Journal of Computer-Aided Molecular Design, **15:** 741–752, 2001.
KLUWER/ESCOM

Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure

David J. Livingstone[a,*], Martyn G. Ford[b], Jarmo J. Huuskonen[c] & David W. Salt[d]
[a]*ChemQuest, Delamere House, 1, Royal Crescent, Sandown, Isle of Wight, PO36 8LZ, UK;* [b]*Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants, PO1 2EG, UK;* [c]*Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, FIN-00014 University of Helsinki, Finland;* [d]*School of Computer Science & Mathematics, University of Portsmouth, Portsmouth, Hants, PO1 2DY, UK*

Received 20 November 2000; accepted 11 June 2001

Key words: canonical correlation, electrotopological descriptors, log *P*, log *S*, neural networks, regression analysis

Summary

It has been shown that water solubility and octanol/water partition coefficient for a large diverse set of compounds can be predicted simultaneously using molecular descriptors derived solely from a two dimensional representation of molecular structure. These properties have been modelled using multiple linear regression, artificial neural networks and a statistical method known as canonical correlation analysis. The neural networks give slightly better models both in terms of fitting and prediction presumably due to the fact that they include non-linear terms. The statistical methods, on the other hand, provide information concerning the explanation of variance and allow easy interrogation of the models. Models were fitted using a training set of 552 compounds, a validation set and test set each containing 68 molecules and two separate literature test sets for solubility and partition.

Introduction

A wide variety of properties, experimental, empirical and theoretical, may be estimated or calculated for the description of chemical structures and their interactions with solvents, other small molecules and complex molecular structures such as membranes and pharmaceutical receptors [1]. Of the experimental properties, perhaps the two most important are water solubility (log *S*) and octanol/water partition coefficient (log *P*). Because of this importance several groups have reported methods for the calculation of solubility [2–7] and a rather larger number of studies have examined the prediction of log *P* [4, 8–18]. The compounds used in these reports have been characterised directly by molecular structure (group contributions) or by properties derived from structure and the mathematical models relating log *S* or log *P* to these descriptors have mainly been constructed using multiple linear regression (MLR).

Recently, non-linear models have been constructed using artificial neural networks (ANN) [19–22] and we have reported methods for the calculation of log *S* [23–26] and log *P* [27, 28] using electrotopological descriptors [29] coupled with ANN. Since both of these properties may be predicted successfully using the same set of descriptors it should be possible to model them simultaneously. An artificial neural network may be trained to compute two target values from the same set of inputs by simply including an extra neuron in the output layer. Statistical methods may also be used to model two or more response variables simultaneously, one such technique is known as canonical correlation analysis (CCA). Although there are often circumstances in drug discovery where multiple responses or measurements are available for a set of compounds, CCA has been applied rarely to analyse the data [30–33]. This paper reports a comparison of the use of MLR, ANNs and CCA to model log *S* and log *P* using a set of 37 electrotopological descriptors.

*To whom correspondence should be addressed; E-mail: davel@chmqst.demon.co.uk

Methods

Data

A set of 900 drug and pesticide-like compounds were chosen from the AQUASOL dATAbASE [34] and the aqueous solubility values at 25 °C expressed as log S, were S is the solubility in moles per litre, were used. The compounds were checked in the 'starlist' (preferred measured values) of Biobyte [35] and those with an entry were selected to give an overall set of 688 compounds with both log S and log P values. A referee has pointed out, however, a potential problem with literature solubility values for ionisable compounds. It appears that reported solubilities are 'invariably' those of the unbuffered solution where the pH is that of the saturated solution. The common data set of 688 compounds contains 82 carboxylic acids and 43 aliphatic amines, solubility values for these compounds will be for a mixture of the neutral and ionised species.

Training/test sets

The set of 688 compounds was subdivided into three sets; a validation set of 10% ($n = 68$), a test set of 10% ($n = 68$) and a training set of the remainder ($n = 552$). The validation and test sets were selected so that they contained molecules presenting all the different chemical structures and functional groups found in the training set. The log S values of the training set ranged between -6.30 and 1.60 and the log P values between -3.89 and 6.50, respectively. Figure 1 shows the distribution of solubilities and partition coefficient values for the training set. Two further test sets were chosen from the literature, a test set of 19 for log P [36] and a test set of 21 for log S [37], allowing comparison with earlier proposed log S and log P estimation methods.

Multiple linear regression analysis

Multiple linear regression (MLR) analysis was performed with the SPSS software (v.8.0, SPSS Inc., Chicago, IL) running on a Pentium PC. The quality criteria of the fit in MLR analysis were squared correlation coefficient, R^2 , standard error, s, and Fischer significance value, F, when all parameters in the model were significant at the 95% confidence level.

Neural network implementation

The artificial neural network simulations were carried out using the NeuDesk software (v. 2.20, Neural Computational Sciences, UK). A three-layered, fully connected neural network was trained by the standard back-propagation learning algorithm with a logistic $f(x) = 1/(1 + e^{-x})$ activation function both for hidden and output nodes. Before the training was started, the input and output values were scaled between 0.1 and 0.9, and the adjustable weights between neurons were given random values of between -0.5 and 0.5. The learning rate and momentum parameters were set at 0.1 and 0.9, respectively. The optimal training endpoint and network architecture was determined on the basis of the validation set of 68 compounds. The network architecture and the training endpoint giving the highest coefficient of determination, R^2_{pred}, and the lowest standard error s for the predictions of the validation set was then used. The predictions were repeated 10 times with different random starting weights in the network and the averaged log S values were calculated.

