J. Mol. Model. **1997**, *3*, 142 – 155

Journal of Molecular Modeling

Prediction of the n-Octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO-Calculations and a Neural Network

Andreas Breindl, Bernd Beck, and Timothy Clark*

Computer Chemie Centrum des Instituts für Organische Chemie der Friedrich-Alexander-Universität Erlangen-Nürnberg,
Nägelsbachstraße 25, D-91052 Erlangen, Germany (clark@organik.uni-erlangen.de).

Robert C. Glen

Wellcome Research Laboratories, Langley Park, Beckenham, Kent, BR3 3BS, U. K.
Present address: Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA,

Received: 21 November 1996 / Accepted: 4 March 1997 / Published: 17 March 1997

Abstract

A back-propagation artificial neural net has been trained to estimate logP values of a large range of organic molecules from the results of AM1 and PM3 semiempirical MO calculations. The input descriptors include molecular properties such as electrostatic potentials, total dipole moments, mean polarizabilities, surfaces, volumes and charges derived from semiempirical calculated gas phase geometries. These properties can be related to the molecule's solubility in hydrophilic or lipophilic media. The input descriptors were selected with the help of a multiple linear regression analysis. The resulting net estimates the logP values of 105 organic compounds with a standard deviation of 0.53 units from the experimental logP values for AM1 and 0.67 units in the case of PM3.

Keywords: Partition coefficient, logP, AM1, PM3, QSAR, neural net

Introduction

It is now possible to derive accurate molecular properties with semiempirical molecular orbital theory. Especially electrostatic properties such as dipole moments, polarizabilities and electrostatic potentials are important and can often be related to experimental behaviour [1]. On the other hand, there is strong interest in the theoretical prediction of physically measurable properties for the development of new drugs. In this respect the n-octanol/water partition coefficient is an important parameter that is a measure of the extent to which a solute is distributed between water and a water-immisicible liquid phase. The n-octanol/water partition coefficient is the ratio of a chemical's concentration in n-octanol to that in water in a two phase system at equilibrium. Since measured values of the partition coefficient range from less than 10^{-4} to larger than 10^{8} (at least 12 orders of magnitude), the logarithm, logP, is commonly used to characterise its value. LogP is used extensively to describe a compound's lipophilic or

* *To whom correspondence should be addressed*

hydrophobic properties and is therefore a valuable parameter in many QSAR studies that have been developed for pharmaceutical, environmental, biochemical and toxicological sciences [2]. Many studies have shown that logP is useful for correlating a drug molecule's transport properties or its interactions with receptor molecules, and changes in its structure with various biochemical or toxic effects [3]. Although logP is generally easy to determine, the reliability of predicted values is important during the design process. Until now, mostly empirical methods have been developed.

Among others [4], there are two widely used, essentially empirical methods for the estimation of logP, Rekker's f constant method [5] and Leo and Hansch's fragment approach [6]. Rekker first defined an arbitrary set of terminal fragments using a database of about 1000 compounds with known logP. Linear regression analysis was performed with the numbers of the different substructures as the independent variable and logP as the dependent variable. The regression coefficients obtained are used as group contributions. To estimate the partition coefficient of a compound, one simply sums up the group contributions and the appropriate correction factors. Leo and Hansch's philosophy was to determine logP values of a set of small molecules very accurately and calculate the fragmental values from these data. Using the concept of isolating carbons (sp^3 carbon atoms with at least two bonds linked to other carbon atoms), they derived their own set of terminal fragments. This system also includes many correction factors (e. g. for multiple halogenation or different double bonds). Although essentially all logP values for the compounds included in the base set are well reproduced, it is often a problem to divide (fragment) a molecule correctly, especially for complex drug molecules, or to use the many correction factors. In order to overcome this situation, new fragment methods (atomic fragments) were developed [7] but even so, not all problems of these methods could be solved.

Recently, methods have been proposed that utilise properties of the entire solute molecule, e. g. molecular surface area, volume, charge density or electrostatic potential, to predict logP [8]. These methods attempt to overcome various inefficiencies of the fragment constant approach, e. g. the need for correction factors or the inability to estimate logP for unknown fragments. For example, Herges et al. have used a combination of semiempirical self-consistent reaction field calculations (SCRF) and a neural network [9]. There have also been attempts to calculate logP directly from the solvation energies. For instance, Reynolds et al. [10] have used free-energy perturbation calculations for a series of acyclic

Table 1. *Calculated properties for the 194 compounds in the data set.*

Property	Symbol	Reference
total dipole moment	D_t	17
mean polarizability	POL	17
molecular surface	SUR	17,18
molecular volume	VOL	17,18
globularity	GLOB	17,19
sum of the electrostatic potential (ESP) derived atomic charges		
on the nitrogen atoms	NSUM	17,20
on the oxygen atoms	OSUM	17,20
Parameters introduced by Politzer et al.:		
highest electrostatic potential	ESP_{max}	17,21,22
lowest electrostatic potential	ESP_{min}	18,21,22
number of surface points with positive ESP	n_{pos}	17,23
number of surface points with negative ESP	n_{neg}	17,23
mean value of positive ESP	$MEAN_+$	17,21,22
mean value of negative ESP	$MEAN_-$	18,21,22
positive variance	σ_+^2	24
negative variance	σ_-^2	24
total variance	σ_{tot}^2	24
balance parameter	ν	24
histogram including the number of surface points within a specified range of the ESP (8-point)	h1-h8	17,21,22,23

alcohols and find agreement with experimental logP values of ±0.45 units. Cramer, Truhlar et al. [11] and Klamt [12] have used differences in semiempirical calculated SCRF solvation energies. In our approach to estimating logP, semiempirical MO-methods are used to calculate a set of molecular and atomic properties from gas phase geometries in order to use them as descriptors for a back-propagation neural network. Our approach is intended to estimate logP from a single, fast gas phase calculation and should therefore be more appropriate for rapid scans of large numbers of molecules. In an initial test, 194 different organic compounds were used as data set in order to test the reliability of the method. The results obtained were then taken as starting point for the quantitative-structure property relationship (QSPR) on a data set containing 1085 compounds. It includes a wide spectrum of organic compounds, such as nitrogen-, oxygen-, sulfur- and phosphorus-containing molecules, alcohols, ethers, halogenated compounds, amino acids and various aromatic or heteroaromatic molecules.

Computational details

All organic compounds and their experimental logP values were taken from a database of the Wellcome Research Laboratories, Beckenham, Kent. The program CONCORD [13] was used to convert the 2D- into 3D-structures. The geometries were then checked and, if necessary, modified with the help of the molecular modelling program package SYBYL [14]. The structures were optimised using AM1[15] and PM3 [16] included in the semiempirical program VAMP 6.0 [17]. In the case of the small data set, only AM1 was used. Amino acids were calculated in their zwitterionic forms. For the initial approach, a set of molecular properties was calculated using a slightly modified VAMP version. In this version, descriptors developed by Politzer et al. [24] are derived from molecular electrostatic surfaces. In total the 27 MEP-derived descriptors shown in Table 1 were generated for the small data set.

