The Science of the Total Environment, 109/110 (1991) 41–57
Elsevier Science Publishers B.V., Amsterdam

A validation study for the estimation of aqueous solubility from n-octanol/water partition coefficients*

Rainer Brüggemann and Joachim Altschuh

GSF-Projektgruppe Umweltgefährdungspotentiale von Chemikalien, Ingolstädter Landstr. 1, D-8042 Neuherberg, Federal Republic of Germany

ABSTRACT

Physico-chemical properties of chemicals are important data for exposure analysis. They can be estimated from structural information and/or via property–property relationships (PPRs). Often, such PPRs are restricted to distinct chemical classes. A concept has been developed in which molecules are characterized according to the presence or non-presence of typical structural elements (e.g., carbonyl functions, rings, aromatics). All structural elements are represented by binary digits (1 or 0 according to presence or non-presence). Using this tuple of digits, the estimation of aqueous solubility from the n-octanol/water partition coefficient is validated. According to different goodness criteria the PPRs are ranked by applying the graph-theoretical concept of HASSE diagrams. The best are analyzed further. Results indicate that some specific sets of molecules with the OH function should be excluded to keep the PPRs sufficiently accurate. Consequently, new PPRs especially for OH compounds are examined.

1. INTRODUCTION

Quantitative structure–activity relationships (QSARs) have become more and more important in the field of environmental fate and effect modelling of organic chemicals (see, e.g. [1, 2]). Property–property relationships (PPRs) for the estimation of the required data are an essential part of QSARs. They allow a quick estimation of substance properties without tedious analyses of molecular structure. This is particularly helpful if large lists of chemicals have to be evaluated, for example in a systematic evaluation of existing chemicals, for plausibility checks within data sets or for regulatory needs. We use PPRs for the estimation of physico-chemical properties to fill data gaps in the ECDIN data bank [3], which is managed by the Joint Research Centre of the EC in Ispra (Italy). To do this, accurate PPRs are required. The key issue in achieving confidence in the accuracy of PPRs and thus their applicability is

* This paper was presented at the workshop under the title "Automatic search for QSARs by an appropriate bit-pattern of molecular structure".

the evaluation of as many chemicals as possible. Although PPRs need no structural information as input, they are related indirectly to molecular structure. Therefore, the range of applicability of PPRs with respect to chemical classes needs careful examination. A general discussion of this aspect can be found in ref. 4.

In this paper we present a validation study and discuss the problem of applicability of PPRs with respect to different chemical classes. Different types of criteria for the validity of PPRs are applied. The relation between aqueous solubility (WS) and the n-octanol/water partition coefficient (K_{ow}) is chosen as an example. Data for about 800 chemicals were compiled from handbooks, various data banks, and original papers [5]. They are now stored in our data bank and the WS and K_{ow} values have been carefully checked.

2. CLASSIFICATION OF CHEMICAL STRUCTURES

Questions of homology, similarity and classification of chemical structures have been intensively studied [6]. The problem of classification also arises in studies of the validation of PPRs. The result of the validation study should be given in terms of accuracy *and* applicability. In order to derive a classification that additionally covers classification schemes already used in the literature on PPRs, the classes defined in ref. 7 were adopted. Other elements are also included. The classification scheme is given in Table 1. As this scheme was developed for studies of thermodynamic physico-chemical properties it does not explicitly take into account aspects of reactivity.

A tuple c of digits that indicate the presence ("1") or non-presence ("0") of the corresponding structural element characterizes the compounds. The components of c correspond to the numerical order of the structural elements in Table 1. A few examples may illustrate this:

Pentanone:	c = (0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)
Cyclohexane:	c = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
p-Nitrochlorobenzene:	c = (1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
Phenol:	c = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0)
Dichlorocamphene:	c = (0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)

This method can also be seen in more mathematical terms: the mapping, relating the set of molecules with the set of c-tuples, is a multi-to-one mapping as it assigns a given molecule to one and only one c-tuple. Conversely, in general more than one molecule can have the same c-tuple. Thus, the c-tuple is interpreted as a *bit-pattern*, defining corresponding sets of molecules. For example, carboxylic acids are included in the subset that can be obtained by selecting all compounds that have a "1" in the 9th and 10th position of the c-tuple. If acids with a halogen substituent are required, an additional "1" in

TABLE 1

Definition of structural elements

No.	Structural element	Examples (compounds that bear the corresponding structural element)
1	Aromatic	Benzene, phenol, biphenyl
2	Nonaromatic cyclic	Cyclohexane
3	Nonaromatic C=C	Butadiene
4	Nonaromatic C≡C	Acetylene
5	Halogen substitution	Bromobenzene, chloroform
6	N–O function	Nitrobenzene
7	N= or N≡	Benzonitrile
8	NR_3	Aniline, trimethylamine
9	C=O function	Acetone, acetic acid
10	OH function	Ethanol, phenol
11	Phosphororganic	
12	Sulfurorganic	
13	Topological genus[a]	Biphenyl, camphene
14	Heterocyclic	Atrazin
15	R–O–R function	Diethylether, methyl benzoate

[a] More than one ring.

the 5th position is necessary. More examples of the application of the bit-pattern are given in section 5.

As a result of the validation study the applicability of a given PPR will be characterized by 15 digits that can have three different values. For a given structural element these values are defined as follows:

— "1": the structural element is needed to guarantee the accuracy of the PPR

— "– 1": the structural element has to be excluded to maintain the accuracy of the PPR

— "0": the structural element does not seem to be crucial with respect to the accuracy of the PPR

Thus a PPR is characterized by a tuple c' of the above-defined digits. When a PPR is used to estimate a substance property the tuple c' characterizing its applicability should match the tuple c characterizing the given compound ("key–keyhole principle").

