Prediction of Chromatographic Retention Values (R_M) and Partition Coefficients (log P_{oct}) Using a Combination of Semiempirical Self-Consistent Reaction Field Calculations and Neural Networks

J. Grunenberg and R. Herges*,†

Institut für Organische Chemie, Universität Erlangen-Nürnberg, Henkestrasse 42, 91054 Erlangen, Germany

Received June 2, 1995®

A combination of semiempirical solvent effect calculations and artificial neural networks was used to predict the (reversed phase) retention values (R_M) of steroids and the partition coefficients (log P_{oct}) of a diverse set of organic compounds. Eleven selected physical parameters from AM1 self-consistent reaction field calculations were used as input for neural networks of the back-propagation type. The performance (standard error 0.08 R_M units) in predicting retention values of a set of steroids is close to experimental accuracy (0.03 R_M units). The partition coefficients of a heterogeneous set of organic compounds are predicted with a standard error of 0.29 log P_{oct} units. Systematic leave-n-out experiments revealed that the solvation energy obtained by simple solvent calculations (spherical model) is the most important parameter in our 11 parameter set and considerably improves the performance in predicting hydrophobicity at little additional computational cost.

INTRODUCTION

Methods to predict retention values in reversed phase chromatography are useful from different points of view in order (1) to find a target compound with unknown retention in a chromatogram of complex product mixtures, (2) to find a suitable solvent system for separation problems, and (3) to calculate the aqueous solubility of compounds, which is a very important parameter in drug design or pesticide development. Usually the aqueous solubility of compounds is expressed by the logarithmic value of the partition coefficient between *n*-octanol and water (log P_{oct}). Experimentally this parameter is determined via the shakeflask method, which requires large amounts of pure samples. Problems arise in the measurement of log P_{oct} values below −2 or above 4. Therefore reversed phase chromatographic procedures with octanol-like stationary phases were developed, that are more convenient to apply and which provide more accurate results in case of a very high or very low solubility in both solvents. There is a linear relation between R_M and log P_{oct} values

$$R_M = a + b \log P_{oct} \tag{1}$$

R_M values were introduced by Martin for thin layer chromatography

$$R_M = \log k = \log\left[\left(\frac{1}{R_F}\right) - 1\right] \tag{2}$$

R_F is called the retardation factor and has the following definition

$$R_F = \frac{\text{distance migrated by an analyte}}{\text{distance migrated by the solvent front}} \tag{3}$$

Chromatographic retention values and solubilities are macroscopic properties and thus a complex (and in principle unknown) function of a number of microscopic, molecular properties. Various QSAR studies, using both statistical methods as well as neural nets, have been performed to deduce the aqueous solubility of organic compounds. Most of these approaches use group increments or topological features.[1] More straightforward and founded on a stronger physical basis are methods that use empirical molecular properties,[2] parameters derived from semiempirical quantum chemical,[3−10] or ab initio[11,12] quantum chemical calculations. Viswanadhan et al. compared the performance of three different approaches.[13] In the present study we derived R_M and log P_{oct} values from solvent (self-consistent reaction field) AM1 calculations using back-propagation neural networks as nonlinear mapping devices.[14−16]

Physical Interpretation of the Descriptors. The proper choice of input parameters for the neural network is of critical importance from a practical as well as an intellectual point of view. The physical processes of solvation should be modeled as accurately as possible.[17] However, for practical purposes a compromise has to be found between computational cost and accuracy. We used the semiempirical AM1[18] method for geometry optimization. Solvent effects were treated with the self-consistent reaction field (SCRF) approach[19,20] implemented in the VAMP 4.5[21] program package. In this approach the solvent is described as a dielectric continuum, the solute being placed into a spherical cavity. The charge distribution of the solute creates an electric field in the continuum, which reacts on the solute and vice versa. The charge distribution of the solute is represented as a multipole expansion. The perturbation of the solute energy relative to the unperturbed molecule is given by

$$U_d = -\frac{1}{2}\sum_{l=0}^{\infty}\sum_{m=-1}^{1} R_l^m M_l^m \tag{4}$$

R_l^m: components of the reaction field

M_l^m: components of the multipole of first order

† FAX 0049 9131 85 9132; E-mail herges@organik.uni-erlangen.de.

® Abstract published in *Advance ACS Abstracts,* September 1, 1995.