The globularity (GLOB) [19] is often referred to as the deviation from spherity. It is calculated as the ratio of the surface area of a sphere of volume equal to the calculated molecular volume and the surface area of the molecule. If the molecule is perfectly spherical, the globularity is one. The histogram consists of 8 values. These are the number of surface points (generated with a modified „Marsili“ algorithm [23]) having a ESP within a defined range. The ranges were defined as follows:

h1 ESP more negative than 100 kcal mol^{-1}
h2 ESP between -100 and -60 kcal mol^{-1}
h3 ESP between -60 and -20 kcal mol^{-1}
h4 ESP between -20 and 0 kcal mol^{-1}
h5 ESP between 0 and 20 kcal mol^{-1}
h6 ESP between 20 and 60 kcal mol^{-1}
h7 ESP between 60 and 100 kcal mol^{-1}
h8 ESP more positive than 100 kcal mol^{-1}

This approach was used in order to obtain a more detailed description of the calculated electrostatic potential at the surface.

Another data set, also derived from the electrostatic potential, includes the properties introduced by Politzer et al. [24] in their work on the interactions of solute and solvent. The positive (σ_+^2) and negative (σ_-^2) variance are calculated from the positive ($V^+(r_i)$) and negative ($V^-(r_j)$) values of the electrostatic potential $V(r)$ on the molecular surface and their averages:

$$\sigma_+^2 = \frac{1}{m}\sum_{i=1}^{m}\left[V^+(r_i) - \overline{V}_s^+\right]^2 \tag{1}$$

$$\sigma_-^2 = \frac{1}{n}\sum_{j=1}^{n}\left[V^-(r_j) - \overline{V}_s^-\right]^2 \tag{2}$$

In other words, σ_+^2 describes the standard deviation over the positive molecular electrostatic potential (MEP) regions of the molecule’s surface whereas σ_-^2 describes the negative counterpart.

The total variance was then calculated as the sum of σ_+^2 and σ_-^2.

$$\sigma_{tot}^2 = \sigma_+^2 + \sigma_-^2 \tag{3}$$

The total variance, σ_{tot}^2, is a measure of the spread of the surface potential and is particularly sensitive to variations in its magnitude, emphasising positive and negative extremes. It has been interpreted to be indicative of a molecule’s tendency for electrostatic interactions [24]. Finally, the so called balance parameter ν is derived using equation (4):

$$\nu = \frac{\sigma_+^2 \sigma_-^2}{\left[\sigma_{tot}^2\right]^2} \tag{4}$$

ν represents the manner in which σ_{tot}^2 affects interactive tendencies more accurately. It attains its maximum value when σ_+^2 and σ_-^2 are equal. This means that the molecule interacts to a similar extent (whether strongly or weakly) through both its positive and negative regions. The descriptor set was extended for the large data set. The additional MEP-derived parameters are listed in Table 2.

These additional parameters were used in order to obtain a more detailed description of the charge distribution within the molecules. Overall 36 descriptors were created for the large data set.

Table 2. *Additional properties calculated for the 1085 compounds in the large data set*

Property	Symbol	Reference
sum of the electrostatic potential (ESP) derived atomic charges		
on the sulfur atoms	SSUM	17,20
on the phosphorous atoms	PSUM	17,20
on the fluorine atoms	FSUM	17,20
on the chlorine atoms	CLSUM	17,20
on the bromine atoms	BrSUM	17,20
on the iodine atoms	ISUM	17,20
sum of the electrostatic potential (ESP) derived atomic charges of all halides	HalSUM	17,20
histogram including the number of surface points within a specified range of the ESP (12-point)	h1-h12	17,21,22,23

Results

The discussion of the results is divided into two sections. The first describes the results obtained for the small data set using multiple linear regression analyses [25] and a back-propagation neural network [26]. The second part describes similar results for the large data set.

Small data set

Multiple linear regression analyses

Generally, use of multiple linear regression analyses is a very fast method in order to identify the calculated properties that are important for the prediction of experimental quantities. These properties should be useful as descriptors for a neural net. All investigations were performed using the QSAR program TSAR [27]. The best results were obtained for the input parameters ESP_{max}, ESP_{min}, $MEAN_+$, $MEAN_-$, D_t, POL, SUR, VOL, GLOB, NSUM, OSUM, n_{pos}, n_{neg}, σ^2_+, σ^2_-, σ^2_{tot} and and their squares. The linear regression coefficients r, r^2 and the t-values (ratio of the input parameter's regression coefficient and the mean error) were used to select the best regression performance. The following significant 10 term equation (5) results from the regression analysis:

$$\log P = (0.011 \pm 0.001)\ ESP_{min} + (0.036 \pm 0.005)\ SUR - (0.203 \pm 0.023)\ D_t + (0.824 \pm 0.209)\ NSUM + (0.193 \pm 0.190)\ NSUM^2 + (1.126 \pm 0.219)\ OSUM + (0.120 \pm 0.075)\ OSUM^2 - (2.6\cdot 10^{-7} \pm 6.4\cdot 10^{-8})\ n^2_{pos} - (5.1\cdot 10^{-7} \pm 1.6\cdot 10^{-7})\ n^2_{neg} - (0.001 \pm 0.0006)\ \sigma^2_- - (1.279 \pm 0.506) \tag{5}$$

$r = 0.936 \quad r^2 = 0.876$

The t-values of the coefficients are shown in Table 3 and the regression plot is shown in Figure 1.

The linear regression coefficient r is 0.936, the square r^2 0.876 and the cross validation r_{cv} 0.872. The standard deviation is 0.532.

Discussion

The predictive power of our approach is promising, although the mathematical relationship of the different variables is quite simple. The descriptors found in equation (5) were therefore used as starting point for the back-propagation neural network.

Table 3. *t-values of regression equation (5).*

Input parameter	Coefficient	t-value
ESP_{min}	0.011	2.755
SUR	0.036	15.640
D_t	0.203	7.645
NSUM	0.824	6.441
$NSUM^2$	0.193	2.340
OSUM	1.126	8.211
$OSUM^2$	0.120	2.717
${n_{pos}}^2$	$2.6\cdot 10^{-7}$	4.779
${n_{neg}}^2$	$5.1\cdot 10^{-7}$	6.583
σ^2_-	1.279	2.460

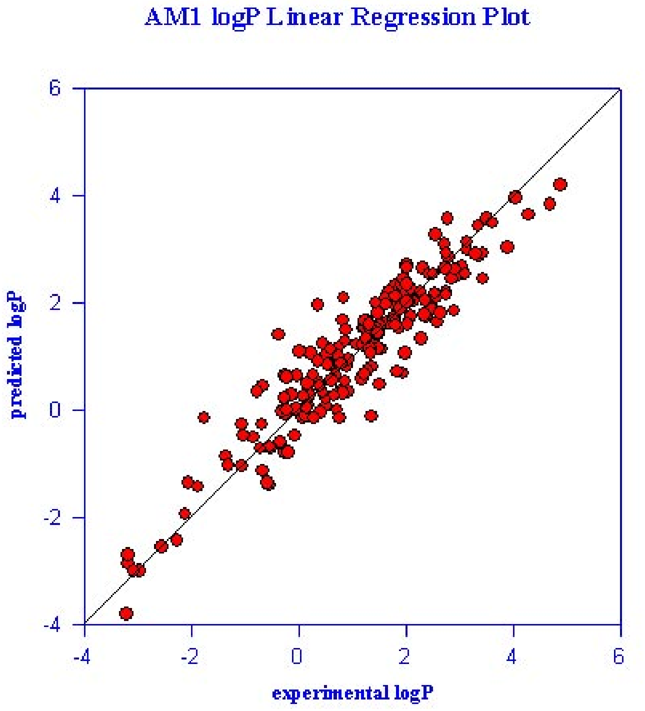


Figure 1. *Best predictions of the logP values by multiple linear regression analysis.*

