Prediction of Henry's Law Constants by a Quantitative Structure Property Relationship and Neural Networks

Niall J. English and Daniel G. Carroll*

Department of Chemical Engineering, University College Dublin, Belfield, Dublin 4, Ireland

Received January 16, 2001

Multiple linear regression analysis and neural networks were employed to develop predictive models for Henry's law constants (HLCs) for organic compounds of environmental concern in pure water at 25 °C, using a set of quantitative structure property relationship (QSPR)-based descriptors to encode various molecular structural features. Two estimation models were developed from a set of 303 compounds using 10 and 12 descriptors, one of these models using two descriptors to account for hydrogen-bonding characteristics explicitly; these were validated subsequently on an external set of 54 compounds. For each model, a linear regression and neural network version was prepared. The standard errors of the linear regression models for the training data set were 0.262 and 0.488 log(H_{cc}) units, while those of the neural network analogues were lower at 0.202 and 0.224, respectively; the linear regression models explained 98.3% and 94.3% of the variance in the development data, respectively, the neural network models giving similar quality results of 99% and 98.3%, respectively. The various descriptors used describe connectivity, charge distribution, charged surface area, hydrogen-bonding characteristics, and group influences on HLC values.

INTRODUCTION

Henry's law relates the equilibrium liquid and vapor phase concentrations of a solute in the limit of low solute concentrations. For dilute solutions of solute *i* at moderate pressures, one obtains Henry's law by equating the expressions for the vapor and liquid-phase fugacities

$$y_i P = x_i H_i(px) \tag{1}$$

where $H_i(px)$ is the HLC with dimensions of pressure. In the limit of infinite dilution, $H_i(px)$ may be defined as

$$H_i(px) = \gamma^{\infty}_{i(LR)} P_i^{SAT} \tag{2}$$

at low to moderate pressures.

Despite the importance of HLCs in a wide variety of engineering applications, experimental values are available for no more than a few thousand compounds.[1,2] The objective of the authors in undertaking this QSPR study was to develop a predictive model for HLC values for a broad spectrum of organic compounds with a set of computationally simple descriptors, which does not require molecular modeling or descriptor generation packages for their calculation on the part of the user, so that the nonspecialist in QSPRs can estimate HLC values quickly and conveniently. A second aim was to investigate the performance of artificial neural networks in comparison with conventional regression analysis, since previous QSPRs for HLCs have tended to concentrate almost exclusively on multiple linear regression.

PREVIOUS WORK

In previous QSPR studies for HLCs of organic compounds in water at 25 °C, the dimensionless gas-to-water concentration ratio, H_{cc}, has been used as the measure of the HLC. Since the range of typical H_{cc} values varies over more than 10 orders of magnitude, the common practice is to estimate $\log_{10} H_{cc}$ at 25 °C for computational tractability, which is usually within the range of −8 to 4 for most organic compounds. Therefore, $\log_{10} H_{cc}$ values will be estimated in this QSPR study. There are two empirical estimation methods: direct experimental determination of the HLC via static or dynamic equilibration techniques, or separate experimental measurement of the vapor pressure and aqueous solubility of the compound at the temperature of interest, and use of the ratio of vapor pressure to the solubility as a definition of the HLC. The appropriate conversion of dimensions to the H_{cc} form is then made. In the rest of this paper, the adjective "experimental" refers to the *direct* determination of the HLC (which is always preferable to the vapor pressure/solubility estimation method).

For a summary of the existing QSPR models for HLCs of organic compounds in water at 25 °C, the reader is referred to Table 1. As regards QSPRs for specific chemical classes, there have been four reported QSPRs for HLCs in water of PCB congeners[13−16] and also a QSPR for nonpolar liquids in molten polyisobutylene[17] between 25 and 150 °C, the exact temperature range depending on the solute.

GENERATION OF DESCRIPTORS

In this QSPR study, a set of 29 descriptors was generated for a dataset of 357 organic compounds composed of various chemical classes and diverse structural features. The descriptors were then used to predict $\log_{10} H_{cc}$ in water at 25 °C using two different models, each of which had a linear regression and neural network version. The various descriptors chosen may be grouped into several categories (refer to Table 2).

* Corresponding author phone: 353-1-706-1897; fax: 353-1-716-1177; e-mail: dan.carroll@ucd.ie.

10.1021/ci010361d CCC: $20.00
Published on Web 08/14/2001

Table 1. Summary of Previous QSPR Models for HLCs in Water at 25 °C

reference	classes of compounds	type of descriptors	statistics
Hine and Mookerjee[3]	292-compound database, comprising: hydrocarbons (aliphatic/aromatic, halogenated/nonhalogenated), ethers, sulfides, alcohols, phenols, aldehydes, ketones, carboxylic acids, esters, amines, nitriles, pyridines, mercaptans	initial model: 34 bond contribution factors; refined model: 50 groups contribution factor	bond contribution model: $n = 255$, $r^2 = 0.946$, $SE = 0.4$ group contribution model: $n = 215$, $r^2 = 0.996$, $SE = 0.11$
Cabani et al.[4]	hydrocarbons and monofunctional compounds	28 group contributions	$n = 209$, $SE = 0.09$
Nirmalakhandan and Speece[5]	hydrocarbons (aliphatic/aromatic, halogenated/ nonhalogenated), alcohols, phenols, esters, sulfides, carboxylic acids	11 group contributions to polarizability, Kier-Hall[6] connectivity indices, hydrogen bond index	initial model: $n = 180$, $r^2 = 0.932$, $SE = 0.445$ refined model: $n = 180$, $r^2 = 0.976$, $SE = 0.332$
Meylan and Howard[7]	345-compound database, distributed among 31 classes (218 empirical values, 127 by v.p./solubility method)	59 bond contributions, 15 correction factors	$n = 345$, $r^2 = 0.97$, $SE = 0.34$ test set of 74: $SE = 0.46$
Russell et al.[8]	72-compound database (subset of the data of Hine and Mookerjee[3]), not representative of all groups	five descriptors selected from 165 ADAPT[9] ones	training set of 63: $SE = 0.356$ test set of 7: $SE = 0.414$
Suzuki et al.[10]	229-monofunctional compound set (from ref 3) principal components analysis used	31 atomic/group contributions and ${}^1\chi$[6]	$n = 229$, $r^2 = 0.984$, $SE = 0.20$
Abraham et al.[11]	408-compound database, covering a broad range of organic compounds (exclusively experimental HLCs)	excess refraction, α_2^H[23] dipolarity, volume, β_2^H[24]	$n = 408$, $r^2 = 0.995$, $SE = 0.15$ $F = 16\,180$
Nirmalakhandan and Speece[12]	extension of earlier model[5] to aldehydes, ketones, pyridines, amines, nitro compounds	extra polarizability group contributions	test of earlier refined model[5] on 105 compounds: $r^2 = 0.953$

Table 2. Descriptors Used in the Study

bulk	molecular mass
	McCabe volume[20]
	Quale's parachor[18]
	polarizability[21]
	magnetic susceptibility[22]
connectivity	${}^1\chi^v$, χ_{pc}^{4v}[6]
	weighted path count
gross structure	aromaticity index[6]
	number of ring enclosures
charge	average charge per heteroatom
charged surf. area	weighted positively charged surf. area (WPSA)
	relative negatively charged surf. area (RNCS)
bonding	number of single bonds
	number of double bonds
	number of triple bonds
atomic contributions	number of heteroatoms (NHET)
	number of oxygen atoms (OX)
	number of nitrogen atoms (NIT)
	number of sulfur atoms (S)
	number of chlorine atoms (Cl)
hydrogen-bonding	effective H-bond acidity,[23] α_2^H
	effective H-bond basicity,[24] β_2^H
group contributions	COOH, OH, C, CH, CH_2, CH_3

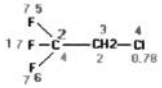
The various descriptors were chosen due to their inclusion in previous QSPR studies or because they were felt to be important in distinguishing pertinent topological or electronic features in molecules or particular classes of compounds. It ought to be noted that the numbers of double and single bonds in each hydrogen-suppressed molecular skeleton is counted using the Kekulé structure for benzene (i.e. three single and three double bonds).