0095-2338/95/1635-0905$09.00/0

Table 1. Compilation of the Descriptors Used in This Study

descriptor	physical interpretation
Solv	calculated perturbation in H_2O (spherical SCRF-model)
Hd	sum of the Mulliken charges on hydrogen atoms; ability to act as H donor
Ha	sum of the Mulliken charges on heteroelements; ability to act as H acceptor
Vib	sum of calculated harmonic frequencies ≤ 200 cm^{-1}; conformational flexibility
Ent	calculated entropy (300 K)
Epot−	electrostatic potential—minimum; dipole/dipole interactions
Epot+	electrostatic potential—maximum; dipole/dipole interactions
D	permanent dipole moment; dipole/dipole interactions
P	polarizability; dipole/induced dipole interactions
SA	surface area; cavity energy
V	volume of the molecule; cavity energy

The calculated perturbation energy U_d from eq 4 was used as the first input parameter (Solv, see Table 1) for our neural net studies. The cavity was calculated using the spherical model developed by Rivail and Rinaldi.[22] The algorithm is very efficient (only slightly more time consuming than the geometry optimization in vacuo); however, it is based on the simple assumption that the solvent cavity can be approximated by a globe. This is a crude approximation in some cases (e.g., steroids). To compensate for this oversimplification, we therefore added the molecular surface area[23,24] (SA) and the molecular volume[25] (*V*) calculated by the marching cube algorithm of Marsili[26] as further input parameters.

Additional parameters were introduced to account for both ion/dipole, dipole/dipole, (permanent dipole moment *D*), and dipole/induced dipole interactions (polarizability *P*). The sum of Mulliken charges on hydrogen atoms (Hd) and heteroelements (Ha) accounts for hydrogen bonding and electron pair donor/acceptor interactions, that are not recognized within our continuum SCRF model. The maximum (Epot+) and minimum (Epot−) electrostatic potential[27] was calculated using the NAO-PC[28,29] method, which is also implemented in the VAMP 4.5 program.

Although neglected in most approaches, the conformational flexibility of the solute is important in the description of the solvation process. Flexible molecules (e.g., crown ethers or proteins) are able to maximize favorable interactions by orientating their nonpolar molecular sites away from the polar solvent molecules and by stretching polar groups toward the surrounding solvent. Particularly in solvents of high polarity, flexible molecules change conformation in such a way as to move their hydrophilic parts toward the molecular surface and to "hide" their hydrophobic groups in the "interior". Conformational changes can have a large impact on solubility and partition coefficients. In order to obtain a parameter describing the conformative flexibility, harmonic frequency calculations were carried out for each molecule of the dataset. The harmonic frequencies of vibrations lower than 200 cm^{-1} were added up and used as a further input parameter (Vib) for the neural net. These frequencies are excited at ambient temperatures to a considerable extent and might account for conformational processes.

In order to model the chromatographic partition process, entropic effects have to be considered. The analytes partition between the stationary and the mobile phase. The sorption−desorption process occurs many times during the chromatographic procedure. When entering the stationary phase, the molecules lose degrees of freedom and their entropy decreases; the larger the change in entropy, the smaller the retention. Rigid molecules such as fluorene (**1**)show much

(1) (2)

larger retention than structurally related but nonrigid compounds, such as diphenylmethane (**2**).[30] We calculated the entropy by thermochemical analysis of the vibrational frequencies and obtained another input parameter (Ent). Together, these 11 microscopic quantum chemical parameters were used to train neural networks in order to map them on a single macroscopic value, the hydrophobicity in terms of R_M or log P_{oct} values. We tried to keep the number of input parameters as small as possible, so we discarded higher orders of single input parameters (SA^2, SA^3, V^2, V^3, ...). In preceding studies, they turned out to make the neural net more susceptible to overtraining without improvement of the general performance.

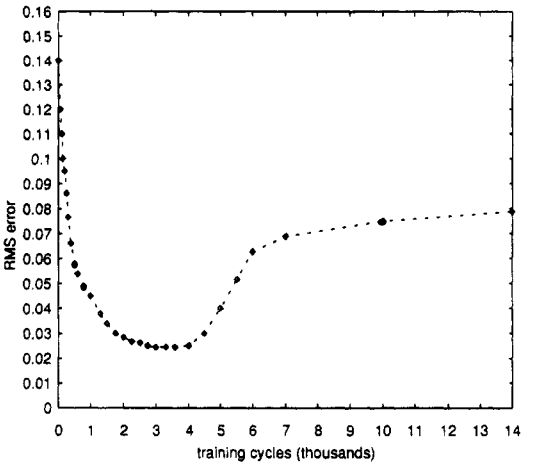


Figure 1. RMS error in the test set as a function of training epochs.

COMPUTATIONAL DETAILS

The starting geometries were obtained using PCMODEL.[31] This program uses the MMX force field of Gajewski and Gilbert. MMX is based on the MM2 force field of Allinger and co-workers.[32] After force field geometry preoptimization the files were converted into the VAMP 4.5 format. We used the AM1 Hamiltonian for the SCRF geometry optimization with a dynamic cavity (the cavity size is updated during the optimization). The multipole expansion was of third order (octapol). Harmonic force constants and vibrational frequencies were evaluated via the force[33] method at the equilibrium molecular geometry.