Canonical correlation analysis

Canonical correlation analysis was carried out using the 6M routine of the BMDP package [38] running on a PC. Since canonical correlation is a somewhat less well known technique than other statistical methods a brief explanation is presented here:
Canonical correlation analysis (CCA) is a generalisation of multiple linear regression analysis. In the latter a weighted sum of the predictor variables is found which maximally correlates with a single response variable. CCA is a technique, which finds a weighted sum of several response variables that is maximally correlated with a weighted sum of the predictor variables. Whereas in multiple regression, where the responses are analysed independently ignoring any covariance amongst the response set of variables, CCA utilises this shared information and provides an analysis of all response variables simultaneously.

Canonical variates

Denote the variables in the response set $Y_1, Y_2, \ldots, Y_q$ and the variables in the predictor set $X_1, X_2, \ldots, X_p$. In CCA the coefficients $a_{11}, a_{12}, \ldots, a_{1q}$ and $b_{11}, b_{12}, \ldots, b_{1p}$ are found such that the two linear constructs W_1 and Z_1, where

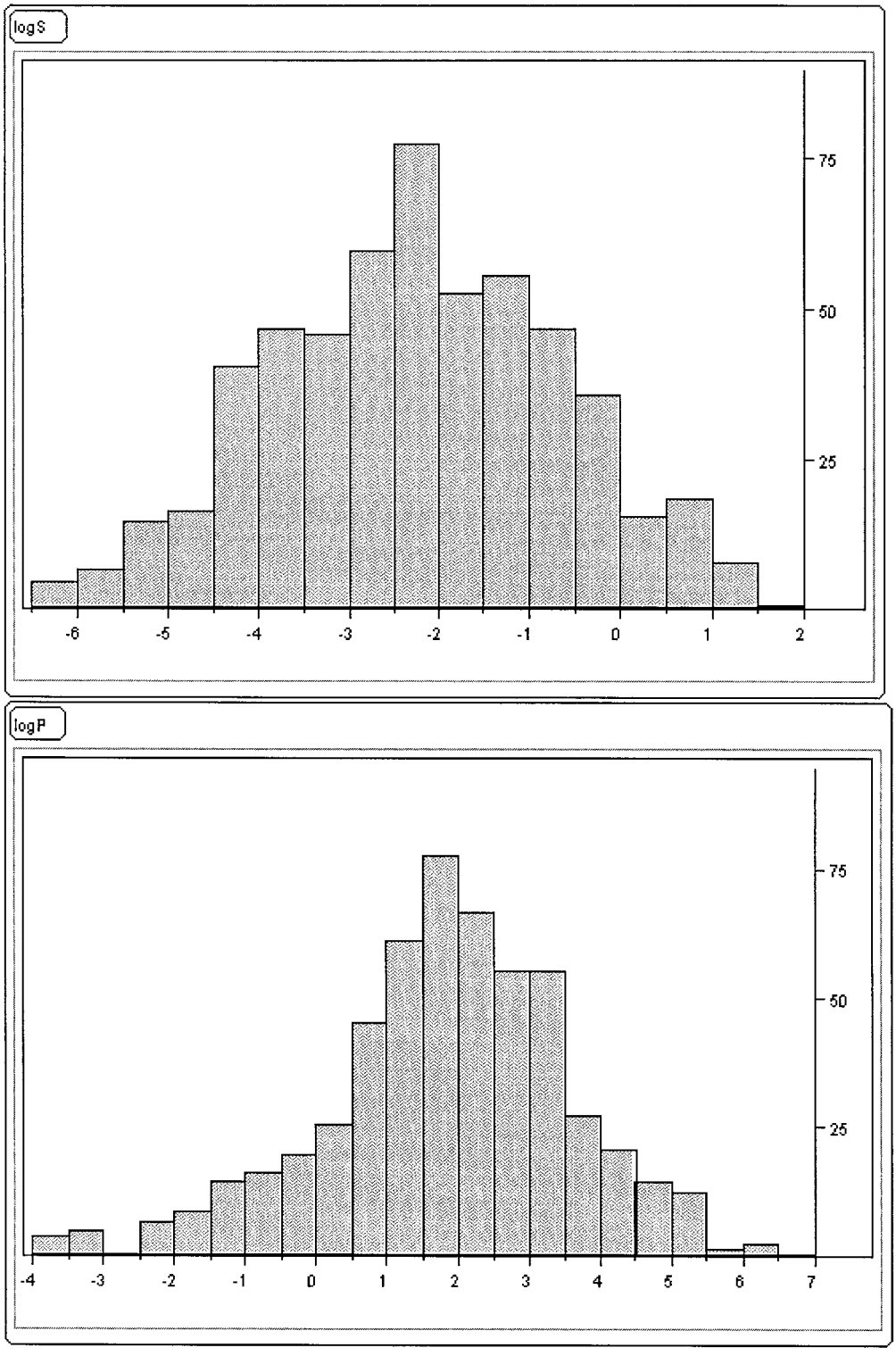

Figure 1. Distribution of log *S* and log *P* values for the training set.