The weighted path count descriptor was developed by the authors as a method of structural classification for atoms, as an alternative to the connectivity indices of Kier and Hall.[6] The algorithm to generate the path count is as follows: using a graph of the molecular skeleton, the non-hydrogen atoms are numbered from one to the last non-hydrogen atom, and each atom is assigned the appropriate Kier-Hall δ^v. For each pair of atoms, the number of bonds between the two atoms of interest is counted and attached a weighting $1/\sqrt{\Pi\delta^v}$, where $\Pi\delta^v$ is the product of the δ^v values along the particular path chosen. In the case of cyclic or aromatic molecules, the mean of the various weighted path lengths between the two molecules of interest is chosen. Each of these results for each pair of atoms is then added to give the overall weighted path count. The reader is referred to Chart 1 for example calculations for a halogenated ethane and 4-chloroaniline.

The path count is essentially a weighted type of Wiener index[25] adjusted for the presence of heteroatoms, not too dissimilar to the molecular identification number proposed by Randic.[26] However, Randic's moleculer I.D. number was derived only for molecules consisting entirely of carbon skeletons. As Randic[26] pointed out, the intrinsic weakness of connectivity indices is that they are too localized in nature and fail to account for the relationship of a given atom with the whole molecule. The fact that connectivity indices are sometimes identical for different isomers means that for unique classification of molecules, several indices must be generated. None of the weighted path counts were found to be identical for any of the compounds in the database of

Chart 1

2-chloro-1,1,1-trifluoroethane



The δ^V values are given in the diagram above (0.78 for Cl, 2 for CH2, 4 for C and 7 for F). Non-hydrogen atoms are numbered from 1 to 6.

Bond 1-2 : $1/\sqrt{7\cdot4} = 0.18898$

Bonds 1-3: $2/\sqrt{7\cdot4\cdot2} = 0.26726$

Bonds 1-4: $3/\sqrt{7\cdot4\cdot2\cdot0.78} = 0.45392$

Bonds 1-5: $2/\sqrt{7\cdot4\cdot7} = 0.14286$

Bonds 1-6: $2/\sqrt{7\cdot4\cdot7} = 0.14286$

Bond 2-3 : $1/\sqrt{4\cdot2} = 0.35355$

Bonds 2-4: $2/\sqrt{4\cdot2\cdot0.78} = 0.80064$

Bond 2-5 : $1/\sqrt{4\cdot7} = 0.18898$

Bond 2-6: $1/\sqrt{4\cdot7} = 0.18898$

Bond 3-4: $1/\sqrt{2\cdot0.78} = 0.80064$

Bonds 3-5: $2/\sqrt{2\cdot4\cdot7} = 0.26726$

Bonds 3-6: $2/\sqrt{2\cdot4\cdot7} = 0.26726$

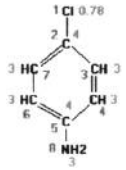
Bonds 4-5: $3/\sqrt{0.78\cdot2\cdot4\cdot7} = 0.45392$

Bonds 4-6: $3/\sqrt{0.78\cdot2\cdot4\cdot7} = 0.45392$

Bonds 5-6: $2/\sqrt{7\cdot4\cdot7} = 0.14286$

$\therefore$ Path Count = 5.11390

4-Chloroaniline



The δ^V values are given in the diagram above (0.78 for Cl, 4 for C, 3 for CH, 3 for NH2). Non-hydrogen atoms are numbered from 1 to 8.

1-2: $1/\sqrt{0.78\cdot4} = 0.56614$

1-3: $0.5(2/\sqrt{0.78\cdot4\cdot3} + 6/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4\cdot3\cdot3}) = 0.42122$

1-4: $0.5(3/\sqrt{0.78\cdot4\cdot3\cdot3} + 5/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.41926$

1-5: $0.5(4/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4} + 4/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4}) = 0.37742$

1-6: $0.5(5/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4\cdot3} + 3/\sqrt{0.78\cdot4\cdot3\cdot3}) = 0.41926$

1-7: $0.5(6/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4\cdot3\cdot3} + 2/\sqrt{0.78\cdot4\cdot3}) = 0.42122$

1-8: $0.5(5/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4\cdot3} + 5/\sqrt{0.78\cdot4\cdot3\cdot3\cdot4}) = 0.2724$

2-3: $0.5(1/\sqrt{4\cdot3} + 5/\sqrt{4\cdot3\cdot3\cdot4\cdot3\cdot3}) = 0.21378$

2-4: $0.5(2/\sqrt{4\cdot3\cdot3} + 4/\sqrt{4\cdot3\cdot3\cdot4\cdot3}) = 0.26290$

2-5: $0.5(3/\sqrt{4\cdot3\cdot3\cdot4} + 3/\sqrt{4\cdot3\cdot3\cdot4}) = 0.25000$

2-6: $0.5(4/\sqrt{4\cdot3\cdot3\cdot4\cdot3} + 2/\sqrt{4\cdot3\cdot3}) = 0.26290$

2-7: $0.5(5/\sqrt{4\cdot3\cdot3\cdot4\cdot3\cdot3} + 1/\sqrt{4\cdot3}) = 0.21378$

2-8: $0.5(4/\sqrt{4\cdot3\cdot3\cdot4\cdot3} + 4/\sqrt{4\cdot3\cdot3\cdot4\cdot3}) = 0.19245$

3-4: $0.5(1/\sqrt{3\cdot3} + 5/\sqrt{3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.23611$

3-5: $0.5(2/\sqrt{3\cdot3\cdot4} + 4/\sqrt{3\cdot4\cdot3\cdot3\cdot4}) = 0.26290$

3-6: $0.5(3/\sqrt{3\cdot3\cdot4\cdot3} + 3/\sqrt{3\cdot4\cdot3\cdot3}) = 0.28868$

3-7: $0.5(4/\sqrt{3\cdot3\cdot4\cdot3\cdot3} + 2/\sqrt{3\cdot4\cdot3}) = 0.27778$

3-8: $0.5(3/\sqrt{3\cdot3\cdot4\cdot3} + 5/\sqrt{3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.21378$

4-5: $0.5(1/\sqrt{3\cdot4} + 5/\sqrt{3\cdot3\cdot4\cdot3\cdot3\cdot4}) = 0.21378$

4-6: $0.5(2/\sqrt{3\cdot4\cdot3} + 4/\sqrt{3\cdot3\cdot4\cdot3\cdot3}) = 0.27778$

4-7: $0.5(3/\sqrt{3\cdot4\cdot3\cdot3} + 3/\sqrt{3\cdot3\cdot4\cdot3}) = 0.28868$

4-8: $0.5(2/\sqrt{3\cdot4\cdot3} + 6/\sqrt{3\cdot3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.21478$

5-6: $0.5(1/\sqrt{4\cdot3} + 5/\sqrt{4\cdot3\cdot3\cdot4\cdot3\cdot3}) = 0.21378$

5-7: $0.5(2/\sqrt{4\cdot3\cdot3} + 4/\sqrt{4\cdot3\cdot3\cdot4\cdot3}) = 0.26290$

5-8: $1/\sqrt{4\cdot3} = 0.28868$

6-7: $0.5(1/\sqrt{3\cdot3} + 5/\sqrt{3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.23611$

6-8: $0.5(2/\sqrt{3\cdot4\cdot3} + 6/\sqrt{3\cdot3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.21478$

7-8: $0.5(3/\sqrt{3\cdot3\cdot4\cdot3} + 5/\sqrt{3\cdot4\cdot3\cdot3\cdot4\cdot3}) = 0.21378$

$\therefore$ Path Count = 7.99697

this study, and its less localized nature, relative to Kier-Hall[6] indices, makes it an alternative to one, or perhaps more, connectivity indices in QSPR studies.

