambiguities were removed, but which is still not applicable to 25 compounds of the original training set due to missing fragments. With this ChemProp version of the Tao model, the calibration statistics for the remaining 560 compounds of their original set (560 = 592 – 7 – 25) and their original log K_{oc} values are r^2 (squared correlation coefficient) = 0.816, *se* (standard error) = 0.533, and bias = –0.01.

ChemProp Implementation of Platts Methods

The published version [9] lists a common table of 82 fragments and correction factors for *E*, *S*, B^H, and B^O, differing only in the respective increment values, and another scheme for *A* with 51 parameters, both including an intercept. From the viewpoint of Eqn. (7), the *A* prediction model is a pure correction factor model, because there are no basic fragments according to which the molecule has to be fully decomposed, and because a given atom may (in principle) belong to more than one method-relevant substructural feature. By contrast, the first 42 parameters of the prediction method for *E*, *S*, B^H, and B^O can be regarded as a classical fragments (mostly in the particularly simple form of atom contributions). The remaining multiple-atom contributions represent correction factors except for the intercept and parameter #58 (steroid structure), the latter of which is to be applied only once per molecule if appropriate, and thus serves as an indicator variable.

With regard to the chemical domain, there is no restriction for the *A* prediction model (except that it would formally yield 0 for compounds that do not contain any of the hydrogen bond donor groups listed as correction factors), and the other model could formally be applied to all compounds that consist of atom types covered by the first 42 fragments. However, while the *V* prediction method of Abraham and McGowan [8] is clearly described without any particular issues concerning implementation and use, this was not the case for the Platts methods. Unfortunately, the few example calculations given for phenol derivatives are not explained in detail, and some other results can obviously not be obtained from the published fragment table. This also holds true for a number of calculations given in a subsequent paper [25], and even reference calculations by means of a demo version of the computerized model [10] cannot help to clarify the apparent inconsistencies of the method description.

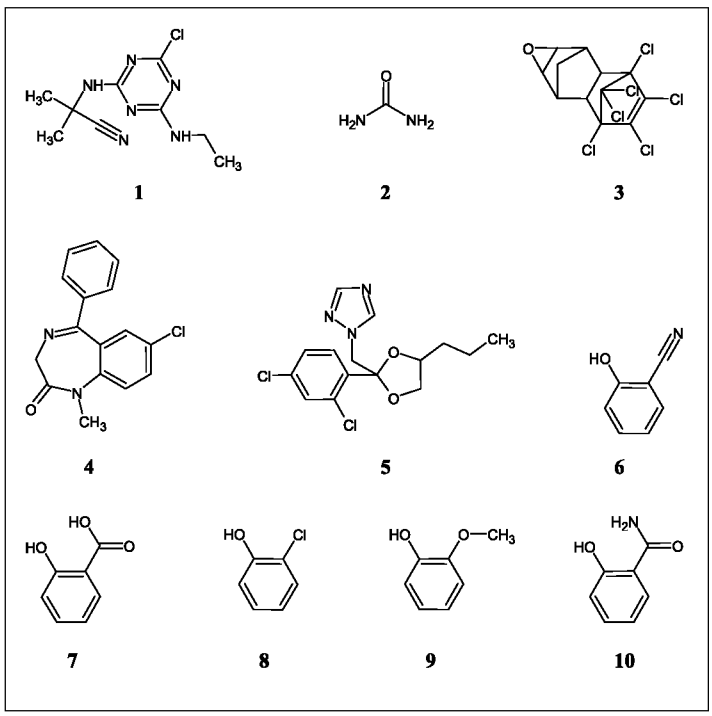
The following examples for B^O illustrate the problem. For endrin (compound **3** in the Scheme), both manual application of the scheme and the commercial software yield the same results of 0.9, while the value given in [25] is 0.7. On the other hand, for diazepam (no. **4**) the manual result of 1.13 is almost the same as in [25] (1.15), while the software calculates 1.31. For propiconazole (no. **5**), there are three different results (manual application: 1.29, published value: 1.32, commercial software: 1.43).

Moreover, the published calculation result *A* = 0.74 for 2-cyanophenol (**6**) implies application of parameter #47 of the prediction method (2-CX substitution at phenol, with X = halogen, NO_2, CN or CF_3), although introduction of X = CN would lead to the wrong substituent CCN. With salicylic acid (**7**), the published calculation result requires consideration of both the carboxylic fragment (#9) and the OH group attached to aliphatic carbon (#1) for the COOH substructure, which appears at least unusual.