$$W_1 = a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1q}Y_q \quad (1)$$

and

$$Z_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p \quad (2)$$

are maximally correlated. The two new variables, W_1 and Z_1, are referred to as canonical variates (CV) and the correlation between them (R_1) is known as the canonical correlation. A second pair of CVs, (W_2, Z_2) is then selected to account for a maximum amount of the relationship between the two sets of variables unaccounted for by the first pair of CVs, and so on. The number of pairs of canonical variates constructed equals the smaller of q and p. Thus, the linear relationships,

$$\begin{aligned} W_1 &= a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1q}Y_q, \\ W_2 &= a_{21}Y_1 + a_{22}Y_2 + \ldots + a_{2q}Y_q, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ W_s &= a_{s1}Y_1 + a_{s2}Y_2 + \ldots + a_{sq}Y_q \end{aligned}$$

and

$$\begin{aligned} Z_1 &= b_{11}X_1 + B_{12}X_2 + \ldots + b_{1p}X_p, \\ Z_2 &= b_{21}X_1 + b_{22}X_2 + \ldots + b_{2}X_p, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ Z_s &= b_{s1}X_1 + b_{s2}X_2 + \ldots + b_{sp}X_p \end{aligned} \quad (3)$$

can be found, where s is the smaller of q and p. The pairs of canonical variates are extracted so that the correlation R_1 between the first pair of CV's (W_1, Z_1) is a maximum; the correlation R_2 between the second pair (W_2, Z_2) is a maximum subject to these variables being uncorrelated with (W_1, Z_1) and $R_2 < R_1$; and so forth.

Procedure for CCA

The method of extracting the successive pairs of CVs involves an eigenvalue-eigenvector analysis. The eigenvalues (R_i^2) and associated eigenvectors constructed by the CCA are based on the combined $(p+q) \times (p+q)$ correlation matrix, **C**, between the descriptor variables and the response variables, where

$$\mathbf{C} = \begin{bmatrix} p \times p\text{matrix}\mathbf{R}_{XX} & \vdots & p \times q\text{matrix}\mathbf{R}_{XY} \\ \ldots\ldots\ldots\ldots & \vdots & \ldots\ldots\ldots\ldots \\ q \times p\text{matrix}\mathbf{R'}_{XY} & \vdots & q \times q\text{matrix}\mathbf{R}_{YY} \end{bmatrix} \quad (4)$$

From this matrix a $s \times s$ matrix $\mathbf{R}_{YY}^{-1}\mathbf{R'}_{XY}\mathbf{R}_{YY}$ is constructed, and its eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_s$ are the squares of the canonical correlations between the pairs of canonical variates and represents the amount of variance in the canonical variate W_i that is accounted for by the other canonical variate Z_i. The corresponding eigenvectors allow the canonical variate coefficients (a_{ij} and b_{ij} in equation (3)) to be calculated.

Canonical weights and canonical loadings

Canonical weights, a_{ij} and b_{ij} in equation (3), are analogous to the coefficients in multiple linear regression analysis (MRA) and indicate the contribution of each variable to the variance of the respective canonical variate. Canonical loadings are more useful in identifying the nature of the canonical relationships. Canonical loadings give the simple product moment correlation of the original variable and its respective CV and reflect the degree to which the variable is represented by a CV. The canonical loadings can easily be found by correlating the raw variable scores with the canonical variate scores.

Predicting a response variable

There are several ways in which the results of a CCA can be used to obtain a prediction for a response variable. One approach is to regress each of the original responses on the appropriate set of canonical variates which have been constructed. Because each variate is a new, orthogonal variable with a known functional relationship to the original variables, the procedure is straightforward. Each original y variable is regressed on the set of canonical variates constructed from the molecular descriptors (the X block) using a standard multiple regression procedure.

A second approach is to base prediction on a method analogous to that used to solve simultaneous equations. The various canonical variates are regarded as i equations in i unknowns which can be solved analytically; i is the number of variables in the smallest set.

Results and discussion

Regression analysis

Stepwise and backward methods were employed in the regression analysis. The following regression equations were calculated for log S with 28 and log P with 26 significant descriptors at 95% confidence limits:

$$\log S = \Sigma(a_i S_i) + 1.128$$
$$n = 552,\ R^2 = 0.783,\ s = 0.753,\ F = 68.18 \quad (5)$$

$$\log P = \Sigma(a_i S_i) - 0.635$$
$$n = 552,\ R^2 = 0.870,\ s = 0.645,\ F = 134.8 \quad (6)$$

In both equations n is the number of the compounds used in the fit, F is the overall F-statistic for the addition of each successive descriptor, and a_i and S_i are the regression coefficients and the corresponding electrotopological indices. The regression coefficients in equations 5 and 6 are indicated in Table 1 with the t-scores of significant parameters. Fitting and prediction results for these two regression equations are shown in table 2.

Neural networks

In order to allow direct comparison with the multiple linear regression models, neural networks were trained separately to reproduce log S and log P values using the same topological descriptors as selected by the stepwise regression procedure. In other words, the neural network equivalents of equations 5 and 6. In the case of solubility the 'best' network architecture was a 28-6-1 network and for partition coefficient a 26-5-1 network. Fitting and prediction results for these networks can be seen in Table 2 where the network fitting statistics are improved over the MLR results for the training and test sets and are equivalent to MLR for the validation sets.