Prior to the calculation of charged surface area descriptors, it was necessary to develop a simple method to assign partial charges to atoms in the organic molecules typical of this

Table 3. STO-3G-Based Group Contributions for Partial Heteroatom Charges[a]

atom	partial charge
O (in C = O)	−0.2366
O (in C−O−C)	−0.2448
O (in C−O−H)	−0.2814
O (in C−N−O2)	−0.1859
N (in C−N−O2)	+0.1311
N (in C−N−H2)	−0.3952
N (in C2−N−H)	−0.3310
N (in C−N−C2)	−0.2640
N (in Car-Nar-Car)	−0.2349
N (in C ≡ N)	−0.1922
S (in C−S−H)	+0.1195
S (in C−S−C)	+0.1532
S (in Car-Sar-Car)	+0.2485
Cl (in C−Cl)	−0.1608
Br (in C−Br)	−0.0179
I (in C−I)	−0.0015
F (in C−F)	−0.1387

[a] Nar, Sar, and Car denote aromatic nitrogen, sulfur, and carbon atoms, respectively.

study, from which a set of group contribution factors for partial charges on heteroatoms in various functional groups could be defined. For a subset of 70 compounds of the development set, the STO-3G basis set method was used to generate electrostatic potentials (and hence find the associated partial atomic charges from a least-squares fit), while the PEOE (partial equalization of orbital electronegativity) method of Gasteiger and Marsili,[28] dependent only on atomic connectivity, was used also to compute the partial charges. In addition, the extended Hückel method was used to estimate the partial atomic charges for the subset. These compounds were chosen to be representative of the various functional groups of atoms and their interaction in the molecules of this study. The Hyperchem 6 software package was used to implement the PEOE algorithm, STO-3G basis set geometry optimizations and electrostatic potential calculations and the extended Hückel method. It was found that the partial charges derived from the STO-3G method were the most physically "reasonable", and it was noted in subsequent regression analysis that the charged surface area descriptors and the average charge on the heteroatoms (i.e. the sum of group contribution partial charges on the heteroatoms divided by the number of heteroatoms present) derived from the STO-3G method gave better statistics. Consequently, the STO-3G-based group contribution scheme for partial charges was adopted for this QSPR study. A summary of results for partial charges on heteroatoms in various functional groups is given as Supporting Information. From these results, the arithmetic average charge was taken on each heteroatom in a particular functional group to prepare the group contribution table (see Table 3).

The relative negatively charged surface area (RNCS) and weighted positively charged surface area (WPSA) were included for calculation, because they were found to be statistically significant for HLC prediction in the QSPR study of Stanton and Jurs.[27] The group contribution method developed by the authors was used to assign partial charges to heteroatoms in their calculation (defined below).

The WPSA is defined as

$$WPSA = (SA)\sum SA_i^+ \qquad (3)$$

where SA is the total molecular surface area, and SA_i^+ denotes the surface area of the positively charged atom i. To obtain an estimate for the solvent-accessible cavity area of the molecule in water at 25 °C, the bond contribution method of Bondi[29] was used. Once the total water-accessible cavity surface area has been computed, the $\sum SA_i^+$ term is found by adding together the Bondi area contributions of the positively charged atoms, as found from experience with the results of the STO-3G model.

The RNCS is defined as

$$RNCS = \frac{(QNEG)(SA_{mneg})}{\sum Q_i^-} \qquad (4)$$

where $QNEG$ is the absolute value of the partial charge on the most negatively charged heteroatom, SA_{mneg} is the Bondi surface area[29] of the most negatively charged heteroatom, and $\sum Q_i^-$ is the absolute value of the sum of the partial charges on the negative heteroatoms. For the heteroatoms, the partial charges were determined using the derived table of partial charge contribution factors, while $QNEG$ was determined from the set of partial charges generated by the STO-3G method. RNCS is defined as zero for all molecules in which negatively charged heteroatoms are absent.

The reason the Bondi bond contribution technique was used to calculate the WPSA and the RNCS is because of its relative simplicity, meaning that the user does not have to resort to computer molecular modeling or descriptor generation packages; further, it was found that reasonable intuition about whether atoms have a positive or negative partial charge tended to coincide well with the results of STO-3G calculations. This simplifies the calculation of the WPSA and RNCS, making it more convenient for the nonexpert in descriptor generation or QSPRs.

The effective hydrogen bond acidities[23] α_2^H and basicities[24] β_2^H were used as descriptors to enhance the predictive capacity of the models for sets of compounds where it was suspected that polar interactions, attributable to hydrogen bonding of acidic and basic functional groups with the surrounding water molecules and from their intramolecular interactions, played a significant role in determining the aqueous solubility (and hence the HLC). It was found that the hydrogen bond acidities and basicities improved the statistical performance of the models significantly, since none of the other descriptors account for hydrogen bonding explicitly, although Abraham et al.[23,24] have found that there is no general correlation between α_2^H and β_2^H and proton-transfer characteristics for organic compounds, although some relationships can be established for particular classes of compounds. Nevertheless, the inclusion of α_2^H and β_2^H descriptors is a reasonably effective measure to account for polar interaction effects on HLCs at 25 °C. These hydrogen bond acidities and basicities are derived from log K values for complexation of the organic compound in tetrachloromethane and as such are not strictly nonexperimental descriptors, but values are tabulated for many "typical" organic compounds.[11,23,24] Therefore, two predictive models were developed, one (model 1) which does not contain these hydrogen bond descriptors, meaning that one must select an optimum subset of descriptors from a set of 27 and another (model 2) which may well involve these descriptors (i.e.

Table 4. Classes of Compounds Used as a Database and Distribution among Development and Test Sets

chemical class	train	validation	test
alkanes	17	3	4
cyclic alkanes	5	1	1
alkenes	14	2	3
cyclic alkenes	4	0	0
alkynes	4	1	1
halomethanes	7	1	1
haloethanes	8	1	1
halopropanes	7	1	1
halobutanes	4	1	1
other haloalkanes	9	2	2
haloalkenes	4	0	1
ethers	6	1	2
aldehydes	8	2	2
ketones	13	3	3
esters	22	4	4
nitriles	3	0	1
aliphatic amines	10	2	2
aliphatic nitrogenous	4	1	2
aliphatic sulfurous	5	1	1
aliphatic acids	5	0	1
aliphatic alcohols	18	3	3
monoaromatics	17	3	3
PAHs	6	1	1
pyridines	13	2	3
haloaromatics	9	1	3
anilines	9	1	2
phenols	16	3	2
other aromatic alcohols	4	0	1
other aromatics	6	1	1
PCBs	4	0	1
total	261	42	54

variable selection from the whole set of 29 descriptors), so that the user can invoke model 1 in the case where the α_2^H and β_2^H descriptors are unknown for a particular compound.

CONSTRUCTION OF DATABASE

Reliable experimental data for the HLCs in water at 25 °C, in the form of log(H_{cc}), were selected for inclusion in this study from various sources.[3,4,16,30−36] The range of the log(H_{cc}) values was from −7.08 to 2.32.

These compounds were assigned to 30 separate classes, defined in Table 4. For the development of a general QSPR for organic solutes, balance, in terms of adequate representation of structural diversity in organic molecular topographies, is essential. With this in mind, a development set of 303 compounds and a test set of 54 chemicals was selected. In the case of neural network analysis, the development set was divided into a training set of 261 compounds and a validation set of 42 compounds. In constructing the three sets, the general consensus for neural analysis is that about 70% of a mixed set of data ought to be used for the training set, roughly 10−15% as a validation set, and approximately 10−15% as a test set. In the case of this study, these proportions were 73%, 12%, and 15%, respectively. This general approach in terms of establishing the number of compounds in various sets was also applied to each of the 30 chemical classes individually. As regards which compounds from a particular class were assigned to a given set, the distribution was chosen to be balanced so that a particular isomer of some compound might appear in the test set when one or more examples occurred in the training set; the same principle was extended to the selection of compounds for the validation set. For instance, in the case of the halogenated butanes, 1-chlorobutane is present in the training set, leaving one free to put 2-chlorobutane in the test set. 1,4-Dichlorobutane is in the training set, which allows one to put 1,1-dichlorobutane in the validation set. This approach ensured that the test and validation sets were representative of the training set.