From the results of these calculations the parameters for network training were derived. A standard back-propagation neural network[34] with an input layer of 11 neurons, one hidden layer (three neurons) and one neuron as the output layer was used for mapping the input parameters onto R_M or log P_{oct} values. The number of hidden neurons was optimized by trial and error. In each case, the training was continued until the minimum in the RMS error curve for the testset (Figure 1) was achieved in order to avoid

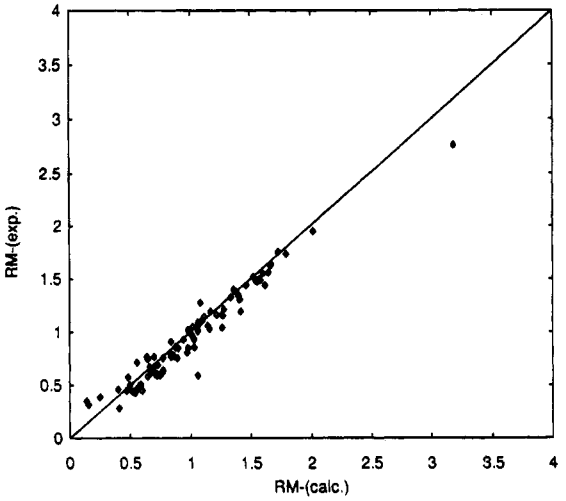


Figure 2. Calculated (neural network) R_M values of the training set (85 steroids) compared to experimental data.

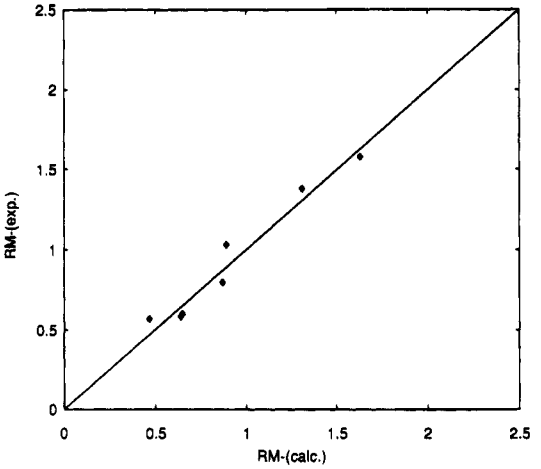


Figure 3. Calculated (neural network) R_M values of one of the test sets (seven selected steroids) compared to the experimental data.

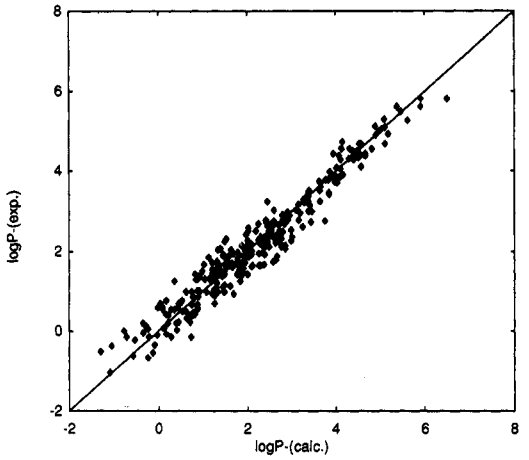


Figure 4. Calculated (neural network) log P_{oct} values of the training set (302 organic compounds) compared to experimental data.

overtraining. The input data were normalized between −1 and +1.

RESULTS AND DISCUSSION

The molecules in the first part of our investigation were taken from a steroid dataset.[35] They can be divided into four classes: (I) 16a-substituted estrogen-1,3,5(10)-triene; (II) 17a-CH_2X-substituted estrogen-1,3,5(10)-triene; (III) 17b-carbamoyloxy-estrogen-1,3,5(10)-triene, structure I; and (IV) 19-nortestosteron derivatives, structure II. The training set

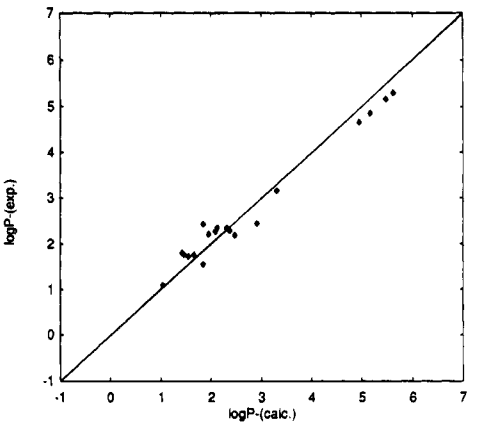
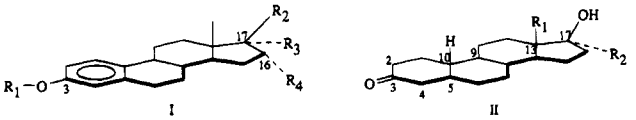