Training a neural network with all 37 topological descriptors using both log S and log P as targets resulted in a 'best' network architecture of 37-10-2 with a training set R^2 of 0.894 for water solubility and $R^2 = 0.931$ for partition coefficient. Statistics for training, validation and test sets for this network are shown in Table 2. One of the drawbacks of neural network modelling is that the final model is hidden in the set of overall weights and thus it is difficult to estimate the contribution that individual descriptors make to the model. Neural network pruning techniques do allow an estimation of the importance of individual input variables [39]. Application of one such pruning method [40] resulted in the removal of seven of the topological descriptors as shown in Table 1. Training a network with the remaining 30 descriptors also resulted in a 'best' network architecture with 10 hidden neurons, i.e. a 30-10-2 ANN. This network gave improved statistics for all the data sets, except the test set for partition, as shown in Table 2. Figures 2 and 3 show plots of calculated against observed for the training set, validation set and test set compounds.

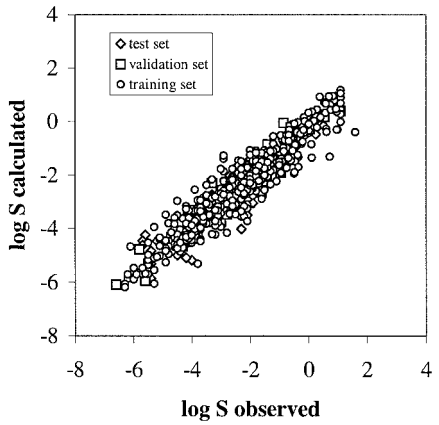

Figure 2. Plot of calculated vs. observed values of log S for the training set, validation set and test set compounds.

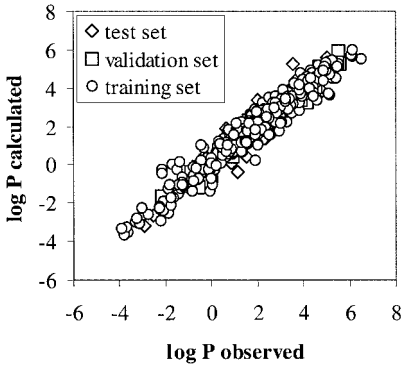

Figure 3. Plot of calculated vs. observed values of log P for the training set, validation set and test set compounds.

Calculated values for the compounds in the two literature test sets of solubility and partition coefficient are shown in Tables 3 and 4 respectively. It can be seen from Table 3 that the ANN gives improved results compared with both the MLR model and the two published comparisons for this set [41, 42]. Two of the compounds in this test set (2,2′,4,5,5′-PCB and 4,4′-DDT) have log S values which are well outside the range of values of the training set (−6.3 to 1.60). These two compounds are poorly predicted by both the MLR and ANN models but are quite well predicted by the two literature methods. For the partition coefficient test set, ANN performs better than XLOGP, MLOGP and the Rekker method but not as well as CLOGP or KOWWIN. Figures 4 and 5 show the correlation between predicted against observed log S values in the test set of 21 drug and pesticide compounds and

Table 1. Parameters used[a] in multiple linear regression and neural network models.

No.	Symbol	Atom-type	log *S*	*t*-score	log *P*	*t*-score	ANN	obsd	min	max
1	SsCH3	$-CH_3$	−0.208	11.752	0.322	21.685	X	348	0	13.905
2	SdCH2	$=CH_2$			0.253	6.083	X	16	0	7.020
3	SssCH2	$-CH_2$-	−0.271	14.568	0.337	21.025	X	299	−1.691	12.161
4	StCH	≡CH	−0.239	2.901	0.368	1.919		4	0	5.703
5	SdsCH	= CH-	−0.222	5.297			X	72	0	5.982
6	SaaCH	aCHa	−0.177	16.224	0.212	26.156	X	336	0	19.332
7	SsssCH	> CH -	−0.174	3.680	0.189	5.580	X	173	−6.865	2.929
8	StC	≡C -			−0.644	1.716		10	0	3.496
9	SdssC	= C <			−0.272	6.857	X	304	−7.562	2.904
10	SaasC	asCa	−0.109	2.370	0.066	1.928	X	350	−5.362	6.070
11	SaaaC	aaCa	−0.242	3.303			X	41	0	4.765
12	SssssC	> C <	−0.263	4.227	0.080	1.705	X	106	−6.759	0.601
13	SsNH3+	- NH_3+			−0.512	14.179	X	19	0	5.939
14	SsNH2	- NH_2	0.036	1.862	−0.103	6.474	X	77	0	10.790
15	SdNH	= NH						2	0	6.804
16	SssNH	- NH -	0.074	2.711	−0.117	5.584	X	131	0	6.666
17	SaaNH	aNHa						4	0	3.084
18	StN	≡ N	−0.120	4.564	0.152	1.916	X	6	0	17.409
19	SdsN	= N -	−0.119	3.761			X	29	0	8.086
20	SaaN	aNa	−0.097	7.133	0.020	1.906	X	79	0	15.657
21	SsssN	> N -	0.282	7.132	−0.275	13.470	X	124	0	7.444
22	SddsN	- N <<	0.549	4.150	−1.123	11.643	X	31	−3.706	0
23	SsOH	- OH	0.041	7.782			X	225	0	66.476
24	SdO	= O	−0.054	11.728			X	353	0	58.507
25	SssO	- O -	−0.029	2.640			X	170	0	29.004
26	SaaO	aOa						6	0	5.105
27	SsF	- F	−0.098	10.021	0.073	9.601	X	25	0	39.932
28	SdsssP	->P=	−0.202	1.955			X	34	−5.743	0
29	SsSH	- SH						2	0	3.983
30	SdS	= S	−0.320	6.989	0.153	6.703	X	37	0	10.862
31	SssS	- S -	−0.309	3.895	0.294	4.458	X	48	0	3.598
32	SaaS	aSa	−0.563	3.116	0.479	3.118		12	0	1.823
33	SdssS	> S =			0.726	2.807	X	4	−1.861	0
34	SddssS	> S <<	−0.132	3.077	0.082	2.513	X	33	−9.059	0
35	SsCl	- Cl	−0.177	23.065	0.162	26.346	X	86	0	51.006
36	SsBr	- Br	−0.319	6.877	0.365	9.203	X	11	0	9.650
37	SsI	- I	−0.932	4.469	0.851	4.761	X	4	0	2.280

[a]Where an entry is missing from the log *S* and log *P* columns that descriptor was removed during the stepwise MLR analysis for that property. An X in the ANN column indicates that this parameter was retained following network pruning, that is to say these descriptors were used in the 30-10-2 network.

log *P* values in the test set of 19 drug compounds, respectively.