In terms of the number of compounds assigned to each chemical class, there are two key considerations: experimental data (or lack thereof) and balance. The overall numbers of compounds assigned to each class and proportions of compounds in classes were chosen to preserve the generality of the model so that certain molecular features cannot become dominant at the expense of predictive accuracy for other compounds, i.e., balance of characteristics such as aliphatic versus aromatic, halogenated or nonhalogenated, monoaromatic versus polyaromatic, aliphatic alcohols versus phenols. Naturally, the numbers of compounds and exact molecules in each class are limited to the extent that experimental HLCs at 25 °C are available. The reliability of the experimental values themselves must also be taken into account as well as the range of sources from which they originated: depending on the empirical technique and conditions, the standard deviation in HLC measurements can range from less than 0.05 to about 0.5 log(H_{cc}) units.

DEVELOPMENT OF THE MODELS

The 29 descriptors were evaluated for each compound. In the case of the two connectivity indices $^1\chi^V$ and $^4\chi_{pc}^V$ and the weighted path count, Fortran 90 programs were devised. For the remainder, the descriptors were entered/calculated using Microsoft Excel spreadsheets. For the multiple linear regression analysis, a Fortran 90 program was prepared to implement the singular value decomposition (SVD) algorithm, using prewritten subroutines,[37] along with calculation of the variance inflation factors[38] (VIFs) for the given set of independent variables analyzed. The linear regression program also calculated the correlation matrix for the given descriptors, using scaled and both centered and uncentered data,[38] and the associated eigenvalues and eigenvectors for the purposes of finding variance decomposition proportions,[38] to detect multicollinearities in the data. The eigenvalues and eigenvectors of the two correlation matrices were calculated using a Householder reduction, followed by the QL decomposition,[37] implemented in Fortran using a prewritten subroutine.[42] Routine parameters, such as Mallow's C_p, the adjusted r^2, the PRESS statistic, and the Fisher F statistic[38] were computed also, along with t-tests on each of the regression parameters to determine the probability value at which the t-statistic is significant. To implement the t-test, for the given degrees of freedom, prewritten Fortran subroutines[42] were used. The calculation of Mallow's C_p used the residual mean square (RMS) for the complete set of descriptors (i.e. either 27 in the case of model 1 or 29 in the case of model 2) as an estimate for the population variance of the corresponding model. For multiple nonlinear regression, the Levenberg−Marquardt algorithm was implemented in Fortran 90, using prewritten subroutines,[37] along with the various diagnostics outlined above, apart from the variance decomposition proportions and VIFs, which are valid only for linear models. In practice, however, nonlinear regression fits were applied only after "good" subsets, without inter-

correlation problems, had been selected by linear regression. For neural analysis, an SPSS Neural Connection 2.0 software package[39] was used.

Prior to statistical analysis, each descriptor was subject to a simple linear transformation, e.g. division of molar mass in g/mol by 100, to render the subsequent regression and neural analysis more tractable computationally. The various descriptors were thus transformed to the general range of 0 to about 25, so that the computer algorithms would not become ill-conditioned due to manipulation of very large or small numbers. Equal weightings were used for each compound in the regression analysis, since the standard deviations of experimental HLCs were known for few compounds.

Prior to selection of optimal descriptor subsets, a Fortran intercorrelation screening program was written to compute the absolute values of the pairwise correlation coefficient of each pair of the 29 descriptors (i.e. the scaled and centered correlation matrix less the row and column for the intercept). It was found that the five descriptors encoding the molecular bulk were intercorrelated significantly, as one might expect, along with $^1\chi^V$. Ideally, of these six descriptors, the preferred ones for inclusion in a QSPR model would not include $^1\chi^V$, since this is a more abstract concept than the other five quantities. Other collinearities between certain descriptors became apparent at this stage, which is valuable knowledge before application of descriptor selection régimes. There is also the additional problem of arriving at fortuitous correlations if too many descriptors are screened relative to the number of molecules in the development set, as the work of Topliss[40] showed; the authors decided, possibly somewhat conservatively, to limit the maximum number of independent regression parameters allowable to about 15, using a guideline of at least 20 values of the dependent variable per parameter.

To select the optimal subsets of descriptors which are most significant statistically in estimating HLCs, the method of regression by leaps and bounds of Furnival and Wilson[41] was implemented in Fortran using a prewritten subroutine.[42] For the generation of descriptors in model 1 (which omits α_2^H and β_2^H), the leaps and bounds regression was applied to 27 descriptors (excluding the acidity and basicity from the selection), while all 29 descriptors were screened in the selection of parameters for model 2. The tacit assumption of this variable selection procedure is that linear regression is actually an *adequate* method to model the interaction of a set of variables in what may well be a nonlinear process (i.e. the Gibbs free energy for aqueous solubility at infinite dilution), apart from the separate matter of whether the set of 29 descriptors contains sufficient information relevant to the physical process being modeled. As a first approximation, this served as a tool for descriptor selection, which then identified the variables to be used in the more rigorous nonlinear regression and neural models.

Selection of Subset of Descriptors for Model 1 and Linear Regression Version of Model 1. Regression by leaps and bounds was applied to the set of 27 descriptors (i.e. excluding the acidity and basicity descriptors), and three selection criteria were used in this process: models with the highest r^2 per subset size, models of all subset sizes with the lowest Mallow's C_p statistic, and regression subsets with the highest adjusted r^2. It was found that the condition numbers of the scaled and centered correlation matrix for the models with the best C_p and adjusted r^2 values were all in excess of 30, which confirms collinearity problems, according to Belsley et al.[43] This is because these models tended to have a larger number of variables, i.e., 15 or more, and hence more than the authors would desire, as commented earlier. This led to over-fitting of these models. Instead, a strategy was adopted which entailed the examination of the subsets of each size with the highest r^2. The various criteria considered were the *t*-tests on the regression parameters (including the constant term), the variance decomposition proportions of the two correlation matrices, the r^2, Mallow's C_p, F, and PRESS statistics as well as the standard error SE. It was found that the most appropriate model, which still avoided collinearities (i.e. a scaled-centered correlation matrix condition number below 30 and no excessive variance proportions in the eigenvectors from only one or two "dominant" eigenvalues, and all VIFs less than 10), was a 12-descriptor one (see Table 5). [These 12 descriptors (in transformed form, prior to SVD regression analysis), along with descriptors for model 2, are listed as Supporting Information.]

Table 5. Details of the Linear Regression Version of Model 1

variable	coeff	uncertainty	VIF	*t*-test value	probability
constant term	0.7320	0.1197		6.116	≈0
$^1\chi^v$	−0.2994	0.06487	5.863	−4.616	≈0
no. of C groups	−0.3228	0.04161	3.125	−7.758	≈0
no. of CH groups	−0.1478	0.02067	2.494	−7.150	≈0
no. of CH_3 groups	0.2183	0.03654	1.805	5.974	≈0
average charge	6.3888	0.3604	2.890	17.73	≈0
no. of chlorines	0.4050	0.05176	2.453	7.283	≈0
no. of nitrogens	−1.5320	0.09652	1.792	−15.87	≈0
no. of OH groups	−1.6867	0.09424	1.534	−17.90	≈0
no. of O atoms	−0.7224	0.05379	1.958	−13.43	≈0
RNCS	−0.3231	0.05251	1.432	−6.153	≈0
no. of S atoms	−2.2525	0.1794	1.181	−12.56	≈0
WPSA	1.1408	0.1741	5.612	6.552	≈0

The performance indices were $r^2 = 0.943$, adjusted $r^2 = 0.941$, PRESS $r^2 = 0.937$, adjusted PRESS $r^2 = 0.934$, $F = 400$, Mallow's $C_p = 30$, $SE = 0.488$ log units, condition number of the scaled-centered correlation matrix = 28.3, and condition number of the scaled-uncentered correlation matrix = 466.