3. CRITERIA FOR THE VALIDATION OF PPRs

In this paper, we introduce different criteria to test the validity of PPRs:

— Mean square error (MSE), which combines the variance and the bias (for formulas, see [8]). This quantity is related to the number of compounds in the

test set (N). In order to have homogeneous symbols we define:

$$K_1 = \text{MSE} \tag{1}$$

— Number of deterministic "outliers" (N_Q), i.e. the number of compounds for which the quotient between the experimental and the calculated value is < 0.1 or > 10. As for MSE, the number of outliers has to be related to the total number of compounds:

$$K_2 = N_Q/N \tag{2}$$

The smaller the value of K_2 the more accurate the PPR. Each outlier is classified by its c-tuple. Thus not only the aspect of accuracy, but also that of applicability is considered.

— Analyzing the linear equation between measured (Y_m) and estimated values (Y_c):

$$Y_c = a_a Y_m + b_a \tag{3}$$

Ideally, the slope a_a should be close to 1 and the intercept b_a should be 0. From this we consider two additional criteria:

$$K_3 = \text{abs}(1 - a_a) \tag{4}$$

$$K_4 = \text{abs}(b_a) \tag{5}$$

— The calculation of the slope and intercept of Eqn (3) by arithmetic linear regression analysis implies that the predicted quantity is exact. In reality both quantities are stochastic. Therefore, as an alternative to arithmetic regression analysis, the geometric mean technique [9] was used to obtain the slope and the intercept. The criteria correspond to K_3 and K_4 and are denoted as K_5 and K_6.

All criteria are defined in such a way that small numbers for $K_i (i = 1\text{–}6)$ indicate greater accuracy. In the present validation study, K_1 is used as a single criterion and K_2–K_6 are analyzed by vector performance techniques (HASSE diagrams). Other techniques such as the bootstrap method [10] may also be useful and have been discussed elsewhere [29].

4. VALIDATION RESULTS FOR THE RELATION BETWEEN WS AND K_{ow}

4.1 *Overview*

The relation between aqueous solubility and octanol/water partition coefficient [Eqns (6) and (7)] has been studied from a theoretical point of view

[11–13] as well as with established empirical relations.

$$\log (WS) = a \log (K_{ow}) + b + cT_m \tag{6}$$

$$T_m = \begin{cases} T_m \text{ for } T_m > 25°C \\ 25°C \text{ otherwise} \end{cases} \tag{7a}$$

$$T_m = \begin{cases} T_m \text{ for } T_m > 25°C \\ 0 \text{ otherwise} \end{cases} \tag{7b}$$

(here, WS is given in mol/l and the melting point T_m in °C)

Consequently, a number of PPRs for these two properties can be found in the literature. Table 2 summarizes these PPRs and shows some characteristic features. The PPRs are different in at least three respects:

— use of melting point T_m as an additional predictor; this means that the physical state for all compounds is explicitly taken into account (denoted in the following as "state correction")

— specific or mixed chemical classes used as training sets

— number of compounds in the training set to establish the PPR.

From our data base described in Section 1, those compounds were extracted for which WS, K_{ow}, and T_m were available. This subset consists of 355 compounds. Because some relations do not require T_m, the corresponding subset is larger, consisting of 374 substances. These two sets of compounds were used in the validation study. The CAS registry numbers of the compounds are given in the Appendix.

4.2 *Correlation analysis*

Testing all PPRs leads to a tuple of numbers with six components, according to the six criteria K_i of Section 3. Analyzing the 26 PPRs, a 26 × 6 matrix is obtained. The results are summarized in Table 3. Although some parallelism between the K_i is expected, the correlation matrix (Table 4) shows rather small values if the outlying relation for phosphate esters (No. 16, Table 2) is omitted. The only exceptions are given by high correlations between K_1 and K_2, and between K_4 and K_6, which are consequences of their defining equations.

4.3 *Results, applying a single criterion*

The PPRs may be ranked according to the value of K_1 for each relation (cf. Table 3). If those relations are selected for which $K_1 < Q_1$ (first quartile), then *PPRs Nos 2, 23, 21, 24, 25, and 17 are recommended.* Analysis of this result with respect to the characteristic features of the 26 PPRs shows: (i) the

TABLE 2

Characterization of known PPRs between WS and K_{ow} [a, b, and c to Eqn (6); WS in mol/l]

No.	a	b	c	N	Range of applicability, reference
1	−1.37	1.64	−0.0094[a]	15	Halogenated aromatics, [14]
2	−1.12	1.3	−0.015[a]	27	"Mixed classes"[b], [15]
3[c]	−0.922	1.184	0	90	"Mixed classes", [16]
4	−1.49	1.46	0	34	"Mixed classes"[b], [17]
5	−1.113	0.926	0	41	Alcohols, [18]
6	−1.229	0.720	0	13	Ketones, [18]
7	−1.013	0.520	0	18	Esters, [18]
8	−1.182	0.935	0	12	Ethers, [18]
9	−1.221	0.832	0	20	Alkyl halides, [18]
10	−1.294	1.043	0	7	Alkynes, [18]
11	−1.294	0.248	0	12	Alkenes, [18]
12	−0.996	0.339	0	16	Aromatics, [18]
13	−1.237	−0.248	0	16	Alkanes, [18]
14	−1.214	0.850	−0.0095[d]	140	All chemicals from Nos 5–12, [18]
15	−1.339	0.978	−0.0095[d]	156	All chemicals from Nos 5–13, [18]
16	−2.38	6.9	0	11	Phosphate esters, [19]
17	−0.9874	0.7178	−0.0095[a]	35	Halobenzenes, [20]
18	−0.88	−0.012	−0.01[a]	32	PAHs, [21]
19	−0.962	0.5	0	9	Halogenated C_1-, C_2-hydrocarbons, [22]
20	−1.38	1.17	0	300	"Mixed classes", [13]
21	−1.26	1.0	−0.0054[a]	300	"Mixed classes", [13]
22	−1.18	0.84	0	NA	NA, [23]
23	−1.0	0.26	−0.01[a]	~40	"Mixed classes", [23]
24	−1.0	1.05	−0.01[a]		Theoretical relation, [11]
25	−1.05	0.87	−0.012[a]	155	"Mixed classes", [11]
26	−1.016	0.515	0	111	"Mixed classes", liquids, [24]