Canonical correlation analysis

As there are only two variables in the smaller of the two sets of variables (log *P* and log *S*) there is a maximum of two pairs of canonical variates that can be extracted. Application of CCA yielded both pairs to be significant (Bartlett's test: $\chi^2 = 442.1$, 36 degrees of freedom, $p = 0.0000$) and Table 5 shows the canonical correlations along with the canonical coefficients and loadings for the response set of variables (cnvrf1, cnvrf2). The first canonical correlation between the 37 descriptors and the two response variables log *P* and log *S* is large (0.923). The second is, of course smaller

Table 2. Comparison of the predictive ability of multiple linear regression and neural network models.

Model	#	Training set			Validation set			Test set		
		R^2	s	n	R^2	s	n	R^2	s	n
(A) Aqueous solubility, log S										
MLR	28	0.78	0.75	552	0.84	0.63	68	0.78	0.75	68
ANN[a]	28	0.87	0.60	552	0.84	0.63	68	0.83	0.65	68
ANN[b]	37	0.89	0.53	552	0.86	0.60	68	0.83	0.65	68
ANN[c]	30	0.90	0.52	552	0.92	0.44	68	0.84	0.63	68
(B) Partition coefficient, log P										
MLR	26	0.87	0.65	552	0.91	0.50	68	0.86	0.65	68
ANN[a]	26	0.91	0.53	552	0.90	0.51	68	0.89	0.55	68
ANN[b]	37	0.93	0.47	552	0.90	0.51	68	0.89	0.55	68
ANN[c]	30	0.94	0.44	552	0.94	0.40	68	0.89	0.55	68

= number of variables used in the model.
[a]Same parameters as in MLR equation used as input to ANNs.
[b]All atom-type E-state indices used as inputs to ANNs with two outputs.
[c]Significant parameters after pruning used as inputs to ANNs with two outputs.

Table 3. The observed and predicted aqueous solubility values for the test set.

No.	Compound	log S_{exp}	MLR	ANN	Klopman[a]	Kühne[b]
1	antipyrine	0.39	−1.47	−1.31	−2.76	−1.90
2	theophylline	−1.39	−1.10	−1.24	−1.07	0.54
3	acetylsalicylic acid	−1.72	−1.83	−1.84	−1.52	−1.93
4	benzocaine	−2.32	−1.59	−1.73	−1.71	−1.75
5	phenobarbital	−2.32	−2.80	−3.26	−2.08	−2.41
6	prostaglandin E2	−2.47	−4.65	−3.98	−4.21	na
7	phenolphthalein	−2.90	−4.16	−3.79	−4.48	−4.61
8	malathion	−3.37	−3.39	−3.22	−2.94	−3.48
9	nitrofurantoin	−3.38	−2.50	−3.06	−2.19	−2.62
10	diazinon	−3.64	−4.16	−4.13	−5.29	−4.98
11	diazepam	−3.76	−4.17	−3.90	−5.54	−4.51
12	diuron	−3.80	−3.04	−3.26	−2.85	−3.38
13	atrazine	−3.85	−3.12	−3.98	−3.05	−3.95
14	phenytoin	−3.90	−3.68	−4.02	−3.47	−5.25
15	testosterone	−4.09	−4.24	−4.44	−5.17	−4.62
16	lindane	−4.64	−4.66	−4.90	−4.88	−5.08
17	parathion	−4.66	−3.82	−3.90	−3.94	−4.59
18	chlorpyriphos	−5.49	−5.13	−5.16	−5.77	−3.75
19	a-chlordane	−6.86	−6.79	−6.05	−7.55	−6.51
20	2.2′.4.5.5′-PCB	−7.89	−6.27	−6.34	−7.90	−7.47
21	4,4′-DDT	−8.08	−6.65	−6.30	−8.00	−7.75
		r^2	0.79	0.86	0.72	0.76
		s	0.97	0.77	1.11	1.06
		n	21	21	21	20

[a]Ref. 42.
[b]Ref. 41.

Table 4. The observed and predicted partition coefficient values for the test set.