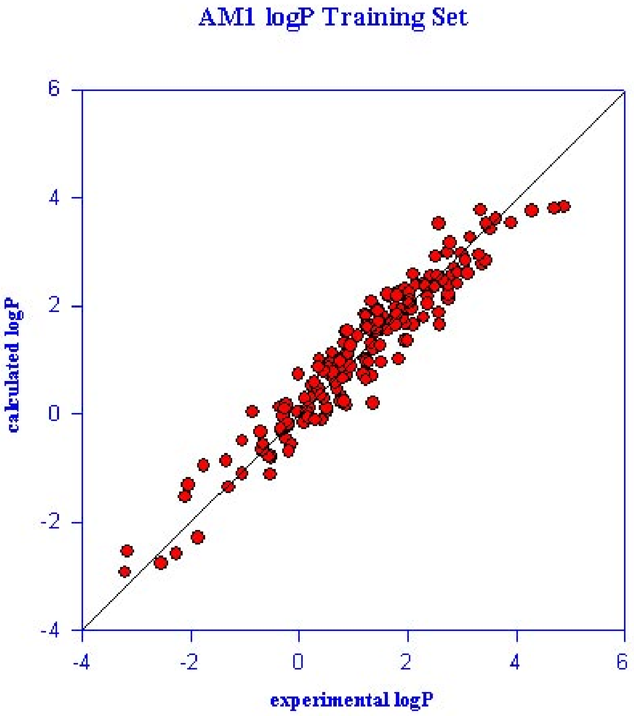


Figure 2. *Best predictions obtained for the logP training set using a 12:4:1 neural net.*

Neural network

We recently introduced a technique to estimate ^{13}C chemical shifts using a combination of semiempirical calculations and a back-propagation network [28]. The concept of a supervised learning algorithm (implemented in back-propagation networks) is well suited as a nonlinear device for linking semiempirically calculated parameters with experimental quantities. Generally, neural networks are able to perform highly nonlinear pattern recognition, classification and regression tasks, the results of which are often superior to traditional approaches. Recent applications of neural networks include the determination of structure-activity relationships in drug design (QSAR) [29-31], the prediction of protein structure [32] and the classification of spectra [33]. Thermodynamic data such as solubilities [34] and boiling points of organic heterocycles [35] have also been the subject of investigation. The network simulation program ANsim [36] was used in this work.

Results of the back-propagation net

The small training set contained mainly aliphatic- and cyclic hydrocarbons, aromatic- and heteroaromatic compounds containing oxygen-, phosphorus- and nitrogen atoms. The number of descriptors was varied within a range of 8 to 13 parameters. The parameters of equation (5) were taken as starting point. Table 4 shows the best descriptor set, which was obtained by trial-and-error variation of the descriptor set starting with those that appear in equation (5). The correlation obtained is illustrated in Figure 2.

The correlation coefficient r obtained is 0.962, the square r^2 is 0.925 with a standard deviation of 0.393.The cross vali-

Table 4. *Descriptor set for the 12:4:1 back-propagation network.*

Number	Descriptor
1	total dipole moment D_t
2	mean polarizability POL
3	molecular surface SUR
4	molecular volume VOL
5	NSUM
6	OSUM
7	n_{pos}
8	n_{neg}
9	$MEAN_+$
10	$MEAN_-$
11	σ^2_{tot}
12	ν

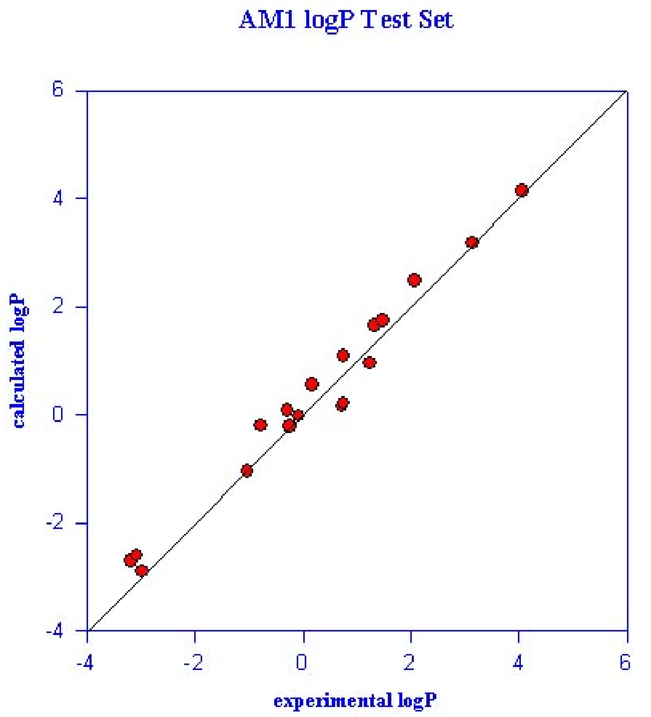


Figure 3. *Best predictions obtained for the small logP test set using a 12:4:1 neural net.*

Table 5. *t-values of regression equation (6).*

Input parameter	Coefficient	t-value
POL	0.063	9.140
$MEAN_+$	0.058	5.243
h6	0.0009	7.142
ESP_{min}	0.031	9.192
GLOB	1.296	2.680
NSUM	0.437	11.660
OSUM	0.815	13.664
PSUM	0.704	3.810
SSUM	0.437	7.899
ClSUM	2.286	7.142
BrSUM	5.883	5.727
ISUM	1.874	2.969

dation r_{cv} is 0.930. A small test set, consisting of 18 molecules, was selected in order to test the trained network. The correlation obtained can be seen in Figure 3.

The correlation coefficient r is 0.986, r^2 is 0.972 with a standard deviation of 0.339. The r_{cv} value is 0.958. The results of this initial investigations were thus very promising. We therefore used a set of 1085 organic compounds in order to refine our approach.