[a] State correction according to Eqn (7a).
[b] In the original paper calculated as log $K_{ow} = f(\text{WS})$.
[c] WS in g/l.
[d] State correction according to Eqn (7b).

six best PPRs are of the state-corrected type; (ii) five of these six were developed for "mixed classes"; (iii) the number of compounds in the training set is lower for four of the six best PPRs than the mean (140) or the median (138) numbers of compounds in the training sets; obviously a small number of compounds in the training set does not necessarily lead to lower accuracy. Thus, this analysis confirms that correction by T_m and the use of carefully selected training sets representing a high variability of chemical structures would lead to PPRs with a high applicability.

However, different types of information (bias and variance) is summed in

TABLE 3

Results of the validation study for the 26 PPR of Table 2

PPR	N	K_1	K_2	K_3	K_4	K_5	K_6
1	355	0.971	0.245	0.133	0.574	0.217	0.791
2	355	0.768	0.177	0.010	0.013	0.095	0.208
3	374	1.341	0.457	0.235	1.371	0.158	1.178
4	374	1.386	0.257	0.123	0.513	0.254	0.842
5	374	1.213	0.340	0.161	0.218	0.063	0.465
6	374	0.911	0.233	0.073	0.062	0.034	0.211
7	374	1.047	0.299	0.236	0.124	0.148	0.100
8	374	1.067	0.283	0.109	0.183	0.005	0.445
9	374	0.953	0.254	0.079	0.056	0.028	0.326
10	374	1.050	0.251	0.024	0.220	0.089	0.507
11	374	1.233	0.353	0.024	0.575	0.089	0.288
12	374	0.942	0.270	0.294	0.294	0.162	0.074
13	374	1.650	0.535	0.067	1.035	0.041	0.761
14	355	0.769	0.175	0.034	0.163	0.111	0.036
15	355	1.098	0.231	0.126	0.122	0.210	0.094
16	374	17.305	0.880	0.794	5.387	1.003	5.913
17	355	0.580	0.166	0.154	0.307	0.091	0.144
18	355	0.787	0.237	0.228	0.980	0.167	0.824
19	374	1.186	0.318	0.275	0.112	0.190	0.101
20	374	1.144	0.243	0.040	0.292	0.161	0.598
21	355	0.729	0.172	0.001	0.068	0.082	0.142
22	374	0.991	0.267	0.110	0.090	0.007	0.351
23	355	0.764	0.186	0.139	0.791	0.074	0.624
24	355	0.701	0.197	0.139	0.001	0.074	0.166
25	355	0.622	0.158	0.078	0.287	0.006	0.101
26	374	1.035	0.297	0.234	0.131	0.145	0.094

TABLE 4

Correlation matrix of the six criteria K_i [defined by Eqns (1)–(5)] calculated for 25 PPRs (PPR No. 16 is omitted)

	K_1	K_2	K_3	K_4	K_5
K_2	0.877				
K_3	0.116	0.268			
K_4	0.415	0.564	0.234		
K_5	0.254	0.042	0.430	0.218	
K_6	0.501	0.484	0.082	0.817	0.231

the mean square error. It is helpful to discuss directly the number of outliers (N_Q) and the properties of Eqn (3) (K_3 and K_4 or K_5 and K_6). The use of K_1–K_4 as single criteria results in four different "best" PPRs. Additionally, there are inversions in the ranking of the PPRs according to the various criteria. This confirms the need for a parallel analysis of more than one criterion. In the next section a ranking technique is applied that simultaneously takes different criteria into account.

4.4 *Ranking by use of HASSE diagrams*

Ranking several objects that are based on more than one property is performed using the theory of POSETS (partially ordered sets). The use of graph theory (HASSE diagrams) leads to a comprehensive picture of the situation. Details of this method can be found in the literature [25]. Here this technique is applied to analyze whether or not the PPRs designed for "mixed classes" and/or with T_m as an additional predictor are really more accurate.

Applying the criteria K_2, K_3, and K_4, the HASSE diagram shown in Fig. 1 results. From the endpoints of the diagram, *four PPRs, Nos 24, 25, 2 and 21*, are given as candidates for an applicable and accurate relation between K_{ow} and WS. With respect to the three criteria they are better than all other PPRs to which they are connected in the diagram. These four PPRs were originally developed for mixed classes of chemicals. On the other hand, five of the six worst PPRs (top of the diagram) were originally developed for specific classes.

Thus, the HASSE diagram technique verifies the general expectation that PPRs developed for mixed chemical classes are more accurate with respect to criteria K_2, K_3, and K_4 than those developed for specific classes if a test set consisting of mixed compounds is applied. Consequently, the PPR for phosphate esters (No. 16) is located at the top of the diagram. The coefficients of this relation differ strongly from the coefficients of the other PPRs (Table 2).

The 26 PPRs can be divided into two groups: 10 relations include T_m as predictor, the others do not take into account the state correction. Among the four best PPRs, only state-corrected PPRs are found, indicating that the *inclusion of melting point gives an improvement with respect to all three criteria.* Again, there are PPRs for which the number of compounds in the training set is rather low.