No.	Compound	Log P_{obsd}	MLR	ANN	XLOGP	MLOGP	Rekker	CLOGP	KOWWIN
1	chlorothiazide	−0.24	−0.74	0.00	−0.58	−0.36	−0.68	−0.30	−0.23
2	cimetidine	0.40	1.90	1.74	0.20	0.82	0.63	0.35	0.57
3	procainamide	0.88	1.60	1.52	1.27	1.72	1.11	1.42	0.97
4	trimethoprim	0.91	1.44	1.27	0.72	1.26	−0.07	0.88	0.73
5	chloramphenicol	1.14	0.96	1.00	1.46	1.23	0.32	1.28	0.92
6	phenobarbital	1.47	1.82	1.91	1.77	0.78	1.23	1.36	1.33
7	atropine	1.83	2.59	2.23	2.29	2.21	1.88	1.31	1.91
8	lidocaine	2.26	2.96	2.82	2.47	2.52	2.30	1.95	1.66
9	phenytoin	2.47	2.91	2.64	2.23	1.80	2.76	2.08	2.16
10	diltiazem	2.70	4.23	3.31	3.14	2.67	4.53	3.64	2.79
11	propranolol	2.98	3.60	3.70	2.98	2.53	3.46	2.75	2.60
12	diazepam	2.99	3.52	3.00	2.98	3.36	3.18	3.16	2.70
13	diphenhydramine	3.27	4.88	4.21	3.74	3.26	3.41	3.54	3.11
14	tetracaine	3.51	3.31	2.78	2.73	2.64	3.55	3.83	3.02
15	verapamil	3.79	6.70	4.55	5.29	3.23	6.15	4.46	4.80
16	haloperidol	4.30	4.76	4.51	4.35	4.01	3.57	3.84	4.20
17	imipramine	4.80	4.70	4.71	4.26	3.88	4.43	5.03	5.01
18	chlorpromazine	5.19	5.13	5.35	4.45	3.86	5.81	5.51	5.15
19	flufenamic acid	5.25	3.99	4.65	4.91	3.77	5.10	5.92	5.65
		r^2	0.74	0.91	0.90	0.88	0.84	0.95	0.96
		s	0.86	0.51	0.54	0.59	0.66	0.37	0.34
		n	19	19	19	19	19	19	19

Table 5. Canonical correlations and canonical weights and loadings for the canonical variates for the response set of variables.

	Canonical weights		Canonical loadings	
	cnvrf1	cnvrf2	cnvrf1	cnvrf2
log S	−0.122	0.850	−0.806	0.592
log P	0.474	0.644	0.990	0.143
Canonical correlations	0.923	0.752		

Table 6. Proportion of variance in the response set accounted for by the predictor set.

Canonical variate	(a) Squared canonical correlation	(b) Proportion of variance accounted for	(a) × (b)
1	0.8518	0.8143	0.6936
2	0.5651	0.1857	0.1049
		Total =	0.7985

Table 7. Fit and prediction statistics (R^2) for the canonical correlation analysis.

	Training set	Validation set	Test set 1	Test[b] Set 2
Solubility				
CCA[a]	0.752	0.828	0.759	0.785
Regression on CVs	0.751	0.828	0.758	0.820
Partition				
CCA[a]	0.846	0.903	0.867	0.738
Regression on CVs	0.846	0.904	0.868	0.767

[a]Obtained by simultaneous equations.
[b]These are the two published test sets for log S (21 compounds) and log P (19).

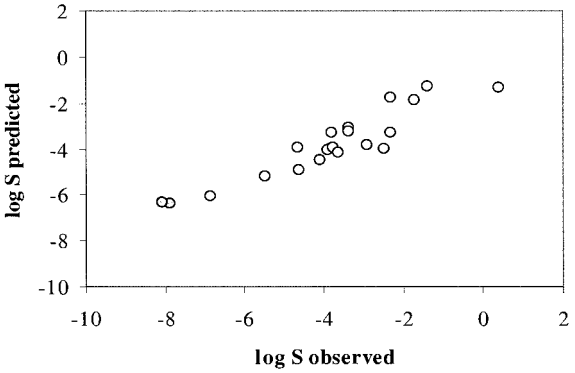

Figure 4. Correlation of predicted vs. observed log S values in the test set of 21 drug and pesticide compounds.

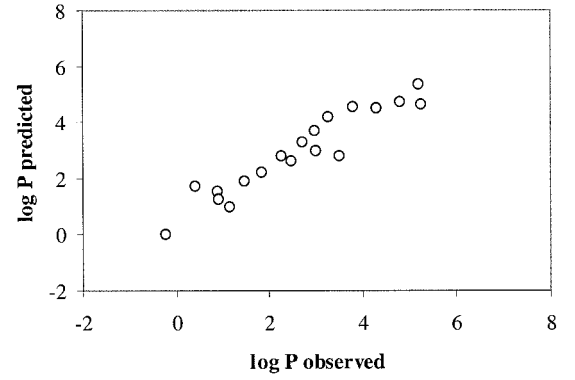

Figure 5. Correlation of predicted vs. observed log P values in the test set of 19 drug compounds.

(0.752), but still indicates high association between the second pair of canonical variates.

Let us now consider the other elements in the table. Essentially the canonical weights are comparable with regression weights and stress the importance of a variable from one set in relation to the other set with regard to maximising the correlation between the sets. The first canonical variate is based on a difference between log S and log P whereas the second variate is based on a weighted sum of the original variables. As the weights are comparable with multiple regression coefficients they are subject to the same problem of multicollinearity and consequently the canonical loadings should be used in conjunction with the weights if reliable interpretations of the canonical variates are to be achieved.

As in multiple regression it is useful to know what proportion of the variance in the response set is accounted for by a particular canonical variate. The proportion of explained variance in the Y-set that is accounted for by a particular canonical variate is given by

$$[R_Y(j)]^2 = \sum_{i=1}^{q} [r_{Y_i}(j)]^2 / q \qquad (7)$$

where $[R_Y(j)]^2$ denotes the proportion of variance in the response set variables accounted for by the jth canonical variate, and $r_{Y_i}(j)$ is the canonical loading of the ith response variable on the jth canonical variate. Similarly, the proportion of variance in the predictor set of variables (X) accounted for by the jth canonical variate can be evaluated. The results of applying equation 7 to the current data set are shown in Table 6 where it can be seen that 81.4% of the variability in log P and log S is accounted for by the first canonical variate. A similar calculation shows that the proportion of the variance accounted for by the response set by the second canonical variate is 18.6%.