Figure 5. Calculated (neural network) log P_{oct} values of one of the test sets (21 organic compounds) compared to the experimental data.



was comprised of 85 steroids with various heterosubstituents (see Table 2) covering R_M values from 0.14 units up to 3.18. In fitting the training set a standard deviation of 0.11 R_M units (Figure 2) was achieved. We checked the influence of a single steroid (no. 58, see Table 2), which seemed to have a disproportionately high value of 3.18 R_M units (see also Figure 2). By leaving this compound out, the quality of the training fit changed imperceptibly. The reasons for the high retention and hence the low hydrophilicity are reflected in the numeric values of the calculated input parameters. Compared to the calculated descriptors of a steroid of approximately the same geometric data but with a much higher solubility in polar solvents (1.46 R_M units), the most striking difference (Table 3) is found in the numeric values of descriptors Ha and Vib. Steroid 58 has a lower H acceptor strength and a lower conformative flexibility than other steroids of similar size.

The trained net was then tested in its ability to calculate R_M values of seven selected steroids, that were not included in the training set. The results of the prediction (standard deviation of 0.08 R_M units) are summarized in Table 4 and plotted as a function of the experimental data in Figure 3. The test steroids were chosen to cover the whole variable space, that means aromatic and aliphatic as well as highly substituted and monosubstituted derivates were included. The accuracy of the experimental R_M values is given to ±0.03 R_M units.

In order to check the stability of our network, we exhaustively interchanged the steroid input data between training and test set. The highest standard deviation observed during this cross-validation process was 0.09 R_M values in the test set.

To obtain information about the relative importance of the individual input parameters, the internal validation in our network was tested by using a leave-one-out method.[3] Training was continued until the minimum in the RMS error curve of the test set (Figure 1) was achieved. A remarkable increase in standard deviation was observed when the

Table 2. Experimental and Calculated R_M Values of 85 Steroids of the Training Set