There is no connection between the two most prominent PPRs, Nos 2 and 21 (cf. Fig. 1). Therefore, some contradictions between the numerical values of the three criteria for the two PPRs must exist. As the data in Table 3 show, relation No. 21 is better than No. 2 with respect to two criteria and worse with respect to the third, the intercept (K_4). As the relative number of outliers (K_2) is approximately the same for both PPRs, the differences in the accuracy have

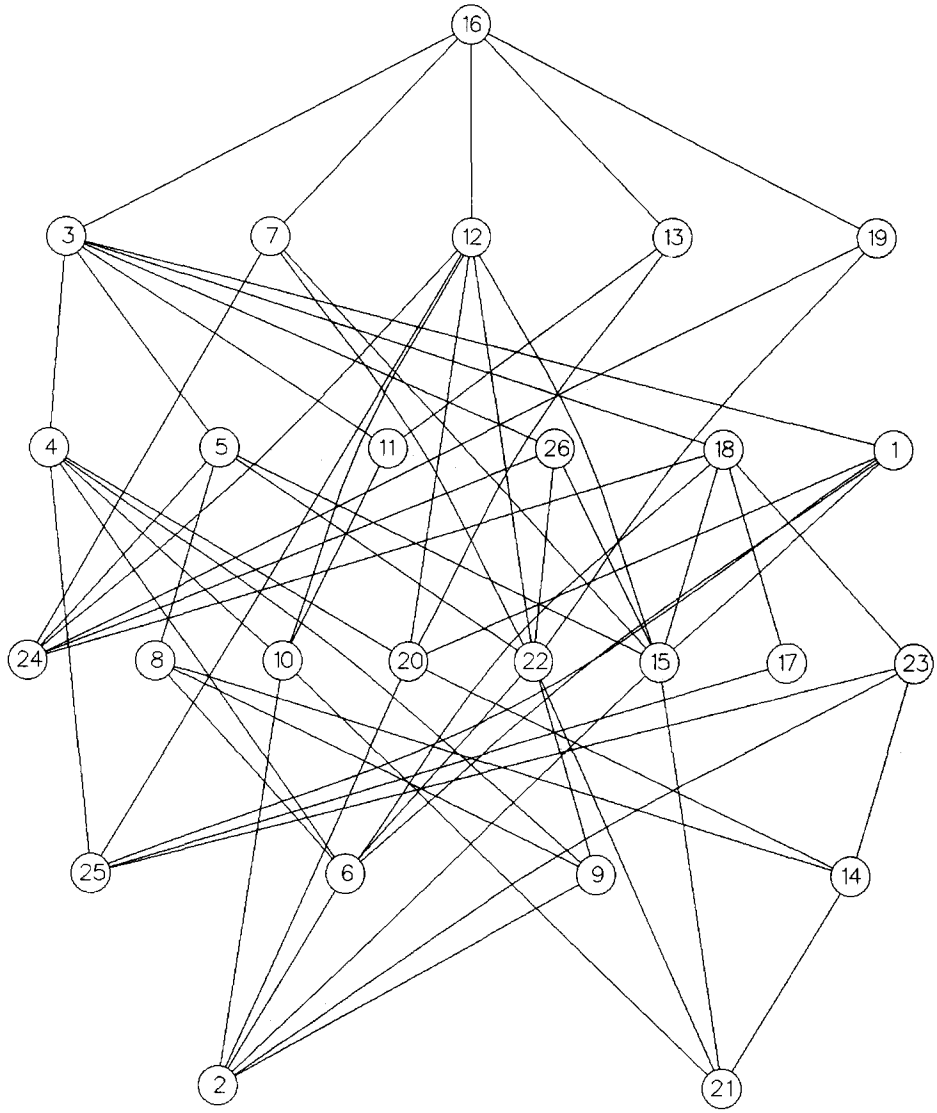

Fig. 1. HASSE diagram for K_2, K_3 and K_4 calculated from the complete test set of 374 and 355 compounds, respectively; all 26 PPRs (cf. Table 2) are considered. (PPRs improve proceeding from the top to the bottom of the figure.)

to be discussed in terms of Eqn (3): PPR No. 21 leads to an enhanced but constant error by underestimating WS. Due to the deviation of a_a from 1, PPR No. 2 is more reliable at low solubilities, but becomes more uncertain for increasing values of WS.

Analyzing HASSE diagrams with K_2, K_5 and K_6 as criteria, there is, in contrast to Fig. 1, a PPR that is distinctly better than the others with respect