As regards the nonlinear regression analysis, it was suspected that the number of certain heteroatoms, e.g. oxygen or nitrogen, may be correlated better as reciprocal terms in the expression for $\log(H_{cc})$: this is because as the mass of a molecule increases it may partition more in the liquid phase (hence having a lower H_{cc}). Various model expressions which were nonlinear in the regression coefficients were attempted with the development database using Levenberg−Marquardt regression, but no real statistical improvement over the model expression linear in coefficients was found: while the r^2 and F were increased and the standard error decreased slightly, it was found that the *t*-tests for some of the regression parameters began to indicate significance, which is not acceptable for a stable regression scheme.

Development of Neural Network Version of Model 1. Artificial neural networks were used to develop another predictive scheme using the descriptors identified for model 1. Essentially, neural networks use a collection of "percep-

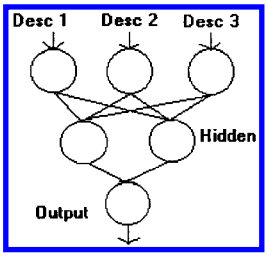


Figure 1. Schematic of a fully-connected, feed-forward neural network with a hidden layer.

Figure 2. Prevention of over-training of the neural network by cross-validation.

trons" to perform processing and logical operations. A typical fully connected, feed-forward neural network would possess the architectural configuration outlined in Figure 1, with the perceptrons arranged in a series of layers.

Various descriptors encoding parameters governing a problem are entered in the input layer and are processed by the hidden and output layers to yield an output signal. The hidden layer of perceptrons is employed to offer greater computational flexibility to the network. An excessive number of perceptrons in the hidden layer(s) leads to a situation in which the target values are predicted perfectly for the development set, but where the network cannot generalize its applicability to an external test set (a condition known as "over-training"). Too few leads to a loss in predictive accuracy.

Signals S_i from the perceptrons in the previous layer are attached corresponding weights w_i and summed to produce a net signal

$$\therefore \text{ net signal} = \sum_i w_i s_i + \theta \tag{5}$$

where θ is an adjustable bias term. A nonlinear activation or threshold function then operates on this net signal to generate an output signal (the most widely used function is perhaps the sigmoidal function)

$$\text{output signal} = \frac{1}{1 + e^{-net \cdot signal}} \tag{6}$$

The value of the output signal is then compared to the target value for that particular set of descriptors. An optimization algorithm is used to adjust the network weights and biases to produce a final set of output values which correspond to the target values as closely as possible. The criterion for the optimum choice of network is the one which minimizes the sum of the squares of the deviations of the target values from the output values of the network. The sum of squares error E is given by

$$E = \sum_i (t_i - o_i)^2 \tag{7}$$

where t_i is the ith target value (e.g. HLC for the ith compound) and o_i is the corresponding output value. There are two common methods for back-propagation training of the network, i.e., steepest descent and the conjugate gradient approach. The steepest descent method is somewhat less effective in comparison to the quasi-Newton conjugate gradient method. The ultimate minimized value of E may not be the global minimum for the error surface: the initial distribution of weights and biases can sometimes play an influential role in determining the ultimate outcome of the error. In the case of this application, the steepest descent minimization method[37] was not used; it was a method very similar to the Broyden, Fletcher, Goldfarb, Shanno (BFGS) quasi-Newton algorithm[37] which was used by the SPSS neural analysis package.[39] In this case, one starts at the kth error $E(\vec{x}_k)$ and then moves to the (lower) error $E(\vec{x}_{k+1})$ in the direction of a "weighted" gradient composed of a combination of the vectors $\nabla\vec{E}_k$ and $\nabla\vec{E}_{k+1}$, where $\nabla\vec{E}_{k+1} = \nabla\vec{E}_k + H_k \Delta\vec{x}_k$ and $\Delta\vec{x}_k = \vec{x}_{k+1} - \vec{x}_k$, H_k being the Hessian matrix at the given $\vec{x}_k$. The error $E(\vec{x}_{k+1})$ is then given approximately by a Taylor series in $\vec{x}$:

$$E(\vec{x}_{k+1}) \approx E(\vec{x}_k) + (\nabla\vec{E}_k)^T \Delta\vec{x}_k + \frac{1}{2}(\Delta\vec{x}_k)^T H_k \Delta\vec{x}_k \tag{8}$$

Initially, a search direction $\vec{d}_k$ is chosen such that $\Delta\vec{x}_k \propto \vec{d}_k$ and this allows an estimate for H_k, G_k, to be found (it would be too expensive in terms of computer time to calculate the Hessian matrix analytically at each step in the optimization). Knowledge of G_k allows a new search direction $\vec{d}_k$ to be initiated which will reduce E_{k+1} further. This iterative process is repeated once E_{k+1} is minimized, the k value is incremented, and the process is continued until convergence of the errors E to a minimum value occurs. The BFGS-style conjugate gradient algorithm is significantly more adept at avoiding false minima than the steepest descent methods and is more much likely, but never *guaranteed*, to achieve global convergence.

One essential requirement for neural network analysis of a set of data is that the development data should be balanced and a validation set of data, representative of the training set, ought to be present to prevent over-training and over-fitting, i.e., the situation in which the neural model predicts all of the target values in the development set perfectly (it has learned the "noise" therein), but is incapable of generalizing to estimate the desired property for an external test set. Cross-validation, with a balanced set of target values and descriptors, chooses the most appropriate model (found from the development set) as the one which minimizes the sum of the squares of the deviations (E) for the validation set, as the number of cycles of the minimization algorithm increases with the passage of time. Since the validation set is not used in the training of the neural model, it acts as a useful indicator of predictive capacity. The "best" model is then used to predict the target values for the external test set. This process is depicted in Figure 2.

As regards stopping criteria for the development process, an RMS error plot of the training and validation sets was monitored in the SPSS package[39] as the number of cycles increased. The number of cycles was set beforehand to 1500 for each run, which was found by experience to provide ample time for the given system of descriptors and the

Table 6. Specifications for the Neural Network Version of Model 1

configuration	12−4−1
descent method	conjugate gradient
weight distribution	uniform
initial weights	range = 0.1, seed = 1
node function	tanh
number of runs	3000
SE (training set)	0.224 (log units)
SE (validation set)	0.244 (log units)

Table 7. Details of the Linear Regression Version of Model 2

variable	coeff	uncertainty	VIF	*t*-test value	prob-ability
constant term	0.49956	0.049105		10.17	≈0
α_2^H	−4.0232	0.091383	1.305	−44.02	≈0
β_2^H	−4.9084	0.11236	2.583	−43.68	≈0
no. of C groups	−0.24292	0.019897	2.479	−12.21	≈0
no. of CH_2 groups	0.12655	0.008623	1.328	14.68	≈0
no. of CH_3 groups	0.35834	0.017676	1.466	20.27	≈0
no. of N atoms	−0.70381	0.051691	1.783	−13.62	≈0
no. of O atoms	−0.47417	0.029497	2.044	−16.08	≈0
no. of ring enclosures	−0.39227	0.041268	2.856	−9.505	≈0
RNCS	−0.41016	0.025868	1.206	−15.86	≈0
no. of sulfur atoms	−0.66568	0.094045	1.127	−7.078	≈0

dependent variable to experience a prolonged plateau in both the training and validation set RMS errors. Various architectures were constructed to find the optimum model, e.g. varying the number of perceptrons in the hidden layer, varying the node activation function from tanh to sigmoidal and the initial distribution of perceptron weights. After numerous runs employing various different combinations of models, the number of cycles was increased to 3000 for the best few, but this led to no change in the results for the runs (i.e. the optimal configuration of weights which minimized the least squares error for the validation set had been located before 1500 cycles, in what was hopefully the actual global minimum of the error surface in hyperspace). The details of the network judged to be the most superior, in terms of lowest RMS error for the validation set (and also, incidentally, the training set), are given in Table 6.

The r^2 values were $r^2 = 0.987$ (training set) and $r^2 = 0.982$ (validation set).

Selection of Subset of Descriptors for Model 2 and Linear Regression Version of Model 2. Regression by leaps and bounds was applied to the set of 29 descriptors (i.e. including the acidity and basicity descriptors), and similar patterns were observed for the regression models with the best adjusted r^2 and Mallow's C_p values (i.e. "too many" descriptors which led to collinearity problems). The approach of using the highest r^2 models for each subset size, as outlined in the development of model 1, produced the model in Table 7.