The proportion of variance accounted for in the response set just evaluated gives the amount of this variance explained by the respective canonical variates. It would be useful, however, to know how much

of the variance in the response set is accounted for by the predictor set. One might think that R_C^2 provides this information. However, although the squared canonical correlation coefficients do have some variance interpretations, they give the variance shared by the canonical variates and not the variance shared by the original X and Y variables. Stewart and Love [43] have proposed an index, called the redundancy coefficient, which represents the amount of variance in the response set that is 'redundant' to the variance in the predictor set. This redundancy coefficient, denoted by $R_{X_{Y/X}}$, is given by

$$R_{C_{Y/X}} = \sum_{j=1}^{s} \lambda_j [R_Y(j)]^2 \quad (8)$$

and is the sum of the product of the proportion of explained variance in the Y set that is accounted for by a particular canonical variate with its associated squared canonical correlation coefficient. A redundancy coefficient, $R_{C_{X/Y}}$, can also be constructed and represents the amount of variance in the X set of variables that is redundant to the variance in the Y set, but this is usually of secondary importance. The redundancy coefficient calculated for the current data set is approximately 80%, which is very high. This figure comes about because the first response set canonical variate contains 81.4% of the response set variability and that variate shares 85.2% of its variance with the first predictor set (X) canonical variate. Considering both canonical variates together, about 80% of the response set variability is accounted for.

The performance of these canonical correlation equations in fitting and prediction are shown in table 7. A comparison of the results from the ANN analyses of the various data sets (Table 2) with those obtained from the two methods used in the CCA (Table 7) show that the neural networks provide better fits to the training data and are better at prediction. This is presumably due to the fact that the ANN are able to accommodate non-linearity in their fitting. The CCA models, on the other hand, do have the advantage that it is possible to 'dissect' the amount of variance explained in and by the response and descriptor sets and the resulting models, like MLR models, can be readily interrogated. Details of the CCA models are shown in the Appendix.

Conclusions

Two important pharmaceutical properties, namely aqueous solubility and partition coefficient, have been modelled simultaneously using ANN and CCA. Both methods produce quite satisfactory models with the ANN outperforming the CCA approach. CCA has the advantage that the proportion of variance explained is easily obtained and the resulting models can be expressed in the form of coefficients that are the equivalent of multiple regression coefficients (see appendix). ANN models suffer from the fact that the model is contained within a set, often large, of network weights although techniques have been proposed by which these models may be extracted [44].

It is, of course, not very surprising that these two properties can be successfully modelled simultaneously since each property can be well described alone using electrotopological descriptors [25, 28]. What these results show, however, is the ease with which two, or more, dependent variables may be modelled using a set or sets of physicochemical properties. The dependent variables may be pharmacological responses, adverse effects such as toxicity measures, desirable (or undesirable) properties such as log P and log S, measures of stability or pharmacokinetcs and so on. In this case a single set of descriptors was sufficient to model both properties but in other applications it may be necessary to use different types [1] of descriptor for the different dependent variables. The models produced by the CCA technique can be most informative since the first canonical variate is based on the difference between the two dependent variables whereas the second canonical variate is based on their weighted sum. It is obvious how such information may be useful in the drug design process.

Appendix. Coefficients of the responses and descriptors on the two canonical variates

Response	coef1	coef2	Descriptor	coef1	coef2
log *S*	−0.122	0.850	SsCH3	0.185978	0.012381
log *P*	0.474	0.644	SdCH2	0.130015	0.117873
			SssCH2	0.204915	−0.017625
			StCH	0.161745	0.020671
			SdsCH	0.044057	−0.244769
			SaaCH	0.114486	−0.034184
			SsssCH	0.126668	−0.068152
			StC	−0.219068	−0.542252
			SdssC	−0.152605	−0.268030
			SaasC	0.093036	−0.035704
			SaaaC	0.129195	−0.249057
			SssssC	0.076709	−0.282966
			SsNH3	−0.265181	−0.431079
			SsNH2	−0.046216	−0.154948
			SdNH	−0.008897	−0.185470
			SssNH	−0.036128	−0.228728
			SaaNH	−0.035715	0.027962
			StN	0.071061	−0.029496
			SdsN	0.008058	−0.163893
			SaaN	0.007526	−0.091930
			SsssN	−0.238424	−0.039155
			SddsN	−0.691107	−0.307503
			SsOH	0.005564	−0.053314
			SdO	0.003980	−0.065451
			SssO	0.002466	−0.040553
			SaaO	0.030563	−0.117513
			SsF	0.051114	−0.051660
			SdsssP	0.006999	−0.300027
			SsSH	0.013365	0.010609
			SdS	0.119694	−0.256348
			SssS	0.193939	−0.114682
			SaaS	0.342971	−0.376335
			SdssS	0.324943	0.776224
			SddssS	0.033125	−0.133106
			SsC1	0.100302	−0.063910
			SsBr	0.241055	−0.106695
			SsI	0.562609	−0.353126