Large data set

Multiple linear regression analyses

With the experience of the initial approach, an enlarged data set with 579 organic compounds was selected covering an experimental logP range from -4 to 8. Parameters shown in Tables 1 and 2 were used for the multiple linear regression analysis. Again, the best results were obtained for the input parameters ESP_{max}, ESP_{min}, D_t, n_{pos}, n_{neg}, $MEAN_+$, $MEAN_-$, POL, SUR, VOL, GLOB, σ^2_+, σ^2_-, σ^2_{tot} , ν, NSUM, OSUM. Surprisingly, no square values were found. Moreover, all parameters with halogen charges were included. The following equation (6) results from the best linear regression:

$$\log P = 0.063\ POL - 0.058\ MEAN_+ + 0.0009\ h6 + 0.031\ ESP_{min} - 1.296\ GLOB + 0.437\ NSUM + 0.815\ OSUM + 0.704\ PSUM + 0.437\ SSUM + 2.286\ ClSUM + 5.883\ BrSUM - 1.874\ ISUM + 3.713 \quad (6)$$

$r = 0.866 \quad r^2 = 0.750$

The t-values are shown in Table 5. Figure 4 shows the regression results graphically.

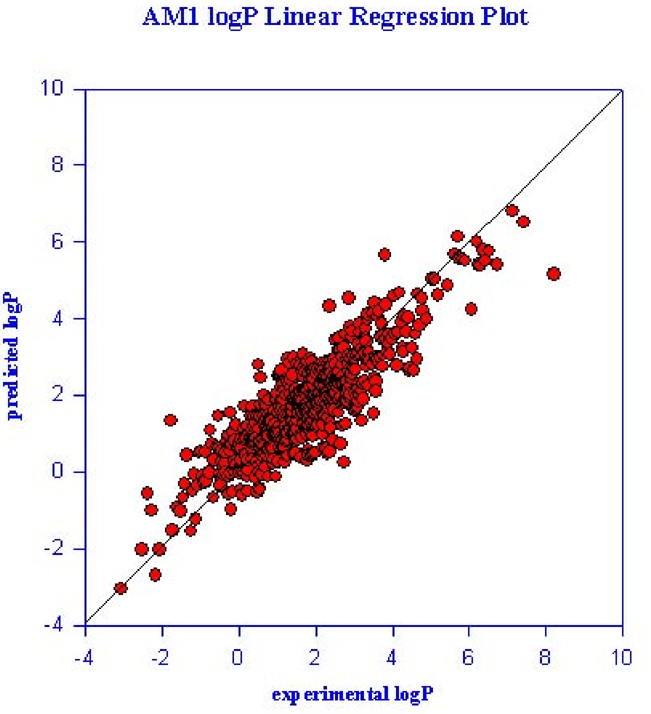


Figure 4. *Best predictions of the logP values by a multiple linear regression analysis (579 compounds).*

The correlation coefficient r is 0.866, r^2 is 0.750 with a standard deviation of 0.799.The r_{cv} value is 0.806. Again the linear regression is fairly accurate. In order to analyse the descriptor set in more detail, a principal component analysis was carried out.

Analysis of the principal components

A principal components analysis was performed within TSAR [27] in order to assess the major contributors to the total variance of the descriptor set. Overall, 16 parameters were used and 12 principal components were obtained that accounted for 100 % of the total variance. It is sufficient to concentrate on the three most important principal components, which contribute 28 %, 22 % and 10 % to the total variance, respectively. The parameters and components are shown in Table 6.

Principal component 1 has the largest influence on the regression's variance. The properties with the largest contributions have the largest impact on the prediction of the logP value. These are D_t, $MEAN_+$, $MEAN_-$, σ^2_{tot}, ν, ESP_{min}, ESP_{max},OSUM and SSUM. The second component adds POL, VOL, SUR, GLOB and NSUM. In the third component no further properties appear. The properties PSUM and HALSUM do not play any role in the first three principal components, but are clearly important for specific compounds. The correlation matrix as a supplemental tool was analysed in order to detect correlations between the different parameters. A direct correlation of the parameters VOL and SUR could be observed, suggesting that one parameter is sufficient for the back-propagation net.

Back-propagation networks

A training set of 980 organic compounds was selected randomly from the total set of 1085. 105 compounds were chosen for the test set. The number of descriptors for the input layer was varied between 16 and 25 and the number of nodes of the three layer back-propagation network between 300 and 500. The networks were trained until the RMS error fall below 4% of the logP range. The danger of overtraining the neural net was checked with the standard deviation of the test set. The best performance was obtained with an input layer of 16 parameters and a total net connectivity of 451 nodes (16-25-1 net). Two networks were trained for AM1 and PM3 data sets, respectively. The final descriptor set is shown in Table 7.

Although the descriptors SUR and VOL are highly correlated, it is surprising that the networks' performances depend on both parameters. The results obtained for the final network are illustrated in Figure 5.

The correlation coefficient r for the training set is 0.965, r^2 is 0.931 with a standard deviation of 0.41. The r_{cv} value is 0.930. In the case of PM3 the performance is slightly worse (r = 0.940, r^2 = 0.883, r_{cv} = 0.905 and the standard deviation

Table 6. *Principal components analysis.*

Variable	property	pc 1	pc 2	pc 3
x_1	D_t	0.353		
x_2	POL		0.504	
x_3	$MEAN_+$	0.397	-0.175	0.151
x_4	$MEAN_-$	-0.389	0.104	0.160
x_5	σ^2_{tot}	0.403		-0.244
x_6	ν	0.186	-0.119	0.490
x_7	ESP_{min}	-0.239	-0.149	0.548
x_8	ESP_{max}	0.422		0.170
x_9	VOL		0.506	0.106
x_{10}	SUR		0.519	0.115
x_{11}	GLOB		-0.251	
x_{12}	NSUM		-0.207	0.355
x_{13}	OSUM	-0.244	-0.174	-0.369
x_{14}	PSUM			
x_{15}	SSUM	0.233		0.132
x_{16}	HALSUM			

pc = principal component

Table 7. *Descriptor set for the neural networks.*

Number	Descriptor
1	total dipole moment D_t
2	mean polarizability POL
3	molecular surface SUR
4	molecular volume VOL
5	globularity GLOB
6	NSUM
7	OSUM
8	PSUM
9	SSUM
10	HalSUM
11	ESP_{max}
12	ESP_{min}
13	$MEAN_+$
14	$MEAN_-$
15	σ^2_{tot}
16	ν