The performance indices were $r^2 = 0.984$, adjusted $r^2 = 0.983$, PRESS $r^2 = 0.982$, adjusted PRESS $r^2 = 0.981$, $F = 1{,}738$, Mallow's $C_p = 82.7$, $SE = 0.262$ log units, condition number of the scaled-centered correlation matrix = 14.5, and condition number of the scaled-uncentered correlation matrix = 82.6.

Various model expressions which were nonlinear in the regression coefficients were attempted with the development database using Levenberg−Marquardt regression, but no real statistical improvement over the model expression linear in coefficients was found: it was found that the *t*-tests for some of the regression parameters began to indicate significance, as was noted in the development of model 1.

Table 8: Specifications for the Neural Network Version of Model 2

configuration	10−3−1
descent method	conjugate gradient
weight distribution	uniform
initial weights	range = 0.35, seed = 1
node function	tanh
number of runs	3000
SE (training set)	0.202 (log units)
SE (validation set)	0.157 (log units)

Table 9. Comparative Summary of Model Performance for the Development Set

model	r^2	*F*	*SE* (log units)
linear (1)	0.943	400	0.488
neural (1)	0.987 (training)		0.224 (training)
	0.982 (validation)		0.244 (validation)
linear (2)	0.984	1,738	0.262
neural (2)	0.999 (training)		0.202 (training)
	0.992 (validation)		0.157 (validation)

Development of Neural Network Version of Model 2. Various architectures were constructed and combinations of network settings employed to find the optimum model in terms of the set of 10 descriptors, e.g. varying the number of perceptrons in the hidden layer, varying the node activation function and the initial perceptron weights, as was outlined in the discussion for model 1. Again, 1500 cycles in each run was found to produce satisfactory convergence for the RMS errors of the training and validation sets. The number of cycles was again increased to 3000 for the best few, as was the case for model 1, but this led to no change in the results for the runs. The details of the network judged to be the best, in terms of lowest RMS error for the validation set (and also, perhaps unsurprisingly, the training set), are given in Table 8.

The r^2 values were $r^2 = 0.99$ (training set) and $r^2 = 0.992$ (validation set).

A comparative summary of the performance of models 1 and 2, for the development set, is given in Table 9.

The predicted values for log(H_{cc}) in water at 25 °C for both the linear regression and neural network versions of models 1 and 2, along with the original experimental values (and their associated literature sources) are listed as Supporting Information, including predictions for the test set.

Defining outliers as compounds for which the estimation error exceeds twice the standard deviation for that particular model, then those for the development set are specified in Table 10.

VALIDATION OF MODELS

There are several ways to assess the validity of the developed models, e.g. 1. plotting of empirical versus predicted log(H_{cc}) values for the development set of each model, 2. plotting of residuals versus empirical log(H_{cc}) values for the development set of each model, and 3. application of the models to a balanced external test set.

It ought to be noted that the plotting of predicted versus actual results and residuals is certainly not as rigorous a validation process as application of the models to an

Table 10. Outliers for the Linear and Neural Versions of Both Models for the Development Set

model 1, linear (11)	1,2-dichloroethane	
	1,1,1,2-tetrachloroethane	
	1,4-dichlorobutane	
	1,4-dioxane	
	cyclohexanone	
	acetonitrile	
	propanonitrile	
	n-methylmorpholine	
	4-acetylpyridine	
	chlorobenzene	
	ethyl benzoate	
model 1, neural (14)	methylcyclohexane	(validation)
	butyne	(training)
	tetrachloromethane	(training)
	1,2-dichloroethane	(training)
	iodoethane	(training)
	1,4-dichlorobutane	(training)
	1,4-dioxane	(training)
	octanal	(training)
	nonanal	(training)
	cyclohexanone	(training)
	4-formylpyridine	(training)
	iodobenzene	(training)
	4-chloroaniline	(validation)
	2-naphthol	(training)
model 2, linear (19)	methane	
	ethene	
	dichloromethane	
	iodomethane	
	1,2-dichloroethane	
	2-chloro-2-bromo-1,1,1-trifluoroethane	
	iodoethane	
	1,2-dichloropropane	
	2-bromo-2-methylpropane	
	1,4-dichlorobutane	
	2-hexenal	
	2-octenal	
	1-cyanobutane	
	2,2,2-trifluoroethanol	
	pyrene	
	iodobenzene	
	3-cyanophenol	
	ethyl benzoate	
	PCB 2,2′,3,3′,4,4′	
model 2, neural (18)	cyclopentene	(training)
	iodomethane	(training)
	iodoethane	(training)
	2-bromo-2-methylpropane	(training)
	1,4-dichlorobutane	(training)
	diethyl ether	(training)
	2-hexenal	(training)
	2-octenal	(training)
	2,4-dimethyl-3-pentanone	(training)
	cyclpropyl-CO2-methyl	(training)
	1-cyanobutane	(training)
	prop-2-en-1-ol	(training)
	1-ethylnaphthalene	(validation)
	iodobenzene	(training)
	4-chlorophenol	(validation)
	1-naphthol	(training)
	ethyl benzoate	(training)
	PCB 2,2′3,3′,4,4′	(training)

independent test set but ought be performed regardless, as a matter of routine, to observe if there are any patterns in the outliers. A plot of the experimental log(H_{cc}) values versus the predicted values for the development set of model 1 is given in Figure 3, while the corresponding plot for model 2 is supplied in Figure 4. Each of the best-fit lines had slopes of unity and a y-intercept of zero, thus indicating the goodness of fit. Plotting of the residuals for each model

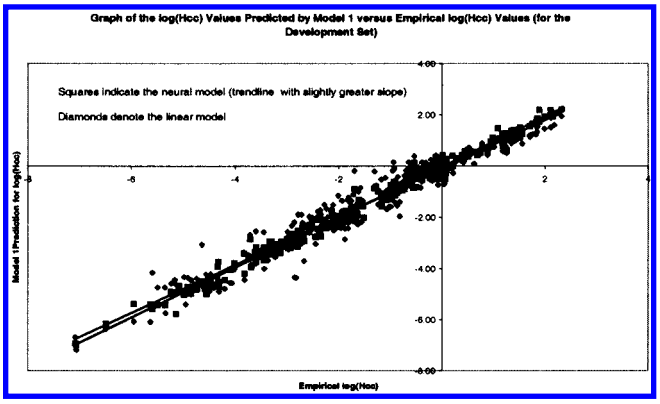


Figure 3.

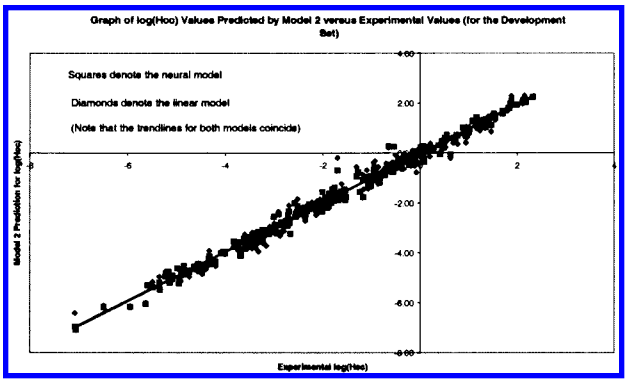


Figure 4.

Table 11. Performance Indices for Application of Models 1 and 2 to the Independent Test Set

model	r^2	*SE* (log units)
model 1, linear	0.925	0.533
model 1, neural	0.979	0.281
model 2, linear	0.980	0.285
model 2, neural	0.985	0.237

appeared to produce random distributions, so these plots are omitted.

The final validation test is to apply each version of both models to the external test set of 54 compounds. The results for the test set have been specified with those of the development set. The following statistics and outliers (residuals in excess of two standard deviations for the relevant version of the model) were obtained for the test set and are supplied in Tables11 and 12.