no.	R1	R2	R3	R4	R_M (exp)	R_M (calc)	ΔR_M
			Structure I				
1	CH_3	OH	H	H	0.86	0.78	0.08
2	CH_3	OH	H	OH	0.52	0.44	0.08
3	CH_3	OH	H	Br	0.98	1.02	−0.04
4	CH_3	OH	H	N_3	0.97	0.81	0.16
5	CH_3	OH	H	SCN	0.84	0.91	−0.07
6	CH_3	OH	H	SC_2H_5	1.02	0.96	0.06
7	CH_3	OH	H	$NHCOCH_3$	0.5	0.50	0.00
8	CH_3	OH	H	NCS	1.22	1.16	0.06
9	CH_3	OH	H	CH_2OH	0.59	0.49	0.10
10	CH_3	OH	H	SO_2CH_2Ph	0.94	0.93	0.01
11	CH_3	OH	H	SH	0.88	0.85	0.03
12	CH_3	OH	H	SCH_2Ph	1.40	1.33	0.07
13	H	OH	H	H	0.54	0.43	0.11
14	H	OH	H	OH	0.16	0.32	−0.16
15	H	OH	H	Br	0.73	0.69	0.04
16	H	OH	H	N_3	0.66	0.68	−0.02
17	H	OH	H	SCN	0.56	0.72	−0.16
18	H	OH	H	SCH_2H_5	0.70	0.77	−0.07
19	H	OH	H	$NHCOCH_3$	0.14	0.35	−0.21
20	CH_3	H	H	H	1.58	1.49	0.09
21	CH_3	H	H	Br	1.60	1.55	0.05
22	CH_3	H	H	N_3	1.54	1.47	0.07
23	CH_3	H	H	SCN	1.27	1.16	0.11
24	CH_3	H	H	$NHCOCH_3$	0.77	0.64	0.13
25	CH_3	H	H	SH	1.42	1.19	0.23
26	CH_3	OH	CH_3	H	0.89	0.75	0.14
27	CH_3	OH	CH_2CN	H	0.77	0.75	0.02
28	CH_3	OH	CH_2OH	H	0.59	0.51	0.08
29	CH_3	OH	CH_2Br	H	1.03	0.93	0.10
30	CH_3	OH	$CH_2NHCOCH_3$	H	0.47	0.45	0.02
31	CH_3	OH	CH_2OCH_3	H	0.84	0.76	0.08
32	CH_3	OH	CH_2Cl	H	0.99	1.00	−0.01
33	CH_3	OH	CH_2N_3	H	1.02	1.05	−0.03
34	CH_3	OH	CH_2OCOCH_3	H	0.77	0.64	0.13
35	CH_3	OH	$CH_2OC_2H_5$	H	0.98	0.85	0.13
36	CH_3	OH	CH_2NHCH_3	H	1.14	1.07	0.07
37	CH_3	OH	$CH_2NPhCOCH_3$	H	1.06	1.01	0.05
38	CH_3	OH	CH_2-R*	H	1.01	0.98	0.03
39	CH_3	OH	$CH_2O(CH_2)_2OH$	H	0.65	0.74	−0.09
40	CH_3	OH	$CH_2S(CH_2)_2OH$	H	0.72	0.60	0.12
41	CH_3	OH	CH_2NHCH_2Ph	H	1.28	1.21	0.07
42	CH_3	OH	CH_2SCH_2Ph	H	1.36	1.40	−0.04
43	H	OH	CH_2SCN	H	0.65	0.58	0.07
44	H	OH	CH_2Br	H	0.77	0.62	0.15
45	H	OH	CH_2CN	H	0.48	0.57	−0.09
46	H	OH	CH_22N3	H	0.75	0.60	0.15
47	H	OH	C≡CH	H	0.56	0.48	0.08
48	CH_3	OTHP**	CH_2SCNH	H	1.38	1.37	0.01
49	CH_3	OTHP	CH_2OCH_3	H	0.90	0.85	0.05
50	CH_3	OTHP	CH_2CN	H	1.11	1.15	−0.04
51	CH_3	$OCOC_2H_5$	CH_2N_3	H	1.57	1.50	0.07
52	CH_3	$OCOCH_3$	CH_2SCN	H	1.26	1.17	0.09
53	CH_3	$OCOCH_3$	CH_2OCH_3H	H	1.06	1.60	0.00
54	CH_3	$OCOCH_3$	CH_2CN	H	1.06	1.09	−0.03
55	CH_3	$OCOCH_3$	CH_2N_3	H	1.41	1.31	0.10
56	H	$OCOCH_3$	CH_2N_3	H	1.03	0.86	0.17
57	CH_3	$OCONHN(CH_3)_2$	H	Br	1.08	1.28	−0.2
58	CH_3	OCONH-cC_6H_{11}	H	Br	3.18	2.76	0.42
59	CH_3	OCONH-NHTs***	H	Br	1.46	1.44	0.02
60	CH_3	$OCONHCH_2Ph$	H	Br	1.66	1.63	0.03
61	CH_3	$OCONHN{=}C(CH_3)_2$	H	Br	1.02	0.96	0.06
62	CH_3	OCONH-NHPh	H	Br	1.52	1.52	0.00
63	CH_3	OCONHPh	H	Br	1.73	1.75	−0.02
64	CH_3	$OCONH_2$	H	Br	1.06	0.60	0.46
65	CH_3	OCOCl	H	Br	1.67	1.64	0.03
66	CH_3	$OCONHCH_2COOEt$	H	Br	1.34	1.32	0.02
67	CH_3	OCOOEt	H	Br	1.62	1.44	0.18
68	CH_3	$OCON_3$	H	Br	1.56	1.47	0.09
69	CH_3	$OCONHN{=}C(CH_3)_2$	H	H	1.17	1.19	−0.02
70	CH_3	$OCONHN(CH_3)_2$	H	H	1.16	1.03	0.13
71	CH_3	OCONH-cH_6H_{11}	H	H	2.02	1.95	0.07
72	CH_3	OCONHPh	H	H	1.65	1.56	0.09
73	CH_3	OCONH-Napht.	H	H	1.80	1.73	0.07
74	CH_3	$OCONHN(CH_3)_2$	H	H	0.64	0.77	−0.13
75	CH_3	$OCON_3$	H	H	1.09	1.10	−0.01

Table 2. (Continued)

no.	R1	R2	R3	R4	R_M (exp)	R_M (calc)	ΔR_M
			Structure II				
76	CH_3	CH_2CN	Δ4,5 and Δ9,10		0.25	0.39	−0.14
77	CH_3	CH_2N_3	Δ4,5		0.72	0.61	0.11
78	C_2H_5	CH_2N_3	Δ4,5		0.68	0.62	0.06
79	CH_3	CH_2N_3	Δ4,5 and Δ9,10		0.68	0.64	0.04
80	C_2H_5	CH_2N_3	Δ5,6		0.84	0.79	0.05
81	CH_3	CH_2N_3	3-O-CH_3; Δ2,3 and Δ5,10		1.26	1.04	0.22
82	CH_3	CH_2NH_2	Δ4,5		0.70	0.67	0.03
83	CH_3	CH_2OCH_3	Δ4,5		0.40	0.46	−0.06
84	CH_3	CH_2Br	Δ4,5		0.60	0.45	0.15
85	CH_3	H	Δ4,5		0.41	0.29	0.12