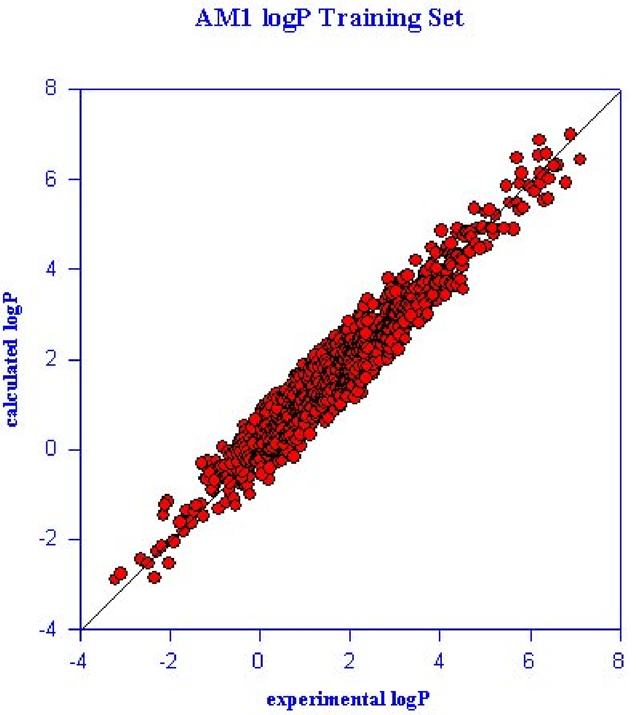


Figure 5. *AM1 training set.*

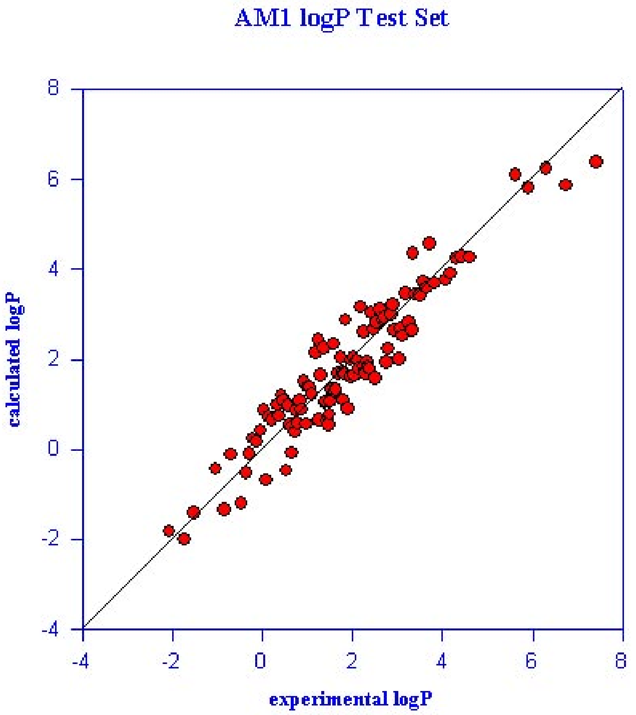


Figure 6. *AM1 test set.*

is 0.45). A test set containing 105 organic molecules was used in order to investigate the predictive power of the network. The result is shown in Figure 6.

A correlation coefficient r of 0.950, r^2 of 0.902, the cross validation r_{cv} of 0.915 with a standard deviation of 0.53 was obtained (maximum error 1.19). For the range of organic molecules within this data set, the predictions are very accurate. In the case of PM3, the performance of the network is slightly worse. The result is shown in Figure 7.

A correlation coefficient r of 0.910, r^2 of 0.830, the cross validation r_{cv} of 0.837 with a standard deviation of 0.67 (maximum error: 2.15) was obtained. Although the training performance is rather similar for the two semiempirical methods, it is surprising that there are such large deviations in the prediction for certain molecules in the case of PM3. In order to compare the efficiency of the two networks, all predicted logP values of the test set are given in Table 8.

While the performance of AM1 is quite accurate, there are some large deviations in the case of PM3. During a further investigation of test molecules not included in the original test set, a systematic weakness of our present networks was detected. LogP for compounds with large alkyl chains is estimated poorly. Some examples with a large calculated error are given in Scheme 1.