References

1. Livingstone, D.J., J.Chem. Inf. Comput. Sci., 40 (2000) 195.
2. Nirmalakhandan, N.N. and Speece, R.E., Environ. Sci. Technol., 22 (1988) 328.
3. Bodor, N. and Huang, M-J., J. Pharm. Sci., 81 (1992) 954.
4. Patil, G.S., J. Hazard. Mater., 36 (1994) 35.
5. Sutter, J.M. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 36 (1996) 100.
6. Huibers, P.D.T. and Katritzky, A.R., J.Chem.Inf.Comput.Sci., 38 (1998) 283.
7. Mitchell, B.E. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 38 (1998) 489.
8. Rekker, R. E., Hydrophobic Fragment Constant; Elsevier: New York, 1977.
9. Hansch, C. and Leo, A., Substituent Constants for Correlation Analysis in Chemistry and Biology ; Wiley: New York, 1979.
10. Klopman, G. and Iroff, L., J. Comput. Chem., 2 (1981) 157.
11. Bodor, N. and Huang, M-J., J. Pharm. Sci., 81 (1992) 272.
12. Leo, A., Chem. Rev., 93 (1993) 1281.
13. Klopman, G.; Li, J-Y.; Wang, S. and Dimayuga, M., J. Chem. Inf. Comput. Sci., 34 (1994) 752.

14. Meylan, W. M. and Howard, P. H., J. Pharm. Sci., 84 (1995) 83.
15. Wang, R.; Fu, Y. and Lai, L., J. Chem. Inf. Comput. Sci. 37 (1997) 615.
16. Haeberlin, M. and Brinck, T., J. Chem. Soc. Perkin Trans. 2, (1997) 289.
17. Bodor, N. and Buchwald, P., J. Phys. Chem., 101 (1997) 3404.
18. Buchwald, P. and Bodor, N., Current. Med. Chem., 5 (1998) 353.
19. Bodor, N. and Huang, M-J., J.Am.Chem.Soc., 113 (1991) 9480.
20. Breindl, A.; Beck, N.; Clark, T. and Glen, R. C., J. Mol. Model., 3 (1997) 142.
21. Schaper, K.-J. and Samitier, M. L. R., Quant. Struct.-Act. Relat. 16 (1997) 224.
22. Devillers, J.; Domine, D. and Guillon, C., Eur. J. Med. Chem. 33 (1998) 659.
23. Huuskonen, J. J.; Salo, M. and Taskinen, J., J. Pharm. Sci. 86 (1997) 450.
24. Huuskonen, J. J.; Salo, M. and Taskinen, J., J. Chem. Inf. Comput. Sci., 38 (1998) 450.
25. Huuskonen, J.J., Rantanen, J. and Livingstone, D., Eur. J. Med. Chem., 35 (2000) 1081.
26. Huuskonen, J. J., J. Chem. Inf. Comput. Sci., 40 (2000) 773.
27. Huuskonen, J. J.; Villa, A. E. P. and Tetko, I. V., J. Pharm. Sci. 88 (1999) 229.
28. Huuskonen, J. J.; Livingstone, D.J. and Tetko, I. V., J. Chem. Inf. Comput. Sci., 40 (2000) 947.
29. Hall, L.H. and Kier, L.B., J. Chem. Inf. Comput. Sci. 35 (1995) 1039.
30. Szydlo, R.M., Ford M.G., Greenwood, R.G. and Salt, D.W., In Dearden, J.C. (ed.), Quantitative Approaches to Drug Design, Elsevier, Amsterdam, 1983, pp. 203–14.
31. Laass, W., In Seydel, J.K. (ed.), QSAR and Strategies in the Design of Bioactive Compounds, VCH, Weinheim, 1985, pp. 285–289.
32. Bordas, B., In Seydel, J.K. (ed.), QSAR and Strategies in the Design of Bioactive Compounds, VCH, Weinheim, 1985, pp. 389–392.
33. Ford, M.G. and Salt, D.W., In van de Waterbeemd, H. (ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 265–282.
34. Yalkowsky, S.H. and Dannenfelser, R-M. (ed.), AQUASOL dATAbASE of Aqueous Solubility, College of Pharmacy, University of Arizona, Arizona, USA, 1990.
35. Biobyte, Corp., 201 W. Fourth St., Suite #204, Claremont, CA 91711, USA.
36. Moriguchi, I., Hirono, S., Nakagome, I. And Hirono, H. Chem. Pharm. Bull., 42 (1994) 976–978.
37. Yalkowsky, S., Chemosphere 26 (1993) 1239–1261.
38. BMDP Statistical Software Manual, Dixon, W.J. (ed.) University of California Press, 1990.
39. Tetko, I.V., Villa, A.E.P. and Livingstone, D.J., J. Chem. Inf. Comput. Sci, 36 (1996) 794.
40. Wikel, J.H. and Dow, E.R., BioMed. Chem. Lett. 3 (1993) 645.
41. Kühne, R., Ebert, R.-U., Kleint, F., Scmidt, G. and Schüürmann, G. Chemosphere 30 (1995) 2061
42. Klopman, G., Wang, S., Balthasar, D.M., J.Chem.Inf.Comput.Sci. 32 (1992) 474
43. Stewart, D. K. and Love, W. A. Psychol. Bull., 70 (1968) 160.
44. Roadknight, C.M., Palmer-Brown, D. and Mills, G.E., In Liu, X., Cohen, P. and Berthold, M. (eds), Advances in Intelligent Data Analysis, Springer-Verlag, Berlin, 1997, pp. 337–346.