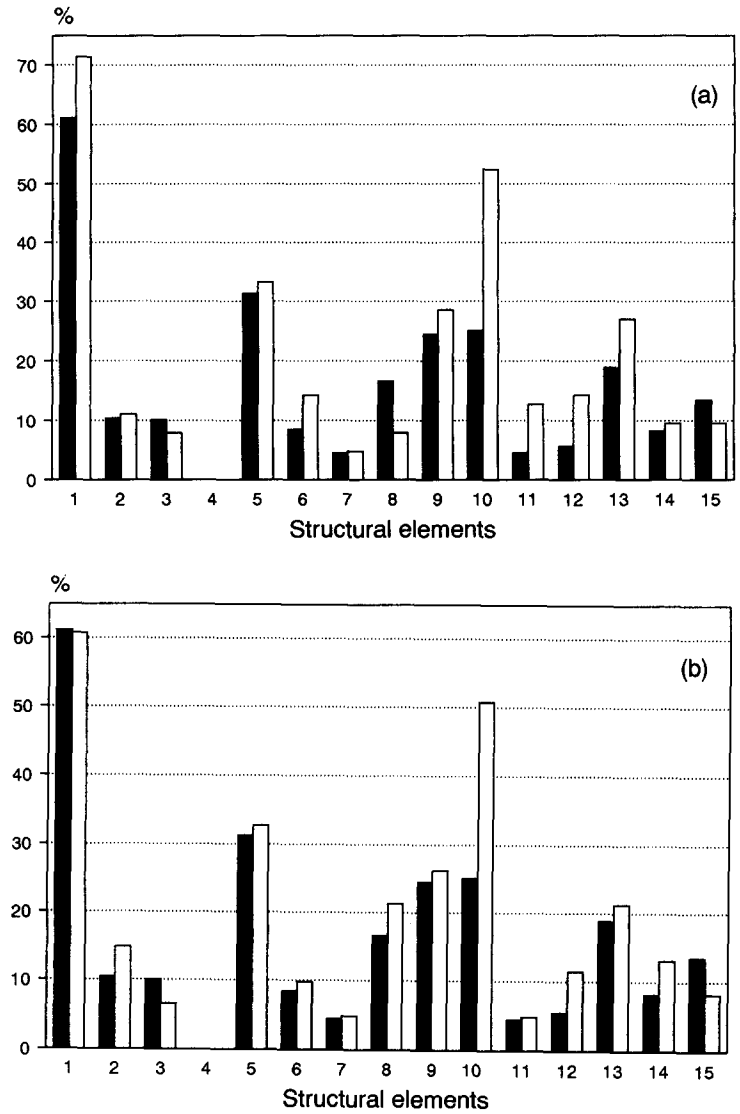

Fig. 2. Histograms of the structural elements present in the test set (■) and in the set of outliers (according to K_2) (□) for PPRs No. 2 (a) and No. 21 (b).

to all three criteria, i.e. *PPR No. 25*. By use of the bit-pattern of molecular structure the compounds that are outliers for the various PPRs are analyzed in more detail in the next section.

4.5 *Analysis of outliers*

The classification scheme defined in Section 3 will be used for a search of distinct patterns of chemical classes within the set of outliers. As an example,

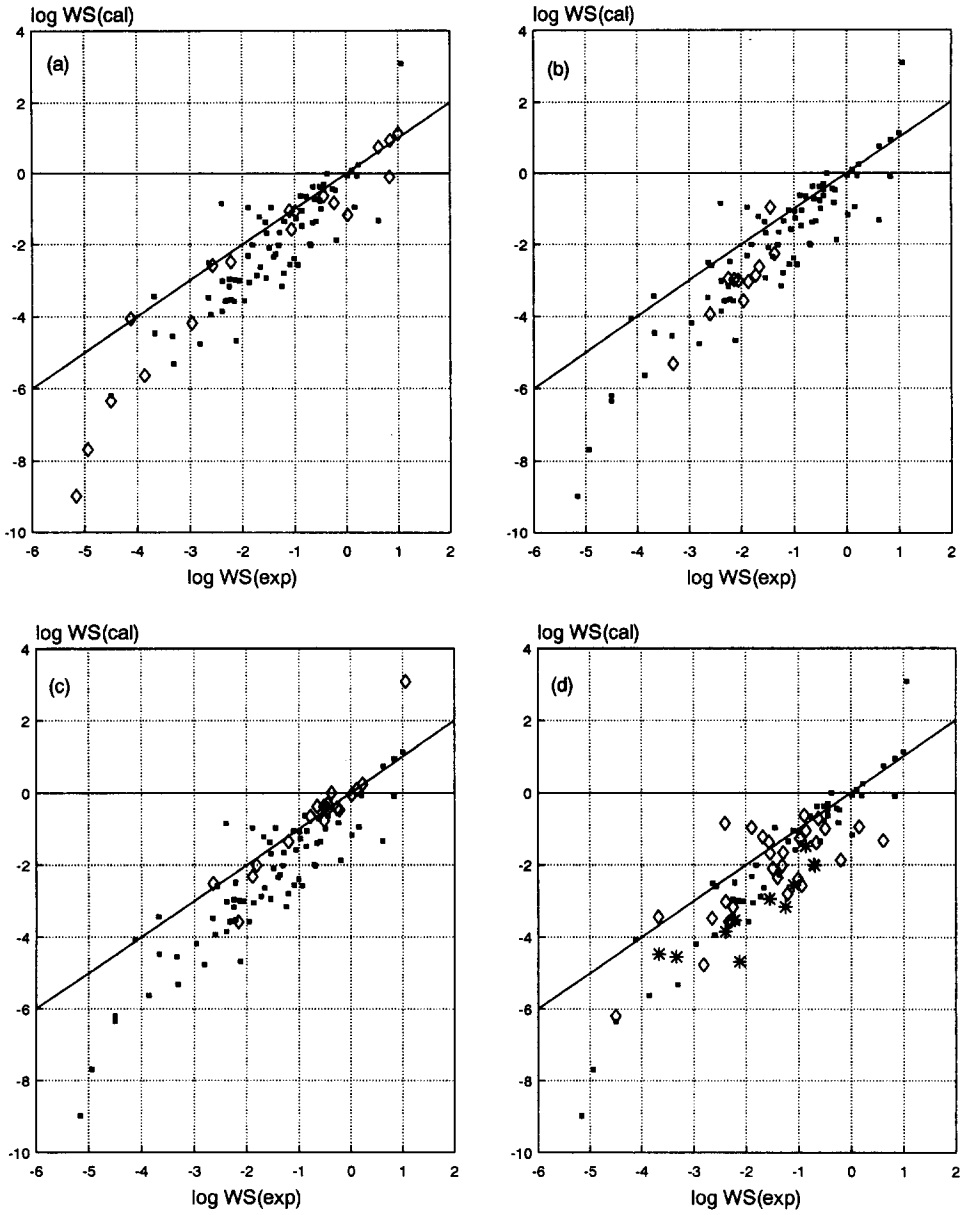

Fig. 3. Calculated (from PPR No. 2) versus measured WS for compounds bearing structural element 10 (OH function). Highlighted (◇) are (a) aliphatic carboxylic acids, (b) aromatic carboxylic acids, (c) aliphatic alcohols, and (d) phenols (phenols with a halogen substituent are indicated by *). The solid line corresponds to $a_a = 1$ and $b_a = 0$ in Eqn (3).