With regard to the method description for predicting *E*, *S*, B^H, and B^O, another inconsistency can be illustrated with *o*-chlorophenol (**8**): Here, correction factor YccY (parameter #74, Y = any heteroatom, c = aromatic carbon) must not be applied to reproduce the calculation result, perhaps because the respective structural constellation is already covered by parameter #67 (H bond 9). Interestingly, the corresponding restriction to only one of two possible correction factors does not apply for *o*-methoxyphenol (**9**) and several other *o*-substituted phenols, where both the YccY correction and the intramolecular H bond (parameter #61, H bond 3) must be used to reproduce the results.

Concerning the nine H-bond correction factors, in the case of ambiguous opportunities only one correction factor has to be used, in the order of preference #59, #60, #63, #61, #62, #64 ... #67. In the case of non-aromatic H-bond corrections, these factors must not be applied to non-branched chains, but only to branched groups or ring members. The same holds true for the H-bond corrections in the A prediction scheme, with an additional preference of the H-bond correction factor 10 (#37 in that scheme).

Note further that the aromatic amide functionality (parameter #48) is not uniquely defined, because it is left open whether the carbon or nitrogen side or both should be attached to aromatic substituents. Interestingly, *o*-hydroxybenzamide (**10**) is not treated as aromatic amide, but as noncyclic aliphatic amide (parameter #49). Finally, there appear to be also some inconsistencies with the treatment of the basic fragments



Scheme. Chemical structures of compounds discussed in the text: **1** = cyanazine, **2** = urea, **3** = endrin, **4** = diazepam, **5** = propiconazole, **6** = *o*-cyanophenol, **7** = salicylic acid, **8** = *o*-chlorophenol, **9** = *o*-methoxyphenol, **10** = *o*-hydroxybenzamide

#1–42: Taking the cyano group CN as an example, it is covered by parameters #21 and #22, and nevertheless the triply bound carbon atom (but not the triply bound nitrogen) has to be considered separately (parameter #9). For amines, there are different fragments for non-aromatic and aromatic attachment. In case of both attachments for the same group, the aromatic fragment applies. In Tables A2–A4 of the appendix (see supplementary material), the correspondingly modified (or specified) versions of the Platts prediction models for the Abraham parameters A as well as E, S, B^H, and B^O are listed as implemented in ChemProp [11] and used for our comparative analysis.

Comparative Statistics for Henry's Law Constant

The statistical performance of the three methods to predict log K_{aw} from molecular structure is summarized in Table 1. While the fragment scheme of Meylan and Howard [16][17] can be applied to all compounds except dimethyl diselenide ($CH_3Se-SeCH_3$) and achieves the best statistics (q^2 = 0.839), the chemical domain of the ALOGS fragment method is much smaller, covering only 1300 of the 2070 compounds. Clearly, the overall result of the Abraham approach is rather poor when employing calculated descriptors (q^2 = 0.179), yielding a systematic underestimation of Henry's law constant by a factor of *ca.* 10 (bias = –1.048) and a standard error of more than 2.5 orders of magnitude. In Fig. 1, the plots of calculated *vs.* experimental value and calculation error *vs.* experimental value are shown for both the Meylan and Howard method (top) and the Abraham equation employing calculated input parameters (bottom).

Analysis of the separate statistics for the four compound subsets defined according to their general intermolecular interaction characteristics reveals distinct features of the method performances. All three methods yield best results for the group of nonpolar and weakly polar compounds, while significantly inferior statistics are observed for compounds with stronger hydrogen bond donor or acceptor functionalities. For these two subsets, even the overall best fragment model by Meylan and Howard performs only moderately, and yields systematic underestimations of Henry's law constant for both H bond donors and acceptors. This finding suggests that the latter two subgroups of compounds have, on the average, more complex structures and are thus more difficult to handle for prediction methods.

The substantial negative biases point to a more specific difficulty associated with compounds that possess hydrogen bond donor or acceptor sites. The underestimation of Henry's law constant corresponds to an overestimation of the free energy gained

Table 1. Statistics for the estimation of log K_{aw}.