DISCUSSION

It has been seen that the neural versions of both models for log(H_{cc}) at 25 °C possess a more significant degree of predictive accuracy vis-à-vis the linear regression versions: the standard errors of about 0.224 (model 1) and 0.202 (model 2) for the 261-compound training set are reasonable for 12- and 10-descriptor correlations, respectively. In comparison with the other QSPR models in the literature, the simple models developed in this study perform reasonably well: the linear model of Nirmalakhandan and Speece[5] requires 19 parameters (17 optimized atomic and group contribution factors in the polarizability term, along with a connectivity descriptor and hydrogen bond index) to give an *SE* of 0.262 log units in the development set, while that of Meylan and Howard[7] requires 59 bond contribution factors

Table 12. Outliers for the Linear and Neural Versions of Both Models for the Test Set

model 1, linear (3)	methyl cyclpropyl ketone
	morpholine
	2-methoxyphenol
model 1, neural (5)	2-chlorobutane
	methyl cyclpropyl ketone
	morpholine
	3-formylpyridine
	2-methoxyphenol
model 2, linear (4)	1,1-dichloroethane
	1-cyanopropane
	1,4-dibromobenzene
	methyl benzoate
model 2, neural (5)	1,1-dichloroethane
	methyl cyclpropyl ketone
	1-cyanopropane
	1,4-dibromobenzene
	methyl benzoate

to result in an *SE* of 0.34 in its development set. The model of Abraham et al.[11] compares well against model 2 of this study (since model 2 and that of Abraham et al. contain α_2^H and β_2^H as descriptors), offering a standard deviation of 0.151 for a linear five-descriptor model, based on a 408-compound training set (but with no validation on an external test set). Ideally, the neural and linear models ought to be compared with the other QSPR models for the 150-compound set in Brennan et al.,[44] to give an entirely objective basis for comparison.

The weaknesses of the models are illustrated by somewhat lackluster performances by model 1 for a few halogenated alkanes, two iodine-substituted compounds (overestimation of the HLC), two cyclic ketones (overestimation), two nitriles (underestimation), two morpholines and 4-acetylpyridine (overestimation) and by model 2 for a few halogenated alkanes, three iodine-substituted compounds (overestimation), a few cyano-compounds (overestimation), and a few aldehydes and ketones. Interestingly, there are some different classes of compounds for which the models falter: model 2, for instance, is superior for nitriles, cyclic ketones, morpholines, and 1,4-dioxane but is less reliable than model 1 for cyano-compounds. Further, there is general agreement between the outliers for the linear and neural versions of model 2, while there are some differences in the types of outliers for the linear and neural versions of model 1. Examination of Tables 9 and 11 shows that the gross predictive performances of the neural and linear versions of model 2, as characterized by r^2 and the standard error, are quite similar; this may suggest that the neural scheme does not depart too greatly from a linear form, and the fact that the outliers are similar for both versions lends some credence to this notion. However, it is seen that the neural version of model 1 offers a marked improvement in overall performance relative to the linear version, implying that the neural scheme may depart significantly from a linear form and that nonlinear interactions between the descriptors become more important. Therefore, it is perhaps not unexpected that the structure of outliers should differ more in the case of model 1: for instance, the linear scheme overestimates HLCs for morpholines and 4-acetylpyridine as well as underestimates HLCs for some nitriles, while the neural method overestimates for two iodine-substituted compounds.

There are essentially five reasons underlying these predictive aberrations: 1. inaccurate experimental data and variation of data, 2. errors in the descriptors, 3. the set of descriptors used does not contain sufficient information relevant to the physical process being monitored, 4. distant polar interactions, and 5. insufficient number of compounds with given structural feature in training set.

In connection with the issue of likely accuracy of the empirical HLC data, one must take into account the problem of using different literature sources for different isomers of compounds, in addition to the type of experimental technique used in HLC determination and its probable range of standard error. For instance, the HLC of −0.51 for 1,1-dichlorobutane, from Hine and Mookerjee,[3] differs a great deal from the value of −1.7 for 1,4-dichlorobutane, from ref 31: this large difference in HLCs for isomers from different literature sources may explain why both versions of both models overestimate the HLC for 1,4-dichlorobutane. Similar remarks may apply for *cis*-1,2-dichloroethene (HLC of −0.86, from ref 31) and *trans*-1,2-dichloroethene (HLC of −0.56, from Hine and Mookerjee[3]), even though neither of these compounds are outliers. Variation in empirical data is an important issue, in that different experimental techniques in different laboratories produce a range of standard errors of measurement, as mentioned previously in the "Construction of Database" section.

The last point seems to be less of a problem in the case of this study, due to the balanced nature of the database: the fact that different models perform differently for different chemical classes suggests that it is the model per se which may not be suited to a particular subset of compounds. However, it is also perhaps true that a greater number of compounds are required in the alkyne, ether, and sulfide classes. Class-specific descriptors were added to the set of descriptors after initial trials, to improve predictive accuracy: this was very effective for nitrogenous compounds (number of nitrogen atoms) and for sulfur-containing bonds (number of sulfur atoms) but less successful for double and triple bond descriptors.

The third point is of great relevance to any QSPR study: the philosophy of the authors in undertaking the research was, as stated previously, to develop simple descriptors which could be calculated by the (nonspecialist) user to eschew the use of molecular modeling or QSPR descriptor generation software. The motivation to calculate some of the descriptors for the database of compounds stemmed from their use in previous QSPR studies of HLCs, so it is quite possible, even probable, that there are some as yet unidentified molecular properties which exert a strong statistical influence on the physical chemistry of vapor−liquid partitioning. Bearing this point in mind, a set of 29 descriptors cannot claim to represent every possible aspect representing this physical process. Further, the nature of the descriptors such that their calculation can be rendered manual or semiautomatic with spreadsheets or very simple computer programs makes their calculation for a large database a lengthy process, albeit much more attractive for users with little experience in the field of QSPRs and descriptor generation or access to molecular modeling packages. This observation impacts also on the second issue outlined above, namely that of the accuracy of the descriptors: the approximations involved in the calculation of the RNCS, for instance, could conceivably lead to a predictive error in the case of-iodine-substituted compounds. The group contribution factor for the partial charge on an

iodine atom in organic compounds similar to the ones of this study is −0.0015: it was found that the partial charges on the iodine atoms in the compounds used to derive this group contribution factor were both positive and negative, within the range −0.04 to 0.04, i.e., the contribution factor is within the "noise", as it were, of the original data. Consequently, it is sometimes difficult to establish whether an iodine atom is likely to have a net positive or negative charge, which would affect the calculation of the RNCS. However, the authors feel that the advantage of greater computational simplicity of the descriptors more than compensates for some of the approximations embodied therein.

As explained above, the models exhibit some systematic underestimation or overestimation for some types of compounds. These systematic deviations can, in some cases, be rationalized qualitatively by consideration of polar interactions not accounted for explicitly in the descriptors, especially in model 1, for which the hydrogen bond acidities[23] α_2^H and basicities[24] β_2^H are absent. For instance, the HLC of 1,4-dioxane was overestimated by both versions of model 1: both of the oxygen atoms are basic and are therefore more likely to engage in hydrogen bonding but are slightly less basic than the case where the other oxygen atom is absent. Even so, the enhanced hydrogen bonding due to the basic oxygen atoms increases the aqueous solubility, thereby leading to a decrease in the HLC: this could explain why model 1 overestimates the HLC, while there is no such shortcoming for model 2. A similar argument could be made for the two morpholines and 4-acetylpyridine.

One point of relevance to the descriptor selection procedure of regression by leaps and bounds is that it was the linear regression model with the highest r^2 for each subset size which was examined finally with a view to the selection of descriptors for a particular model: the outliers for particular classes of compounds were not considered during variable selection, as this can be a somewhat subjective consideration, requiring qualitative consideration as well as quantitative consideration. It is likely that models of lesser gross statistical performance will have a different distribution of outliers, some of which might be more preferable to the final subset of variables chosen. It would be a very time-consuming and tedious task to attempt to examine this for some of the possible $2^k - 1$ regressions (where k is 27 or 29) in the leaps-and-bounds regression algorithm, but perhaps this could be performed by the analyst for a number of candidate models, assuming that less importance is associated with some of the more established selection criteria, such as r^2, standard error or collinearity diagnostics (as outlined earlier in the discussion for the choice of descriptors for model 1).