two prominent PPRs of Fig. 1, Nos 2 and 21, are analyzed more closely. The two histograms of Fig. 2 show the percentage distributions of the structural elements in the actual test set of 355 compounds and in the set of outliers according to criterion K_2. The higher proportion of compounds bearing the 10th structural element (OH functional group; see Table 1) in the set of outliers ($\sim 50\%$) compared with the complete data set ($\sim 25\%$) is most striking. In other words, the compounds with an OH function (whether aliphatic or aromatic alcohols or acids) are "accumulated" in the set of outliers. This fact could not be derived from the pure statistics of regression analysis. Thus, it is concluded that PPRs Nos 2 and 21 should not be applied to chemicals with an OH functional group. Inspection of the other PPRs also shows a considerable "accumulation" of OH compounds in the set of outliers. To maintain the accuracy of all the "best" PPRs (resulting from criterion K_1, and criteria groups K_2, K_3, K_4, and K_2, K_5, K_6, respectively), OH compounds should be excluded from the range of applicability if estimations for mixed classes are performed. This can easily be done within the framework of the substance classification presented here by applying the appropriate bit-pattern of chemical structure.

However, within the classification scheme presented here, aliphatic and aromatic alcohols as well as acids are included as OH compounds. A more detailed analysis (cf. Fig. 3) shows that the outliers are mainly *not* aliphatic alcohols except for compounds with a high aqueous solubility (e.g. 1,2-ethanediol). Some long-chained aliphatic carboxylic acids and some halogenated phenols are outliers for almost all of the PPRs. As some specific mechanism (e.g. ionization, hydrogen bonding) may be involved for compounds of this type (cf. the paper of Schwarzenbach et al. on nitrophenols [26] and the results of Boethling et al. [27]), a more detailed and thorough study including analysis of individual compounds is required. This is beyond the scope of the present paper and will be discussed elsewhere [28]. To overcome such problems, the data set used here to test existing PPRs is used as a learning set to derive new PPRs, especially for OH compounds.

5. SUGGESTIONS FOR NEW PPRs

The large compilation of experimental data can be used as a learning set to derive new relations between WS, K_{ow}, and T_m. Starting from the complete testing sets of 374 and 355 compounds, respectively, various subsets can be obtained by applying an appropriate bit-pattern of molecular structure. Table 5 shows this bit-pattern for the different subsets, together with the results of the linear regression analysis.

For the sets with a large number of compounds (S_1, S_1' and S_2), fairly good regression equations are derived. However, the model is considerably

TABLE 5

Results of the regression analysis to develop new PPRs [a, b and c refer to Eqn (6); the state correction is according to Eqn (7a)]

Learning set	Bit-pattern	N	a	b	c	r^2_{DF}	s^a	K_2^b
All compounds (S_1)	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	374	−1.06	0.178		0.802	0.905	0.24
All compounds (S_1')	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	355	−1.03	0.676	−0.008	0.866	0.753	0.15
No OH functional group (S_2)	0 0 0 0 0 0 0 0 0 −1 0 0 0 0 0	266	−1.07	0.593	−0.009	0.912	0.633	0.08
Compounds with OH groups (S_3)	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0	89	−0.737	0.589	−0.007	0.834	0.548	0.06
Alcohols (S_4)	0 0 0 0 0 0 0 0 −1 1 0 0 0 0 0	59	−0.745	0.549	−0.005	0.733	0.578	0.08
Aromatic alcohols (S_5)	1 0 0 0 0 0 0 0 −1 1 0 0 0 0 0	42	−0.770	0.614	−0.005	0.634	0.652	0.12
Alkylphenols (S_6)	1 −1 −1 −1 −1 −1 −1 −1 −1 1 −1 −1 −1 −1 −1	19	−0.882	1.02	−0.006	0.818	0.519	0.11
Aliphatic alcohols (S_7)	−1 −1 −1 −1 −1 −1 −1 −1 −1 1 −1 −1 −1 −1 −1	15	−0.705	0.359		0.854	0.398	0.00

[a] Standard deviation.
[b] See Eqn (2).

improved if (i) correction for the solid state is taken into account (S_1'), and (ii) compounds with OH groups (regardless of other substituents) are excluded (S_2) from the regression analysis. This confirms the results of Section 4.5.

The results of the regression models for specific classes of compounds vary substantially: for S_3 an acceptable r^2 value is obtained and the number of outliers is small. On the other hand, the exclusion of carboxylic acids (S_4) and the restriction to aromatic alcohols (S_5) make the models worse. Even for alkylphenols (S_6) the model is worse than for the subset including all compounds with OH groups (S_3). Obviously the enlarged number of compounds improves the statistics. Additionally, the coefficients of the regression equations for the more specific subsets S_3–S_6 deviate strikingly from those for the sets S_1 and S_2, indicating some specific mechanism involved in the partitioning processes. However, for the unsubstituted aliphatic alcohols (S_7) a most accurate, but least applicable, PPR can be derived.

6. SUMMARY

The validity of PPRs has to be discussed in terms of applicability *and* accuracy, which are closely related, but not identical. A purely phenomenological study is performed, taking the rather important relation between aqueous solubility and n-octanol/water partition coefficient (K_{ow} as predictor) as an example. From a practical point of view, statistical criteria considering the variance of the estimator and the bias simultaneously are comfortable (e.g. mean square error). However, the numerical values obtained from such a combined criterion may hide important information. For example, a given value for such a criterion may arise from a high variance and a low bias or vice versa.

Therefore, parallel to a discussion of the mean square error, five criteria are introduced and analyzed with the help of a graph theoretical technique (HASSE diagrams). Parallel with the results derived by use of the mean square error, they lead to a set of the best PPRs:

— Nos 2, 17, 21, 23, 24 and 25 from K_1
— Nos 2, 21, 24 and 25 from K_2, K_3 and K_4
— No. 25 from K_2, K_5 and K_6

Counting those PPRs occurring at least twice in the above rows we arrive at the set of Nos 2, 21, 24 and 25 as the "best". However, there is no definite recommendation for just one "best" PPR, because the numerical values derived for the criteria conflict with each other. Applying the bit-pattern of chemical structure, phenols, aromatic acids and long-chained aliphatic acids turn out to be overrepresented in the set of outliers for almost all of the PPRs. To maintain accuracy, they have to be excluded from the range of applicability. For these substance groups, new PPRs should be derived.