Model	No. compounds	q^2	*bias*	*se*
Total data set				
Meylan and Howard [16][17]	2069	0.839	–0.197	1.182
ALOGS [18]	1300	0.716	–0.452	1.201
Abraham approach [22]	2041	0.179	–1.048	2.638
Subsets with respect to polarity and hydrogen bonding				
Subset of nonpolar or weakly polar compounds				
Meylan and Howard	564	0.923	+0.072	0.512
ALOGS	519	0.853	–0.340	0.692
Abraham approach	561	0.696	–0.169	1.010
Subset with strong H bond donor sites				
Meylan & Howard	542	0.690	–0.377	1.492
ALOGS	250	0.249	–0.697	1.757
Abraham approach	540	–0.529	–1.445	3.312
Subset with H bond acceptor sites but without H bond donor sites				
Meylan & Howard	889	0.655	–0.287	1.296
ALOGS	478	0.381	–0.442	1.296
Abraham approach	875	–0.893	–1.411	2.945
Subset of polar compounds without strong H bonding sites				
Meylan & Howard	74	0.899	+0.154	0.794
ALOGS	53	0.743	–0.488	1.037
Abraham approach	65	0.555	–0.440	1.454

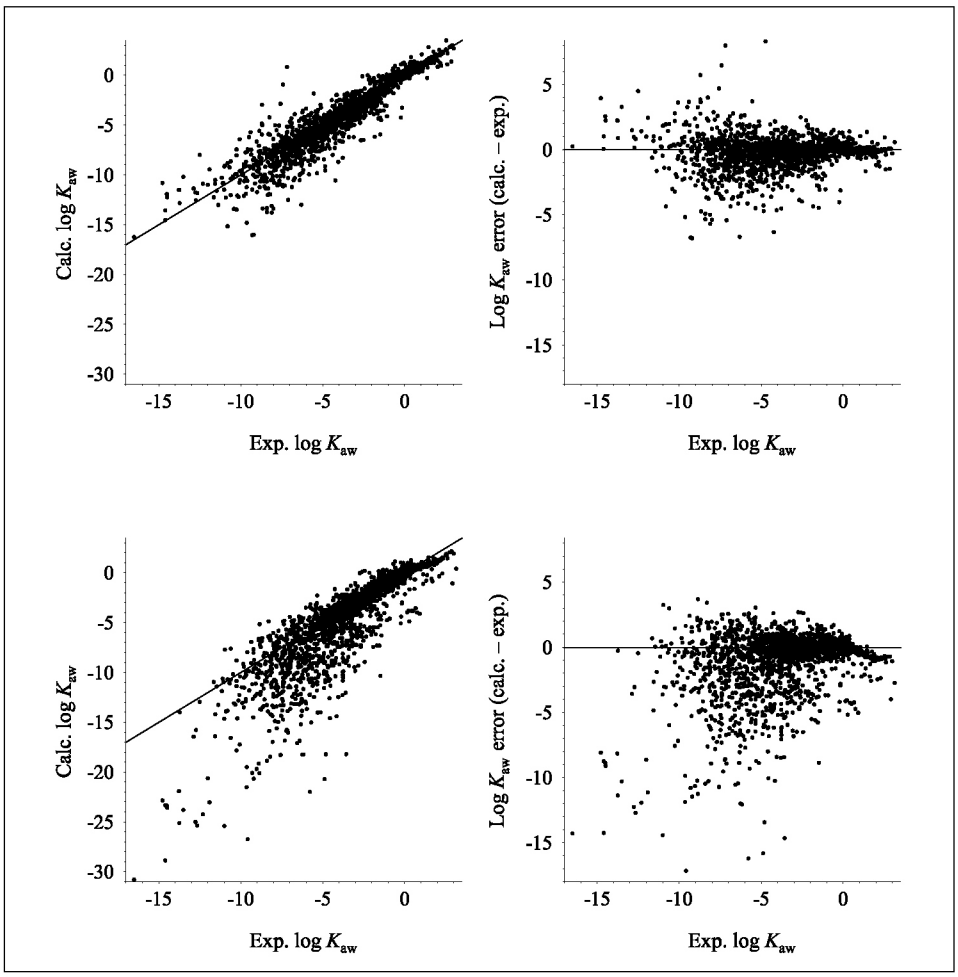


Fig. 1. Log K_{aw} calculation (left) and calculation errors (right) *vs.* experimental value for Meylan and Howard [16][17] (top) and for the Abraham approach [9][22] (bottom)