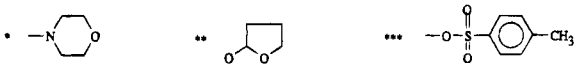


Table 3. Comparison of the Calculated Input Parameters of Two Selected Steroids with Similar Geometry but Largely Different Retention Values

steroid	Solv	*V*	SA	Ha	Vib	Hd	*D*	*Pol*	Epot−	Ent	Epot+	R_M
58	**0.3**	435	724	**1.5**	**0.5**	0.5	3.3	59	60	200	55	3.18
59	**0.6**	467	804	**2.4**	**1.0**	0.5	3.8	63	60	233	70	1.48

Table 4. Experimental and Calculated R_M Values of One of the Test Sets of Seven Selected Steroids

no.	R1	R2	R3	R4	R_M (exp)	R_M (calc)	ΔR_M
1	CH_3	OH	H	CH_3	1.03	0.89	0.14
2	CH_3	H	H	OH	0.80	0.87	−0.07
3	H	OH	CH_2OCH_3	H	0.58	0.64	−0.06
4	CH_3	$OCONHC_2H_5$	H	Br	1.38	1.31	0.07
5	CH_3	$OCONHCH_2$-Ph	H	H	1.58	1.63	−0.05
6	C_2H_5	CH_2N_3	Δ4,5 and 9,10		0.60	0.65	−0.05
7	CH_3	CH_2Cl	Δ4,5		0.57	0.47	0.10

Table 5. Assessment of the Relative Importance of Input Parameters Determined by the Leave-One-Out Method[3]

missing input parameter	change in std deviation of calc R_M values	missing input parameter	change in std deviation of calc R_M values
Solv	+0.033	Ha	+0.010
V	+0.029	Epot+	+0.017
SA	+0.020	Epot−	+0.010
D	+0.015	Vib	+0.012
P	+0.011	Ent	+0.010
Hd	+0.010		

"leading" descriptor Solv, which is the calculated SCRF energy in water, was left out (see Table 5). From a physical point of view this input parameter should contain most of the information necessary to describe the macroscopic R_M value. However, additional geometric input parameters like SA or *V* seem to be essential to improve the quality of the calculations. Molecular surface and volume obviously compensate for part of the errors due to the crude approximation of the spherical cavity model. Additional parameters describing hydrogen bond donor and acceptor abilities (Hd, Ha), surface charge (Epot−, Epot+), dipole moment (*D*), and polarizability (*P*) further improve the accuracy of our predictions.

The dataset in the second part of our investigation contained 323 organic molecules identical with those of a regression analysis of Bodor et al.[36] The dataset includes simple hydrocarbons, halogenated hydrocarbons, multiply substituted benzenes, polynuclear aromatics, ethers, alcohols, aldehydes, ketones, esters, nitriles, amines, nitro compounds, and organosulfur compounds. The range of log P_{oct} values covers a large scope from very soluble (acetone: −0.24; methylamine: −0.57) up to highly unsoluble compounds (*tert*-butylbenzene: 4.11; 2,3,4,5-PCB: 5.91). Descriptors and neural network architecture were identical to those used in the R_M value prediction of the steroid dataset.

Table 6. Performance of the 11 Parameter AM1/Neural Net Approach Compared to a 18 Parameter Regression Analysis[26] in Predicting log P_{oct} Values of a Set of 21 Organic Compounds

molecule	log P_{oct} (exp)	log P_{oct} (calc)	log P_{oct} (regression analysis)[26]
testosterone	3.31	3.15	3.63
prednisone	1.46	1.75	1.95
penicillin	1.83	2.39	1.91
phenytoin	2.47	2.20	2.48
triamcinolon	1.03	1.09	1.59
dexamethasone	1.83	1.53	1.90
betamethasone	1.94	2.17	1.96
p-cresol	1.96	2.22	1.70
2,4,4′-pcb	5.62	5.29	5.52
2,5-pcb	5.16	4.85	5.14
2,6-pcb	4.93	4.67	5.12
2,4,6-pcb	5.47	5.14	5.55
deoxycorticosterone	2.90	2.45	3.17
cortisone	1.42	1.80	2.07
3-chlorophenylacetic acid	2.09	2.28	2.45
4-chlorophenylacetic acid	2.12	2.35	2.51
3-bromophenylacetic acid	2.37	2.30	2.78
4-bromophenylacetic acid	2.31	2.36	2.83
3-fluorophenylacetic acid	1.55	1.73	2.21
4-fluorophenylacetic acid	1.65	1.76	2.18
4-phenylbutanoic acid	2.42	2.70	2.89