ACKNOWLEDGEMENT

We would like to thank Dr E. Halfon (Burlington) for providing programs to calculate the HASSE diagrams and for carefully reading the manuscript, and Prof. S. Yalkowsky (Tucson) for valuable discussions. Thanks are due to S. Sixt for his excellent assistance in collecting the experimental data from the literature. This work was supported by the Commission of the European Community under contract No. 3501-88-11 ED ISP D.

APPENDIX

CAS registry numbers of the 374 compounds used in the validation study. For the distribution of the structural elements within this set, see Fig. 2

50-29-3	50-32-8	51-28-5	53-70-3	55-38-9	56-23-5
56-55-3	57-10-3	57-11-4	57-13-6	58-89-9	59-50-7
60-11-7	60-29-7	60-51-5	60-57-1	62-23-7	62-44-2
62-53-3	62-56-6	63-25-2	64-19-7	65-85-0	67-66-3
67-72-1	69-72-7	70-55-3	71-36-3	71-41-0	71-43-2
71-55-6	72-20-8	72-54-8	72-55-9	74-11-3	74-82-8
74-83-9	74-85-1	74-87-3	74-88-4	74-89-5	74-98-6
75-00-3	75-01-4	75-09-2	75-21-8	75-28-5	75-34-3
75-35-4	75-37-6	75-50-3	75-52-5	75-56-9	75-71-8
75-85-4	76-01-7	76-03-9	76-13-1	77-47-4	77-92-9
78-40-0	78-59-1	78-78-4	78-83-1	78-87-5	78-92-2
78-93-3	79-00-5	79-01-6	79-06-1	79-20-9	79-24-3
79-34-5	80-05-7	80-62-6	82-68-8	83-32-9	83-34-1
84-66-2	84-74-2	85-01-8	85-41-6	85-44-9	85-68-7
86-50-0	86-73-7	86-74-8	87-61-6	87-68-3	87-86-5
88-06-2	88-19-7	88-72-2	88-73-3	88-74-4	88-75-5
88-85-7	88-89-1	89-83-8	90-02-8	90-05-1	90-12-0
90-13-1	90-15-3	90-43-7	90-64-2	91-15-6	91-20-3
91-22-5	91-23-6	91-57-6	91-58-7	91-64-5	91-66-7
91-94-1	92-52-4	92-87-5	93-76-5	94-75-7	95-47-6
95-48-7	95-49-8	95-50-1	95-51-2	95-53-4	95-54-5
95-55-6	95-57-8	95-63-6	95-87-4	95-93-2	95-94-3
95-95-4	96-09-3	96-18-4	96-22-0	96-33-3	96-37-7
97-00-7	98-06-6	98-54-4	98-82-8	98-86-2	98-95-3
99-04-7	99-08-1	99-09-2	99-62-7	99-94-5	99-96-7
99-99-0	100-00-5	100-01-6	100-02-7	100-09-4	100-17-4
100-41-4	100-42-5	100-44-7	100-46-9	100-47-0	100-51-6
100-61-8	100-66-3	101-21-3	101-42-8	101-81-5	102-70-5
103-29-7	103-33-3	103-65-1	103-84-4	104-40-5	104-51-8
105-37-3	105-60-2	105-67-9	106-35-4	106-37-6	106-41-2
106-42-3	106-43-4	106-44-5	106-46-7	106-47-8	106-48-9
106-49-0	106-89-8	106-98-9	106-99-0	107-02-8	107-06-2
107-13-1	107-21-1	107-87-9	108-05-4	108-10-1	108-11-2
108-21-4	108-38-3	108-39-4	108-41-8	108-42-9	108-43-0

(*continued*)

APPENDIX (*continued*)

108-44-1	108-46-3	108-67-8	108-68-9	108-70-3	108-73-6
108-86-1	108-87-2	108-88-3	108-90-7	108-93-0	108-94-1
108-95-2	108-98-5	109-52-4	109-60-4	109-66-0	109-67-1
109-69-3	109-89-7	109-99-9	110-00-9	110-02-1	110-15-6
110-16-7	110-17-8	110-43-0	110-82-7	110-83-8	111-13-7
111-26-2	111-27-3	111-44-4	111-65-9	111-66-0	111-69-3
111-70-6	111-76-2	111-87-5	115-11-7	117-81-7	118-74-1
118-79-6	118-90-1	118-91-2	119-61-9	119-90-4	120-12-7
120-72-9	120-80-9	120-82-1	120-83-2	121-44-8	121-69-7
121-73-3	121-75-5	121-87-9	121-91-5	121-92-6	122-39-4
122-59-8	122-99-6	123-30-8	123-31-9	123-38-6	123-51-3
123-54-6	123-63-7	123-86-4	123-96-6	124-04-9	124-07-2
126-98-7	126-99-8	127-18-4	129-00-0	131-11-3	133-06-2
133-07-3	134-32-7	135-19-3	140-88-5	141-78-6	142-84-7
143-07-7	143-08-8	149-91-7	150-13-0	150-19-6	150-68-5
150-76-5	198-55-0	206-44-0	207-08-9	208-96-8	217-59-4
218-01-9	238-84-6	243-17-4	260-94-6	298-00-0	298-02-2
298-04-4	299-84-3	309-00-2	315-18-4	330-54-1	330-55-2
462-06-6	470-90-6	502-56-7	526-73-8	534-22-5	535-80-8
538-68-1	540-54-5	541-73-1	544-63-8	544-76-3	552-16-9
554-84-7	563-12-2	563-80-4	573-56-8	575-41-7	576-26-1
581-42-0	583-53-9	583-57-3	584-02-1	589-90-2	591-27-5
591-50-4	608-93-5	611-14-3	613-33-2	622-96-8	629-59-4
634-66-2	634-90-2	638-53-9	646-04-8	691-37-2	709-98-8
732-11-6	779-02-2	821-55-6	939-27-5	1194-65-6	1563-66-2
1689-84-5	1746-81-2	1912-24-9	1918-00-9	1918-16-7	1929-82-4
2051-60-7	2051-61-8	2051-62-9	2104-64-5	2310-17-0	2385-85-5
2463-84-5	2921-88-2	3209-22-1	5902-51-2	6032-29-7	6915-15-7
15972-60-8	21609-90-5				