through stabilizing hydrogen bond interaction of the solute with water. Because most solutes of the current data set do not provide opportunities for strong intramolecular hydrogen bonds in competition with corresponding solute–solvent interactions, it appears as if the calculation schemes simply overestimate, on average, their stabilizing interaction with water. As will be shown below, this trend holds true for more complex structures, while for structurally simple compounds with H bond donors and acceptors there is no corresponding bias.

The statistical differences between compounds without and with hydrogen bonding functionalities are particularly striking for the Abraham method (when using calculated parameters). Here, the negative q^2 values for hydrogen bond active compounds (–0.529 and –0.893) indicate that on average, the model prediction is even inferior to taking the subset-specific arithmetic mean as (constant) predicted value for all compounds of the respective subset. Moreover, the unusually bad performance for compounds with H bond donor or acceptor sites indicates that the quality of the presently available increment methods to predict A and B from molecular structure [16–18] is significantly inferior to the ones for other Abraham parameters.

The surprisingly poor results achieved with the Abraham model suggest that this is mainly due to the limited quality of the methods used to predict the input parameters from molecular structure. To address this hypothesis, the subset of all compounds with available experimental Abraham parameters was formed, now including also the Abraham model employing those parameters in the comparative analysis. Note that for this subset, the log K_{aw} data range has decreased from –16.5…+3.1 to –8.8…+3.0.

As can be seen from Table 2, all methods perform significantly better for the subset than for the total compound set. The Abraham model employing experimental parameters is indeed clearly superior to the one using calculated parameters (q^2 of 0.942 *vs.* 0.864), but still slightly inferior to the Meylan and Howard fragment method. However, its bias is the second greatest among the four methods tested for this subset. Again, ALOGS has the smallest application range (677 of the 732 compounds), but the relative portion is much greater than for the total set (96% *vs.* 63%). It reflects the fact that so far, compounds with experimental Abraham parameters are typically relatively simple chemical structures. Indeed, 437 of the 732 compounds have no or one functional group, and multiple occurrence of a single functional group applies for 186 compounds. Only 66 chemicals have two different functionalities, and there is only one compound with more than two functional groups. It follows that for this subset, the generally good performances of the prediction methods are at least partly due to the less demanding chemical structures.

Table 2. Statistics for the estimation of log K_{aw} – subset with available experimental Abraham descriptors

Model	No. compounds	q^2	*bias*	*se*
Total data set				
Meylan and Howard [16][17]	732	0.960	–0.001	0.394
ALOGS [18]	677	0.878	–0.238	0.666
Abraham approach [22]				
experimental descriptors	732	0.942	–0.166	0.476
estimated descriptors [9]	731	0.864	–0.047	0.726
Subsets with respect to polarity and hydrogen bonding				
Subset of nonpolar or weakly polar compounds				
Meylan and Howard	396	0.965	+0.025	0.354
ALOGS	380	0.881	–0.386	0.638
Abraham approach				
experimental descriptors	396	0.910	–0.266	0.565
estimated descriptors	395	0.869	–0.166	0.681
Subset with strong H bond donor sites				
Meylan and Howard	131	0.914	–0.024	0.497
ALOGS	113	0.715	–0.084	0.886
Abraham approach				
experimental descriptors	131	0.955	–0.025	0.361
estimated descriptors	131	0.780	+0.081	0.797
Subset with H bond acceptor sites but without H bond donor sites				
Meylan and Howard	188	0.894	–0.048	0.401
ALOGS	169	0.731	–0.002	0.570
Abraham approach				
experimental descriptors	188	0.924	–0.073	0.340
estimated descriptors	188	0.666	+0.133	0.710
Subset of polar compounds without strong H bonding sites				
Meylan and Howard	17	0.967	+0.087	0.304
ALOGS	15	0.930	–0.291	0.459
Abraham approach				
experimental descriptors	17	0.989	+0.021	0.178
estimated descriptors	17	0.425	–0.279	1.268

Decomposition into the four subsets of nonpolar and weakly polar compounds, H bond donors, H bond acceptors and polar compounds without H bond sites leads to the statistics summarized in the lower part of Table 2. Now, the Abraham model employing experimental parameters outperforms all other methods for all subsets except the group of nonpolar and weakly polar compounds. Here, detailed inspection reveals that this subset of 396 compounds contains a notable number of larger PAHs and PCBs, while such compounds were missing in the original 408-compound training set used to derive the Abraham equation. It shows how the chemical domain of the training set of a regression model affects its application range.