Finally, it was noted that the weighted path count developed by the authors was not found to be statistically significant in this QSPR study. However, it remains a possible alternative to Kier−Hall[6] connectivity indices, as explained during the discussion on descriptor generation.

In conclusion, it is apparent that explicit inclusion of hydrogen bonding parameters in a QSPR for HLCs increases the predictive capacity significantly and that computational neural networks are a very useful tool in enhancing model performance. It is expected that in the future the use of artificial neural networks in QSPRs may well become a routine matter, perhaps superseding conventional methods such as regression analysis.

ACKNOWLEDGMENT

One of the authors (N.J.E.) wishes to thank Labscan Ltd. for a student research grant.

Supporting Information Available: Tables of descriptors for all 357 compounds; experimental HLCs, the associated literature reference, and the HLCs predited by the linear and neural version of models 1 and 2; and the STO-3G partial atomic charges for heteroatoms in the functional groups of a typical set of 70 compounds representative of the 303-compound development set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Sander, R. Experimental HLC Database (Version 3); Air Chemistry Department, Max Planck Institute: P.O. Box 3060, 55020 Mainz, Germany, 1999.
(2) Meylan, W. M. Experimental HLC Database, Syracuse Research Corporation: 1999.
(3) Hine, J.; Mookerjee P. K. Intrinsic Hydrophilic Character of Organic Compounds − Correlations in Terms of Structural Contributions. *J. Org. Chem.* **1975**, *40*(3), 292−298.
(4) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solution Chem.* **1981**, *10*(8), 563−595.
(5) Nirmalakhandan, N. N.; Speece, R. E. QSAR Model for Predicting Henry's Constant. *Environ. Sci. Technol.* **1988**, *22*, 1349−1357.
(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies: Hertfordshire, Great Britain, 1986.
(7) Meylan, W. M.; Howard, P. H. Bond Contribution Method for Estimating Henry's Law Constants. *Environ. Tox. Chem.* **1991**, *10*, 1283−1293.
(8) Russell, C. J.; Dixon, S. L.; Jurs, P. C. Computer-Assisted Study of the Relationship between Molecular-Structure and Henry's Law Constant. *Anal. Chem.* **1992**, *64*, 1350−1355.
(9) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(10) Suzuki, T.; Ohtaguchi, K.; Koide, K. Application of Principal Components-Analysis to Calculate Henry's Constants from Molecular-Structure. *Computers Chem.* **1992**, *16*(1), 41−52.
(11) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapours in Water at 298K, and a New Method for its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777−1791.
(12) Nirmalakhandan, N. N.; Brennan, R. A.; Speece, R. E. Predicting Henry's Law Constant and the Effect of Temperature on Henry's Law Constant. *Water Res.* **1997**, *31*(6), 1471−1481.
(13) Hawker, D. W. Vapour-Pressures and Henry's Law Constants of Polychlorinated-Biphenyls. *Environ. Sci. Technol.* **1989**, *23*(10), 1250−1253.
(14) Sabljic, A.; Gusten, H. Predicting Henry's Law Constants for Polychlorinated-Biphenyls. *Chemosphere* **1989**, *19*(10/11), 1503−1511.
(15) Brunner, S.; Hornung, E.; Santl, H.; Wolff, E.; Piringer O. G.; Altschuh, J.; Bruggemann, R. Henry's Law Constants for Polychlorinated-Biphenyls − Experimental Determination and Structure Property Relationships. *Environ. Sci. Technol.* **1990**, *24*(11), 1751−1754.
(16) Dunnivant, F. M.; Elzerman A. W.; Jurs P. C.; Hasan, M. N. Quantitative Structure Property Relationships for Aqueous Solubilities and Henry's Law Constants of Polychlorinated-Biphenyls. *Environ. Sci. Technol.* **1992**, *26*(8), 1567−1573.
(17) Chiu, R. M. H.; Chen, B. D. Correlation of the Henry's Law Constant for Non-Polar Liquids in Molten Polyisobutylene. *Ind. Eng. Chem. Res.* **1996**, *35*, 4386-4388.
(18) Quale, O. R. The Parachors of Organic Compounds. *Chem. Rev.* **1953**, *53*, 439−589.
(19) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987; Appendix A.
(20) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid-Chromatography. *Chromatographia* **1987**, *23*(4), 243−246.

(21) Miller, K. J., Savchik, J. A. New Empirical-Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*(24), 7206−7213.
(22) Baudet, J.; Guy, J.; Tillieu, J., Calcul des Susceptibilités Magnétiques Principales de la Molécule de Benzène. *J. Phys. Radium* **1960**, *21*, 600−608.
(23) Abraham, M. H.; Greiller, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. Hydrogen Bonding. 7. A Scale of Solute Hydrogen-Bond Acidity Based on LogK-values for Complexation in Tetrachloromethane. *J. Chem. Soc., Perkins Trans. 2* **1989**, 699−711.
(24) Abraham, M. H.; Greiller, P. L.; Prior, D. V.; Duce, P. P.; Morris, J. J.; Taylor, P. J. Hydrogen-Bonding. 10. A Scale of Solute Hydrogen-Bond Basicity Using LogK-values for Complexation in Tetrachloromethane. *J. Chem. Soc., Perkins Trans. 2* **1990**, 521−529.
(25) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(26) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(27) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990**, *62*(21), 2323−2329.
(28) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(29) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*(3), 441−451.
(30) Mackay, D.; Shiu, W. Y. A Critical Review of Henry's Law Constants for Chemicals of Environmental Interest. *J. Phys. Chem. Ref. Data* **1981**, *10*(4), 1175−1199.
(31) Abraham, M. H.; Fuchs, R.; Whiting, G. S.; Chambers, E. C. Thermodynamics of Solute Transfer from Water to Hexadecane. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291−300.
(32) Abraham, M. H. *J. Chem. Soc., Faraday Trans. 1* **1984** , *80*, 153−181.
(33) Abraham, M. H.; Matteoli, E. Thermodynamics of Solution of Homologous Series of Solutes in Water. *J. Chem. Soc., Faraday Trans. 1* **1988**, *84*, 1985−2000.
(34) Wilhelm, E.; Battino, R.; Wilcock, R. J. Low-Pressure Solubility of Gases in Liquid Water. *Chem. Rev.* **1977**, *77*, 219−262.
(35) Bagno, A.; Lucchini, V.; Scorrano, G. Thermodynamics of Protonation of Ketones and Esters and Energies of Hydration of their Conjugate Acids. *J. Phys. Chem.* **1991**, *95*, 345−352.
(36) Gibbs, P. R.; Radzicka, A.; Wolfenden, R. The Anomalous Hydrophilic Character of Proline. *J. Am. Chem. Soc.* **1991**, *113*, 4714−4715.
(37) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling W. T. *Numerical Recipes*, 2nd ed.; Cambridge University Press: 1992.
(38) Myers, R. H. *Classical and Modern Regression Analysis with Applications*, 2nd ed.; Duxbury, 1990.
(39) SPSS Inc., Recognition Systems Inc. *Neural Connection 2.0 User's/ Applications Guide*; SPSS Software Inc.: Chicago, IL, U.S.A., 1997.
(40) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity-Relationships. *J. Med. Chem.* **1979**, *22*(10), 1238−1244.
(41) Furnival, G. M.; Wilson, R. W., Jr. Regression by Leaps and Bounds. *Technometrics* **1974**, *16*(4), 499−511.
(42) Visual Numerics Inc. *IMSL (R) Fortran 90 MP Library, Version 3.0;* Houston, TX, U.S.A., 1998.
(43) Belsley, D. A.; Kuh, E.; Welsh, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley: New York, 1980.
(44) Brennan, R. A.; Nirmalakhandan, N. N.; Speece, R. E. Comparison of Predictive Methods for Henry's Law Coefficients of Organic Compounds. *Water Res.* **1998**, *32*(6), 1901−1911.

CI010361D