The standard deviation in the training set was 0.31 log P_{oct} values (see Figure 4), which is comparable to the results of the 18 parameter regressional analysis.[36] The predictive capability of our model with 11 parameters (standard deviation 0.29, Figure 5, Table 6) is superior compared to the results of an 18 parameter regressional analysis (standard deviation 0.38).[36] Cross-validation of training and test molecules led to a maximum standard deviation of 0.32 in the test set.[4,36]

CONCLUSIONS

Macroscopic parameters (e.g., melting point, solubility, and biological activity) usually are an unknown function of microscopic parameters (quantum theoretical observables). In the present study we have shown that the retention times of 92 steroids in reversed phase chromatography and the partition coefficient of a diverse set of 323 organic com-

pounds can be predicted from 11 microscopic parameters derived from AM1 self-consistent reaction field (SCRF) calculations using a back-propagation neural network as a mapping device.

Assessment of the relative importance of the different parameters proved that the solvation energy obtained from simple spherical SCRF calculations considerably improves the quality of the prediction at little additional computational cost.

However, errors probably due to the spherical approximation have to be compensated by additional geometric parameters. Hydrogen bond donor and acceptor properties, that are not explicitly treated in the continuum solvent model, are accounted for by introducing charges at hydrogen atoms and heteroelements. Minimum and maximum charges on the molecular surface, dipole moment, polarizability, and parameters describing conformational flexibility are also included in the parameter set. We believe that neglect of conformational flexibility is still the main source of errors in our approach and similar treatments[3-10] based on quantum chemical calculations. The input parameters are derived from geometries calculated with the spherical SCRF method, which are only slightly different from the hypothetical gas phase geometries and are only reliable as models for the solutes in solvents of low polarity. In water, however, nonrigid molecules with hydrophilic and hydrophobic parts, like proteins, change their conformation to a large extent, and consequently parameters determined from gas phase geometries are no longer correct.

To account for conformational flexibility we introduced a parameter derived from harmonic frequency calculations. However, this probably does only consider part of the effects. For proper treatment, geometry optimizations would have to be performed in a solvent cavity adapted to the true molecular shape. An algorithm which is able to calculate SCRF energies in such an environment is implemented in the VAMP 5.0[37] program package. Unfortunately, only single point calculations are possible so far.

Hydrophobicity parameters (e.g., retention times, in reversed phase chromatography) for rigid molecules of limited diversity (like steroids) can be predicted with an accuracy close to the error limits of experimental determination. Applied to a heterogenous set of organic compounds our 11 parameter approach including parameters describing conformational flexibility and solvation energy is clearly superior to previous studies using a larger number of parameters but neglecting those effects.[4,36]

REFERENCES AND NOTES

(1) For a recent review, see: Leo, A. L. Calculating log P_{oct}. *Chem. Rev.* **1993**, *93*, 1281−1306 and references therein.

(2) Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Marcus, M. Y.; Taft, R. W. Linear Solvation Energy Relationships. 46. An Improved Equation for Correlation and Prediction of Octanol/Water Partition Coefficients of Organic Nonelectrolytes (Including Strong Hydrogen Bond Donor Solutes). *J. Phys. Chem.* **1988**, *92*, 5244−5255.

(3) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.

(4) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.

(5) Klopman, G.; Namboodiri, N.; Schochet, M. Simple Method of Computing the Partition Coefficient. *J. Comput. Chem.* **1985**, *6*, 28−38.

(6) Klopman, G.; Iroff, L. D. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem.* **1981**, *2*, 157−160.

(7) Koehler, M. G.; Grigorias, S.; Dunn III, W. J. The Relationship Between Chemical Structure and the Logarithm of the Partition Coefficient. *Quant. Struct.-Act. Relat.* **1988**, *7*, 150−159.

(8) Kasei, K.; Umeyama, H.; Tomonaga, A. The Study of Partition Coefficients. The Prediction of log P_{oct} Values Based on Molecular Structure. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 2701−2706.

(9) Kantola, A.; Villar, H. O.; Loew, G. H. Atom Based Parametrization for Conformationally Dependent Hydrophobic Index. *J. Comput. Chem.* **1991**, *12*, 681.

(10) Famini, G. R.; Penski, C. A. Using Theoretical Descriptors in Quantitative Structure Activity Relationships: Some Physicochemical Properties. *J. Phys. Org. Chem.* **1992**, *5*, 395−408.

(11) Sasaki, Y.; Kubodera, H.; Matuszaki, T.; Umeyama, H. Prediction of Octanol/Water Partition Coefficients Using Parameters Derived from Molecular Structures. *J. Pharmacobio-Dyn.* **1991**, *14*, 207−214.

(12) Brinck, T.; Murray, J. S.; Politzer, P. Octanol/Water Partition Coefficients Expressed in Terms of Solute Molecular Surface Areas and Electrostatic Potentials. *J. Org. Chem.* **1993**, *58*, 7070−7073.