Coming back to H bond donor and acceptor compounds, now both the Meylan and Howard fragment method and the Abraham model (when employing experimental input parameters) do not yield substantial biases, which contrasts with findings achieved for the respective subsets of the total compound set (compare Tables 1 and

2). This discrepancy probably reflects the fact that on the average, compounds with experimental Abraham parameters have relatively simple structures, and so are less demanding for structure–activity models as noted above.

Comparative Statistics for the Sorption Constant Normalized to Organic Carbon

As can be seen from Table 3, the application range is significantly smaller for the Tao model (508 compounds) than for both PCKOCWIN and the Abraham approach (if based on calculated input parameters) that can handle all 571 compounds. Apart from that, the Abraham statistics are again inferior to the ones achieved with the fragment models. For the PCKOCWIN and the Abraham model, Fig. 2 contains the data distributions for the total log K_{oc} set.

As with Henry's law constant, the log K_{oc} prediction is significantly inferior for compounds with hydrogen bonding functionalities, and particularly poor for the Abraham equation. This time, however, the biases for the respective subsets are relatively small for the Tao model, while PCKOCWIN and the Abraham equation systematically overestimate the log K_{oc} of compounds with hydrogen bond acceptor sites. Note that this latter subset contains also the only two more polar compounds without hydrogen bond functionalities.

The performance statistics for the log K_{oc} subset of compounds with experimental Abraham parameters are summarized in Table 4. Again, all methods perform significantly better than for the total compound set, which reflects the fact that the subset of 107 compounds consists of generally more simple chemical structures. The Abraham model based on experimental input parameters outperforms all other methods, and now the respective model employing calculated parameters also yields very good results. Note that among the 107 compounds, 62 have indeed very simple chemical structures with at most one functional group, 20 have multiple occurrences of a single functional group, and only 25 have two different functional groups (with no compound having more than two different functional groups).

As compared to Henry's law constant, the prediction capability is lower for log K_{oc} for both the fragment methods and the LFER approach. Possible reasons include the generally lower data quality for K_{oc}, and the considerable complexity of soil organic matter and the associated sorption process. Note, however, that a recently developed fragment model using the present data set yields $r^2 = 0.852$ and $se = 0.469$, and thus outperforms both the currently investigated models as well as other existing prediction methods [15].

Table 3. Statistics for the estimation of log K_{oc}.

Model	No. compounds	q^2	*bias*	*se*
Total data set				
Tao *et al.* [19]	508	0.788	+0.014	0.552
Meylan *et al.* [20][21]	571	0.653	+0.104	0.719
Abraham approach [23]	571	0.581	+0.012	0.790
Subsets with respect to polarity and hydrogen bonding				
Subset of nonpolar or weakly polar compounds				
Tao *et al.*	81	0.847	−0.045	0.534
Meylan *et al.*	81	0.785	−0.036	0.635
Abraham approach	81	0.763	−0.010	0.665
Subset with strong H bond donor sites				
Tao *et al.*	249	0.610	+0.031	0.506
Meylan *et al.*	287	0.416	+0.122	0.679
Abraham approach	287	0.204	−0.076	0.793
Subset with H bond acceptor sites but without H bond donor sites [a]				
Tao *et al.*	178	0.628	+0.016	0.621
Meylan *et al.*	203	0.420	+0.135	0.804
Abraham approach	203	0.373	+0.145	0.836

[a] This subset includes also the only two polar compounds without H bond functionalities.