REFERENCES

1 N.N. Nirmalakhandan and R.E. Speece, Structure–activity relationships. Environ. Sci. Technol., 22 (1988) 606–615.
2 W. Karcher and J. Devillers, in W. Karcher and J. Devillers (Eds), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Toxicology and Chemistry, Kluwer Academic Publishers, Dordrecht, 1990, pp. 1–12.
3 O. Norager, in ref. 6, pp. 195–209.
4 R. Brüggemann, J. Altschuh and M. Matthies, in ref. 2, pp. 197–212.
5 J. Altschuh and R. Brüggemann, in Proc. Workshop Chemical Exposure Prediction, Trois-Epis, June 1990, European Science Foundation, Strasbourgh, 1990, pp. 9–19.
6 W.A. Warr (Ed.), Chemical Structures, Springer-Verlag, Berlin, 1988.
7 W.J. Lyman, W.F. Reehl and D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC, 1990.
8 M. Precht, Bio-Statistik, Oldenbourg Verlag, München, 1987, p. 158.

9 E. Halfon, Regression method in ecotoxicology: a better formulation using the geometric mean functional regression. Environ. Sci. Technol., 19 (1985) 747–749.
10 E. Halfon, The bootstrap and the jackknife in ecotoxicology or nonparametric estimates of standard error. Chemosphere, 14 (1985) 1433–1440.
11 S.H. Yalkowsky and S.C. Valvani, Solubility and partitioning. I. Solubility of nonelectrolytes in water. J. Pharm. Sci., 69 (1980) 91–922.
12 D. Mackay, A. Bobra, W.-Y. Shiu and S.H. Yalkowsky, Relationships between aqueous solubility and octanol–water partition coefficients. Chemosphere, 9 (1980) 701–711.
13 P. Isnard and S. Lambert, Aqueous solubility and n-octanol/water partition coefficient correlations. Chemosphere, 18 (1989) 1837–1853.
14 R. Brüggemann, unpublished results, 1989.
15 S. Banerjee, S.H. Yalkowsky and S.C. Valvani, Water solubility and octanol/water partition coefficients of organics. Limitations of the solubility–partition coefficient correlation. Environ. Sci. Technol., 14 (1980) 1227–1229.
16 E.E. Kenaga and C.A.I. Goring, in J.G. Eaton, P.R. Parrish and A.C. Hendricks (Ed.), Aquatic Toxicology, ASTM STP 707, American Society for Testing and Materials, Philadelphia, 1980, pp. 78–115.
17 C.T. Chiou, V.H. Freed, D.W. Schmedding and R.L. Kohnert, Partition coefficient and bioaccumulation of selected organic chemicals. Environ. Sci. Technol., 11 (1977) 475–478.
18 C. Hansch, J.E. Quinlan and G.L. Lawrence, The linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. J. Org. Chem., 33 (1968) 347–350.
19 V.W. Saeger, O. Hicks, R.G. Kaley, P.R. Michael, J.P. Mieure and E.S. Tucker, Environmental fate of selected phosphate esters. Environ. Sci. Technol., 13 (1979) 840–844.
20 S.H. Yalkowsky, R.J. Orr and S.C. Valvani, Solubility and partitioning. 3. The solubility of halobenzenes in water. Ind. Eng. Chem. Fundam., 18 (1979) 351–353.
21 S.H. Yalkowsky and S.C. Valvani, Solubilities and partitioning. 2. Relationships between aqueous solubilities, partition coefficients, and molecular surface areas of rigid aromatic hydrocarbons. J. Chem. Eng. Data, 24 (1979) 127-129.
22 C.T. Chiou and V.H. Freed, Chemodynamic Studies on Bench Mark Industrial Chemicals, Rep. No. NSF/RA-770286, National Science Foundation, Washington, DC, 1977 (cited in ref. 7).
23 G.G. Briggs, Theoretical and experimental relationships between soil adsorption, octanol–water partition coefficients, water solubilities, bioconcentration factors, and the parachor. J. Agric. Food. Chem., 29 (1981) 1050–1059.
24 S.C. Valvani, S.H. Yalkowsky and T.J. Roseman, Solubility and partitioning. IV. Aqueous solubility and octanol–water partition coefficients of liquid nonelectrolytes. J. Pharm. Sci., 70 (1981) 502–507.
25 E. Halfon and M.G. Reggiani, On ranking chemicals for environmental hazard. Environ. Sci. Technol., 20 (1986) 1173–1179.
26 R.P. Schwarzenbach, R. Stierli, B.R. Folsom and J. Zeyer, Compound properties relevant for assessing the environmental partitioning of nitrophenols. Environ. Sci. Technol., 22 (1988) 83–92.
27 R.S. Boethling, S.E. Campbell, D.G. Lynch and G.D. LaVeck, Validation of CHEMEST, an on-line system for the estimation of chemical properties. Ecotoxicol. Environ. Saf., 15 (1988) 21–30.
28 R. Brüggemann and J. Altschuh, to be published.
29 E. Halfon, J. Altschuh, R. Brüggemann and W. Karcher, Estimations of aqueous solubility from n-octanol/water partition coefficients analyzed by the bootstrap method. Chemosphere, 22 (1991) 953–957.