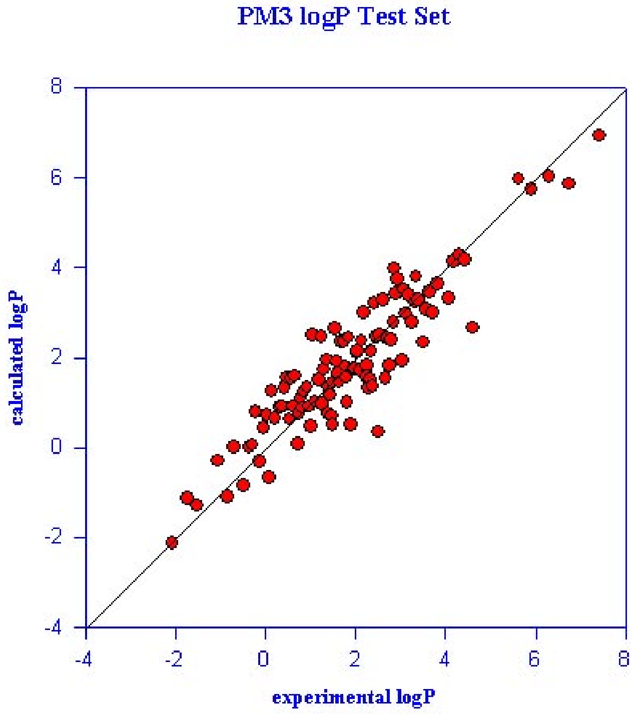


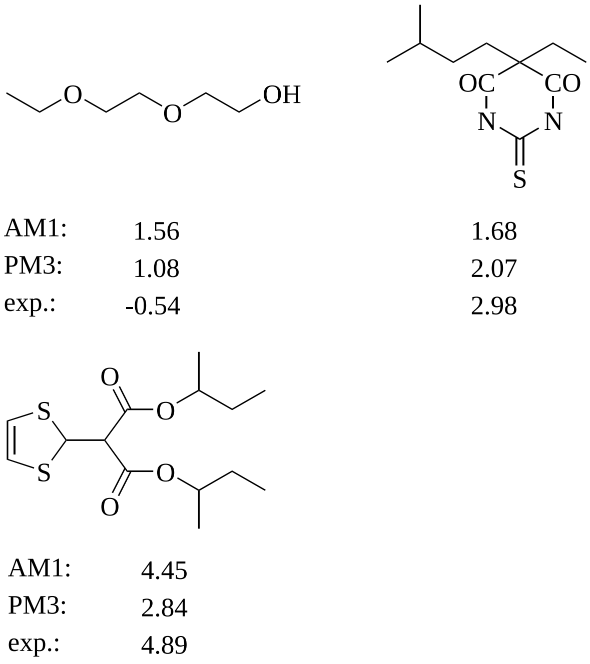
Figure 7. *PM3 test set.*

Table 8. *Table of test molecules with calculated AM1/PM3 logP values.*

	Compound	exp.	AM1	PM3
1:	N,N-Bis(2,3-dihydroxypropyl)-3-N-methyl-acetamido-2,4,6-triiodo-m-phthalamide	-2.06	-1.82	-2.13
2:	Citric acid	-1.72	-2.00	-1.12
3:	Phenylalanine	-1.52	-1.42	-1.29
4:	1,3-Propanediol	-1.04	-0.43	-0.30
5:	Maleic acid-hydrazide	-0.84	-1.34	-1.09
6:	N-Formyl-cyclobutane-carboxamide	-0.70	-0.10	0.00
7:	Nitrofurantoin	-0.47	-1.20	-0.83
8:	2,2-Dimethylpropionic acid-hydrazide	-0.35	-0.53	0.01
9:	3-Fluoropropanol	-0.28	-0.10	0.05
10:	2',3'-Didesoxyadenosine	-0.22	0.24	0.80
11:	3-Mesylphenyl-urea	-0.12	0.16	-0.31
12:	o-Methyl-THPO	-0.04	0.42	0.43
13:	5,6-Dihydro-2-methyl-1,4-oxathiin-3-carboxylic acid	0.04	0.86	0.71
14:	Mercapto-acetic acid	0.09	-0.68	-0.67
15:	Merbarone	0.14	0.72	1.26
16:	o-Methylbenzoyl-hydrazine	0.22	0.65	0.65
17:	Nikethamide	0.33	0.98	0.90
18:	2,2-Dichloroethanol	0.37	0.73	0.92
19:	1-Acetyl-N-(4-fluorophenyl)-hydrazine-carboxamide	0.42	1.19	1.31
20:	Piperazine-2-carboxanilide	0.48	1.07	1.55
21:	2-Nitro-p-phenylenediamine	0.53	-0.46	0.62
22:	2-Amino-5-methoxy-benzimidazole	0.57	0.95	1.52
23:	Glutaric acid-dimethylester	0.62	0.53	0.91
24:	3-(5-Nitro-2-furanyl)-2-propenoicamide	0.65	-0.08	1.59
25:	1-Acethyl-6-dimethyl-7-methoxymitosene	0.72	0.37	0.07
26:	2-Azacycloheptanthione	0.75	0.85	0.75
27:	N-(2-Benzoyl-oxyacetyl)-2-carboxyazetidine	0.79	0.58	1.08
28:	Chloropentazide	0.84	1.08	0.88
29:	4-Pyridine-butaneamine	0.86	0.87	1.22
30:	2-Iodo-benzamide	0.93	1.51	1.33
31:	4-Methylthiazole	0.97	0.55	0.90
32:	6-Cyanoquinoxaline	1.01	1.38	0.47
33:	1-Phenyl-3-cyanoguanidine	1.05	1.37	2.50
34:	m-Acetylamino-acetophenone	1.10	1.21	1.02
35:	Acetylsalicylic acid	1.19	2.14	1.50
36:	2-Nitrobenzyl-alcohol	1.24	2.43	2.47
37:	Benzaldehyde-semicarbazone	1.27	0.65	0.97
38:	4-Oxo-4-phenyl-butanoic acid	1.30	1.64	1.73
39:	2-Phenyl-ethanol	1.36	2.25	1.95

40:	3-Bromo-benzenesulfonamide	1.39	1.04	0.74
41:	Bromochloromethane	1.41	1.05	1.33
42:	2-Imino-3-methyl-5-(5-nitro-2-furfurilidine)-thiazoline-4-one	1.44	0.62	1.16
43:	Trimethylacetic acid	1.47	0.54	0.69
44:	o-Fluorophenylacetic acid	1.50	0.78	0.49
45:	2-(2,6-Dichloro-4-hydroxy-phenylimino)-imidazolidine	1.52	1.06	1.44
46:	N-Phenyl-4-aminophenylsufonamide	1.55	1.32	2.63
47:	N,N-Dimethylcarbamate-p-(n,n-dimethylcarbamate)-benzylester	1.59	2.34	1.90
48:	2-Methylquinoxaline	1.61	1.28	1.65
49:	3,5-Dimethoxyphenol	1.64	1.32	1.44
50:	Indole-3-ethanolcarbamate	1.69	1.67	2.34
51:	3-Indolylpropionic acid	1.75	2.03	2.34
52:	Propylene	1.77	1.72	1.80
53:	2-Oxoisopropyl-5-phenyl-5'-ethylbarbituric acid	1.79	1.08	1.55
54:	4-Dimethylamino-thieno(2,3-D)-pyrimidine	1.82	1.68	1.00
55:	2-Acetyl-oxyethyl-benzoic acid-ester	1.85	2.87	2.45
56:	2-nitro-5-fluorophenol	1.91	0.90	0.51
57:	N-Methyl-2,3-dimethylphenyl-carbamate	1.95	1.98	1.76
58:	o-Methylphenoxy-acetic acid	1.98	1.59	1.77
59:	Acetic acid-m-methoxybenzylester	2.02	2.04	2.09
60:	1,1'-Dioxo-3-cyclohexen-3-yl-1,2,4-benzothiadiazine	2.05	1.64	2.14
61:	3,4-Dimethylacetanilide	2.10	1.98	1.71
62:	Indole	2.14	1.78	2.37
63:	1,2-Dinitro-4-chlorobenzene	2.18	3.15	3.00
64:	1-Hydroxyethyl-2-styryl-5-nitroimidazole	2.25	2.59	1.82
65:	o-Methyl-benzaldehyde	2.26	1.84	1.57
66:	2,6-Dimethoxypyridine	2.30	1.68	1.33
67:	Thiophene-2-carboxylic acid-ethylester	2.33	1.95	1.52
68:	21-Desoxybetamethasone	2.35	1.82	2.13
69:	Thiosalicylic acid	2.39	1.79	1.36
70:	1-Pyrrol-2-yl-pentanone	2.42	3.03	3.21
71:	5,5'-Diphenyl-hydantoin	2.47	2.66	2.45
72:	8-Trifluoromethyl-quinoline	2.50	1.58	0.34
73:	2,17-dihydroxy-3-oxolactone-7,21-dicarboxy-pregan-4-ene	2.54	2.81	2.50
74:	N-Benzyl-N-formylaniline	2.62	3.11	3.29
75:	2-Ethyl-4,6-dinitrophenol	2.67	2.88	1.54
76:	5,6-Diazaphenanthrene	2.71	2.92	2.41
77:	1-Methyl-1,3-dihydro-5-(2-fluorophenyl)-7-chloro-1,4-benzodiazepin-2-one	2.75	1.92	1.83
78:	3-Butyl-RS-1(3H)-isobenzofuranone	2.80	2.23	2.38
79:	2-Anilino-1,4-Naphthoquinone	2.84	3.02	2.78
80:	4-Aminobiphenyl	2.86	2.99	3.97
81:	Dihydromorphanthridine	2.90	3.21	3.42
82:	p-Phenoxy-aniline	2.93	2.64	3.74
83:	Octanoic acid	3.05	1.99	1.92

84:	Deoxycorticosterone-acetate	3.08	2.67	3.51
85:	3,4-Dichloronitrobenzene	3.12	2.51	2.98
86:	N-(3,4-Dichlorophenyl)-difluoroacetamide	3.18	3.46	3.39
87:	Iodobenzene	3.25	2.81	2.79
88:	1-(3,4-Dichlorophenyl)-2-isopropylaminoethanol	3.32	2.65	3.23
89:	3-Methoxy-4-cyclohexyl-methoxy-phenylacetic acid	3.35	4.34	3.79
90:	Anthraquinone	3.39	3.43	3.28
91:	Prometrin	3.51	3.43	2.33
92:	4,7-Dichloroquinoline	3.57	3.73	3.05
93:	9-(N-((N,N'-Diethylamino)acetyl)amino)-fluorene	3.64	3.57	3.45
94:	Indigo	3.72	4.56	3.00
95:	3,4-Dimethylchlorobenzene	3.82	3.69	3.64
96:	1-(4-Cyclohexylphenyl)-3-methoxy-3-methylurea	4.08	3.76	3.30
97:	Propanoic acid-(1-phenyl-1-benzyl-2-methyl-3-(n,n-dimethylamino))-propylester	4.18	3.90	4.13
98:	2,6-Dimethylnaphthalene	4.31	4.24	4.27
99:	1,3-Dimethylnaphthalene	4.42	4.27	4.17
100:	Propanoic acid-1,3-dithiolan-2-ylidine-dibutylester	4.60	4.26	2.65
101:	2,4,4'-Trichlorobiphenyl	5.62	6.09	5.97
102:	2,4,5-Trichlorobiphenyl	5.90	5.81	5.74
103:	1,3,7,8-Tetrachlorodibenzodioxin	6.30	6.24	6.03
104:	1,2,3,6,7-Pentachlorodibenzodioxin	6.74	5.86	5.86
105:	3,3',4,4',5,5'-Hexachlorobiphenyl	7.41	6.38	6.93