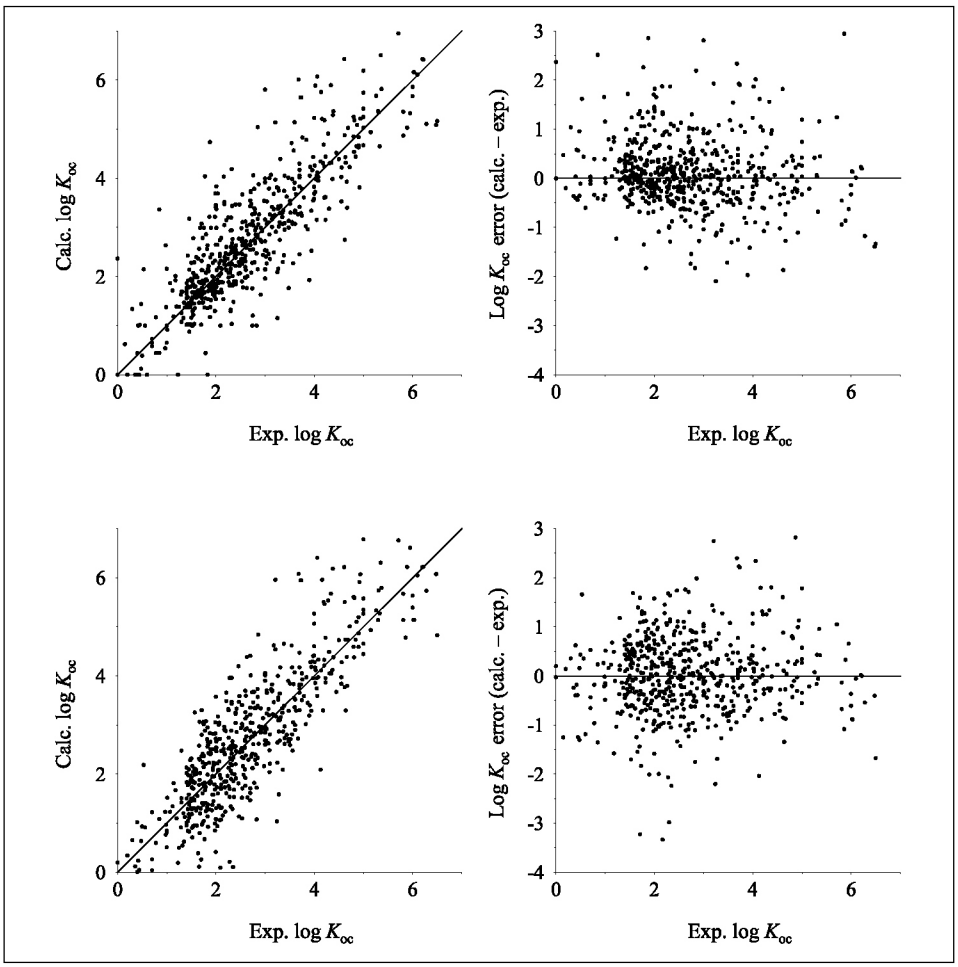


Fig. 2. Log K_{oc} calculation (left) and calculation errors (right) *vs.* experimental value for PCKOCWIN [20][21] (top) and for the Abraham approach [9][23] (bottom)

Table 4. Statistics for the estimation of log K_{oc} – subset with available experimental Abraham descriptors.

Model	No. compounds	q^2	*bias*	*se*
Tao *et al.* [19]	103	0.918	–0.050	0.443
Meylan *et al.* [20][21]	107	0.891	–0.018	0.515
Abraham approach [23]				
experimental descriptors	107	0.923	–0.047	0.432
estimated descriptors [9]	107	0.900	–0.093	0.494

Conclusions

At present, compounds with complex chemical structures are outside the model domain of the Abraham approach. This holds true for both the currently available models to estimate the descriptors and the model equation. The latter is caused by the fact that so far, Abraham equations have typically been calibrated with rather simple chemicals. Consequently, the most serious shortcomings of the Abraham approach are the current lack of reliable compound descriptors due to the limited number of measured data, and the resultant restriction of the chemical domain to relatively simple compounds. At the same time, the solid mechanistic basis of the Abraham model suggests that efforts should be undertaken to generate new experimental descriptors to feed the model equation. An increased availability of such descriptors for more complex compounds would also provide opportunity to improve the LFER prediction methods, and to extend the chemical domain of the Abraham model through respective re-calibration. To this end, it is recommended to focus on compounds with two and more chemical functionalities.

Acknowledgement

Financial support by the European Union (European Commission, FP6 Contract No. 003956, Integrated Project NoMiracle) is gratefully acknowledged.