(13) Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. Assessment of Methods Used for Predicting Lipophilicity: Application to Nucleosides and Nucleoside Bases. *J. Comput. Chem.* **1993**, *14*, 1019−1026.

(14) Maggiora, G. M.; Elrod, D. W. Computational Neural Networks as Model-Free Mapping Devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732−741.

(15) Aoyama, T.; Ichikawa, H. Neural Networks as Nonlinear Structure-Activity Relationship Analyzers. Useful Functions of the Partial Derivative Method in Multilayer Neural Networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 492−500.

(16) Burden, F. R. Mapping Analytic Functions Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1229−1231.

(17) For a review of solute−solvent interactions, see: Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 2nd ed.; Verlag Chemie: Weinheim, Cambridge, New York, 1990; Chapter 2.

(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(19) Rinaldi, D.; Rivail, J. L.; Rguini, N. Fast Geometry Optimization in Self-Consistent Reaction Field Computations on Solvated Molecules. *J. Comput. Chem.* **1992**, *6*, 675−680.

(20) Rauhut, G.; Clark, T.; Steinke, T. A Numerical Self-Consistent Reaction Field (SCRF) Model for Ground and Exited States in NDDO-Based Methods. *J. Am. Chem. Soc.* **1993**, *115*, 9174−9181.

(21) Rauhut, G.; Chandrasekhar, J.; Alex, A.; Clark, T. VAMP 4.5, Oxford Molecular Ltd.: Oxford, 1993.

(22) Rivail, J. L.; Terryn, B.; Rinaldi, B.; Ruiz-Lopez, M. F. Liquid State Quantum Chemistry: A Cavity Model. *J. Mol. Struct. (Theochem)* **1985**, *120*, 387.

(23) Shinanoglu, O. Molecular Interactions Within Liquids, the Solvophobic Force and Molecular Surface Areas. *Molecular Interactions*; Ratajczak, H., Orville-Thomas, W. J., Eds.; John Wiley and Sons, Ltd.: 1982; Vol. 3, pp 283−342.

(24) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure-Activity Relationships. *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Dekker: New York, 1980; pp 321−345.

(25) Khang, Y. K.; Nemethy, G.; Scheraga, H. A. Free Energy of Hydration of Solute Molecules. 1. Improvement of Hydration Shell by Exact Computation of Overlapping Volumes. *J. Phys. Chem.* **1987**, *91*, 4105−4109.

(26) Marsili, M. *Physical Properties Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer Verlag: Heidelberg, 1988; pp 249. The algorithm in a slightly modified form is implemented in the VAMP 4.5 program package (see lit. 19).

(27) Scrocco, E.; Tomasi, J. The Electrostatic Molecular Potential as a Tool for the Interpretation of Molecular Properties. *Top. Curr. Chem.* **1973**, *42*, 95.

(28) Rauhut, G.; Clark, T. Multicenter Point Charge Model for High-Quality Molecular Electrostatic Potentials from AM1 Calculations. *J. Comput. Chem.* **1993**, *14*, 503.

(29) For another approach, see: Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323.

(30) Terada, H. Determination of Log P_{oct} by High-Performance Liquid Chromatography, and its Application in the Study of Quantitative Structure−Activity Relationships. *Quant. Struct.-Act. Relat.* **1986**, *5*, 81−88.

(31) PCMODEL 6.0; Molecular Modeling Software, Serena Software, Box 3076, Bloomington, IN 47402-3076. Copyright 1987, 1988, 1989, 1990.

(32) (a) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms. *J.*

Am. Chem. Soc. **1977**, *99*, 8127. (b) Allinger, N. L.; Yuh, Y. H.; Lii, J.-H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8551 and references therein.

(33) Pulay, P.; Török, F. Calculation of Molecular Geometries and Force Constants from CNDO Wavefunctions by the Force Method. *Mol. Phys.* **1973**, *25*, 1153.

(34) Rumelhart, D. E. Parallel Distributed Processing; MIT Press: Cambridge, MA, USA, 1986; Vol. 1 + 2.

(35) Draffehn, J.; Schönecker, B.; Ponsold, K. Reversed-Phase Thin-Layer Chromatography of Steroids. I. Measurement and Interpretation of R_M values. *J. Chromatogr.* **1981**, *205*, 113−124.

(36) Bodor, N.; Huang, M.-J. An Extended Version of a Novel Method for the Estimation of Partition Coefficients. *J. Pharm. Sci.* **1992**, *81*, 372.

(37) Rauhut, G.; Chandrasekhar, J.; Alex, A.; Steinke, T.; Clark, T. VAMP 5.0; Centrum für Computer-Chemie der Universität Erlangen-Nürnberg, 1993.

CI950051T