Scheme 1. *Molecules with large errors.*

Dependence of the network's predictive power on the descriptors

The dependence of the predicted AM1 and PM3 logP values was investigated by systematically changing the input descriptors. A compound with a small error (acetophenone) was selected for that purpose. With the exception of the parameters polarizability POL, balance ν and charge OSUM (Table 2), only small effects on the predicted logP value resulting from the change of the AM1/PM3 calculated input parameters were found. Changing the calculated value of the parameters ν and POL has a dominant influence on the prediction of logP. Figure 8a shows the dependence of the calculated logP on the descriptor polarizability, POL.

A linear relationship between the logP values and the change of the parameter is evident, with a larger slope for AM1. The effects for the parameter "Balance" are illustrated in Figure 8b.

In this case the network's reaction is nonlinear. Small changes of the calculated value (AM1 and PM3) of the balance cause a positive change in the predicted logP. It is interesting to note that the original value of ν represents the minimum of the curve.

Another sensitive parameter for the evaluation of logP is the descriptor OSUM. The result is illustrated in Figure 8c.

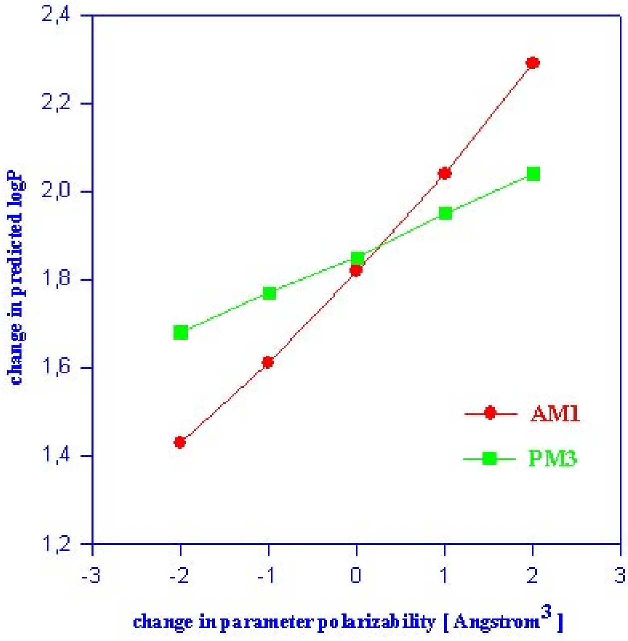


Figure 8a. *Dependence of the AM1/PM3 predicted logP value of acetophenone on the polarizability. The "change of parameter" axis indicates that the numerical value of the parameter was changed by the given amount.*

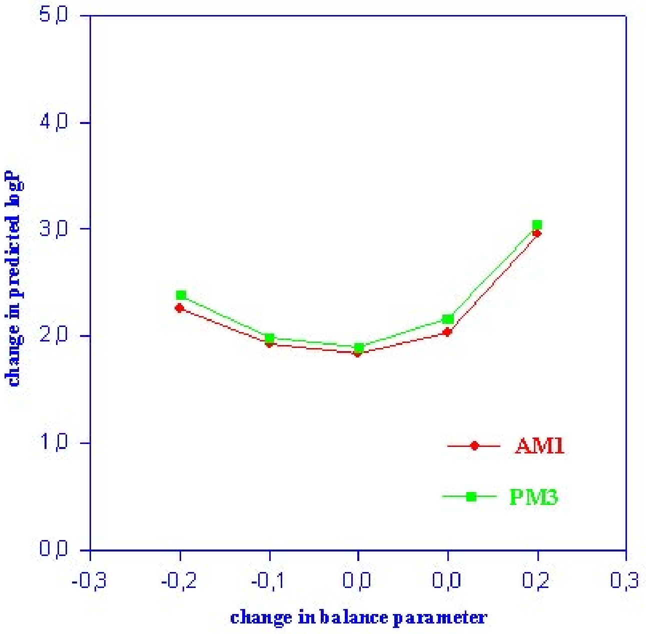


Figure 8b. *Dependence of the AM1/PM3 predicted logP values of molecule acetophenone on the balance. The "change of parameter" axis indicates that the numerical value of the parameter was changed by the given amount.*

While the predicted logP value is nearly constant when the PM3 charge is varied, the effect on the AM1 predicted logP is enormous. Increasing the charge strongly increases the predicted logP, whereas in the reverse direction a limiting value of 1.70 is approached.

The different behaviour of the AM1 and PM3 nets for the descriptor OSUM is perhaps the source of the worse performance of PM3. The same behaviour was found for NSUM and HalSUM. The PM3 net seems to be less sensitive to changes in the atomic charges. The charge distribution within a molecule depends strongly on the molecule's conformation. Consequently, PM3 reacts far less sensitively to conformationally free molecules leading to a worse calculation of logP for bulky molecules (Scheme 1). On the other hand, AM1 is well suited for molecules with less conformational freedom. AM1 calculated Electrostatic Potential-Derived Atomic Charges (VESPA) [37] agree better than PM3 with *ab initio* calculated charges, which helps explain the better performance of AM1 in logP calculations.

Summary and conclusion

Using semiempirically calculated molecular and atomic properties and a back-propagation network is a promising technique for the prediction of logP values. The correlation coefficients and standard deviations obtained are nearly as good as those obtained using ClogP [38] without having problems with unknown fragments and without using any correction factors.The current method is more general than that of Herges et al. [9], because the descriptors are calculated within one gas phase geometry optimisation, whereas the technique of Herges is dependent on several calculation steps. Furthermore, the networks are able to handle a large spectrum of organic compounds, making them compatible with large data base systems. The observed weakness in the prediction of bulky molecules seems to be a computational problem in the derivation of the appropiate conformation from gas phase optimized geometries. Of course, techniques such as that presented here can only work properly for compounds that exist in the same structural form in water and n-octanol and may therefore fail for sugars and ionisable compounds. Further investigations will improve the predictive power of this approach. On the other hand, our approach is accurate enough to be included as a general application for use with the semiempirical program VAMP 6.5 [39].

Acknowledgements. This work was supported by the Deutsche Forschungsgemeinschaft and the Wellcome Foundation Ltd., Langley Park, Beckenham, Kent, BR3 3BS, U. K. We thank A. Rosner for initial investigations.