Supplementary Material

The appendix contains the increment tables of the ChemProp [11] implementations of the Tao model [19] to predict log K_{oc} (Table A1), and of the Platts methods [9] to predict the Abraham parameters E, S, A, B^H and B^O from molecular structure (Tables A2–A4). The tables are available in electronic form through the website of the UFZ Department of Ecological Chemistry, *http://www.ufz.de/index.php?en=1785*, under the entry 'References'.

Received: August 15, 2006

[1] 'Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences', Eds. R.S. Boethling, D. Mackay, CRC Press, Boca Raton, **2000**.
[2] 'The Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs', OECD Document OECDENV/JM/MONO(2004)24, **2004**.
[3] J.C. Dearden, G. Schüürmann, *Environ. Toxicol. Chem.* **2003**, *22*, 1755–1770.
[4] W.J. Doucette, *Environ. Toxicol. Chem.* **2003**, *22*, 1771–1788.
[5] M.H. Abraham, G.S. Whiting, R.M. Doherty, W.J. Shuely, *J. Chromatogr.* **1991**, *587*, 229–236.
[6] M.H. Abraham, *Chem. Soc. Rev.* **1993**, *22*, 73–83.
[7] G. Cornelissen, Ö. Gustafsson, T.D. Bucheli, M.T.O. Jonker, A.A. Koelmans, P.C.M. van Noort, *Environ. Sci. Technol.* **2005**, *18*, 6881–6895.
[8] M.H. Abraham, J.C. McGowan, *Chromatographia* **1987**, *23*, 243–246.
[9] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inform. Comput. Sci.* **1999**, *39*, 835–845.
[10] ADME Boxes 2.5.7 Build 12 (Demo version), Advanced Pharma Algorithms, Inc., Toronto, Canada, **2004**.
[11] G. Schüürmann, R. Kühne, F. Kleint, R.-U. Ebert, C. Rothenbacher, P. Herth, in 'Quantitative Structure-Activity Relationships in Environmental Sciences – VII', Eds. F. Chen, G. Schüürmann, SETAC Press, Pensacola, **1997**, pp. 93–114.
[12] A. Sabljić, *Environ. Sci. Technol.* **1987**, *21*, 358–366.
[13] J. Huuskonen, *J. Chem. Inform. Comput. Sci.* **2003**, *43*, 1457–1462.
[14] T.H. Nguyen, K.U. Goss, W.P. Ball, *Environ. Sci. Technol.* **2005**, *39*, 913–924.
[15] G. Schüürmann, R.-U. Ebert, R. Kühne, *Environ. Sci. Technol.* **2006**, *40*, in press.
[16] W.M. Meylan, P.H. Howard, *Environ. Toxicol. Chem.* **1991**, *10*, 1283–1293.
[17] HENRYWIN 3.1 (EPI-Suite v.3.12), W.M. Meylan, Syracuse Research Corporation, Syracuse, NY, USA, **2000**.
[18] V.N. Viswanadhan, A.K. Ghose, U.C. Singh, J.J. Wendoloski, *J. Chem. Inform. Comput. Sci.* **1999**, *39*, 405–412.
[19] S. Tao, H. Piao, R. Dawson, X. Lu, H. Hu, *Environ. Sci. Technol.* **1999**, *33*, 2719–2725.
[20] W.M. Meylan, P.H. Howard, R.S. Boethling, *Environ. Sci. Technol.* **1992**, *26*, 1560–1567.
[21] PCKOCWIN 1.66 (EPI-Suite v.3.12), W.M. Meylan, Syracuse Research Corporation, Syracuse, NY, USA, **2000**.
[22] M.H. Abraham, J. Andonian-Haftvan, G.S. Whiting, A. Leo, R.S. Taft, *J. Chem. Soc. Perkin Trans. 2* **1994**, 1777–1791.
[23] S.K. Poole, C.F. Poole, *J. Chromatogr. A.* **1999**, *845*, 381–400.
[24] R. Kühne, R.-U. Ebert, G. Schüürmann, *Environ. Sci. Technol.* **2005**, *39*, 6705–6711.
[25] J.A. Platts, M.H. Abraham, D. Butina, A. Hersey, *J. Chem. Inform. Comput. Sci.* **2000**, *40*, 71–80.