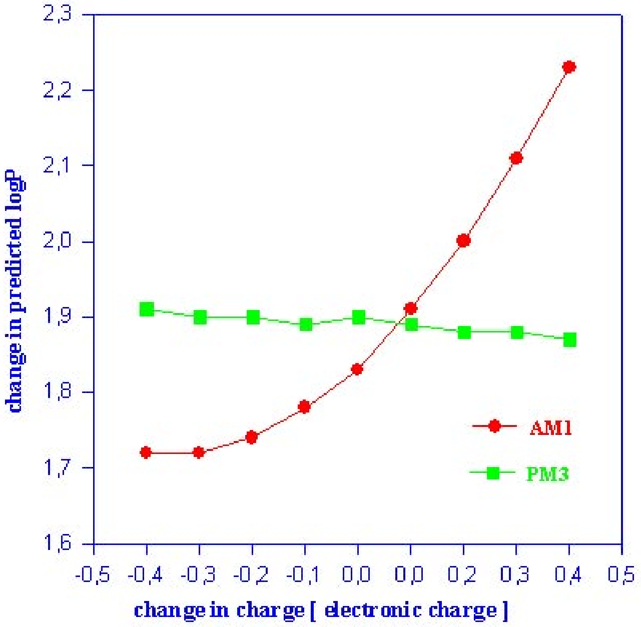


Figure 8c. *Dependence of the AM1/PM3 predicted logP values of molecule acetophenone on the oxygen charge. The "change of parameter" axis indicates that the numerical value of the parameter was changed by the given amount.*

References

1. Rauhut, G.; Clark, T. *J. Am. Chem. Soc.* **1993**, *115*, 4698.
2. (a) Lyman, W. J. In *Handbook of Chemical Property Estimation Methods: Environmental Behaviour of Organic Compounds*; Lyman, W. J., Ed.; American Chemical Society: Washington, DC, 1990, pp 1-50. (b) Sablijic, A.; Guesten, V.; Hermans, J.; Opperhuizen, A. *Environ. Sci. Technol.*, **1993**, *27*, 1394. (c) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979.
3. Suzuki, T.; Kudo, Y. J. *J. Comput.Aided Mol. Des.* **1990**, *4*, 155.
4. (a) Ghose, A. K.; Crippen, G. M. *J. Comp. Chem.* **1986**, *7*, 565. (b) Viswanadhan, V. N.; Ghose, A. K.; Ravankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163. (c) Klopman, G.; Wang, S. *J. Comp. Chem.* **1991**, *12*, 1025.
5. Rekker, R. E. *The Hydrophobic Fragment Constant*, Elsevier, Amsterdam, 1976.
6. Leo, A. J.; Jow, P. Y. C.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1975**, *18*, 865.
7. Meylan, W. M.; Howard, P. H. *J. Pharm. Sci.* **1995**, *84*, 83.
8. Bodor, N.; Huang, M. J. *J. Pharm. Sci.* **1995**, *81*, 272.
9. Grunenberg, J.; Herges, R. *J. Chem. Comput. Sci*. **1995**, *35*, 905.
10. Essex, J. W.; Reynolds C. A.; Richards W. G. *J. Am. Chem. Soc.* **1992**, *114*, 3634.
11. Giesen, D. J.; Gu M. Z.; Cramer C. J.; Truhlar D. G. *J. Org. Chem.* **1996**, *61*, 8720.
12. Klamt A. *J. Phys. Chem*. **1995**, *99*, 2224.
13. Pearlman, R. S.; Balducci, R.; Rusinko, A.; Skell, J. M.; Smith, K. M. CONCORD; available from Tripos Associates Inc., St. Louis, MO.
14. SYBYL 6.2, Tripos Associates, St. Louis, Mo., USA, 1994
15. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
16. (a) Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 209. (b) Stewart, J. J. P. *ibid* 221.
17. Rauhut, G.; Alex, A.; Chandrasekhar, J.; Steinke, T.; Sauer, W.; Beck, B.; Hutter, M.; Clark, T. VAMP 6.0, Oxford Molecular Ltd., Medawar Centre, Oxford Science Park, Sand ford-on-Thames, Oxford, OX4 4GA, England.
18. Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118.
19. Meyer, A. Y. *J. Chem. Soc. Rev.* **1986**, *15*, 449.
20. Beck, B.; Glen, R. C.; Clark, T. *J. Mol. Model.* **1995**, *1*, 176.
21. Rauhut, G.; Clark, T. *J. Comp. Chem.* **1993**, *14*, 503.
22. Beck, B.; Rauhut, G.; Clark, T. *J. Comp. Chem.* **1994**, *15*, 1064.
23. (a) Heiden, W.; Goetze, T.; Brinckmann, J. *J. Comp. Chem.* **1993**, *14*, 246. (b) Lorenzen, W.; Cline, H. *Comp. Graph.* **1987**, *21*, 163.
24. (a) Murray, J. S.; Lane, P.; Brinck, T.; Politzer, P. *J. Phys. Chem.* **1993**, *97*, 5144. (b) Murray, J. S.; Lane, P.; Brinck,T.; Paulsen, K.; Grice, M. E.; Politzer, P. *J. Phys. Chem.* **1993**, *97*, 9369.
25. Draper, N. R.; Smith, H. *Applied regression analysis*, 2nd Edition, John Wiley & sons, New York, 1981.
26. Müller, B.; Reinhardt, J. *Neural Networks. An Introduction*, Springer, Heidelberg, 1990.
27. TSAR 2.1, Oxford Molecular Ltd., Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford, OX4 4GA, England.
28. Clark, T.; Rauhut, G.; Breindl, A. *J. Mol. Model.* **1995**, *1*, 22.
29. So, S.-S.; Richards, W. G. *J. Med. Chem.* **1995**, *35*, 3201.
30. Oinuma, H.; Miyako, K.; Yamanaka, M.; Nomoto, K.-I.; Katoh, H.; Sawada, K.; Shino, M.; Mamano, S. *J. Med. Chem.* **1990**, *33*, 905.
31. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 2583.
32. Lacy, M. E. *Tetrahedron Computer Methodology* **1990**, *3*, 119.
33. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *Chem. Pharm. Bull.* **1989**, *37*, 2558

34. Bodor, N.; Harget, A.; Huang, M.-J. *J. Am. Chem. Soc.* **1991**, *113*, 9480.
35. Egolf, L. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616.
36. Saic ANSim, Saic, 10260 Campus Point Drive, MS71, San Diego, CA 92121, USA, 1989.
37. Beck, B.; Glen, R. C.; Clark, T. *J. Comp. Chem.* **1997**, in the press.
38. Pomona89 Physico-Chemical Database & Medchem Software version 3.54, Daylight Chemical Information Systems Inc., Claremont, CA.
39. Rauhut, G.; Alex, A.; Chandrasekhar, J.; Steinke, T.; Sauer, W.; Beck, B.; Hutter, M.; Gedeck, P.; Clark, T. VAMP 6.5, to be available from Oxford Molecular Ltd. in 1997.