Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets

Pierre Bruneau[†]

AstraZeneca Centre de Recherche, Parc Industriel Pompelle, BP 1050, 51689 Reims, France

Received February 10, 2001

Several predictive models of aqueous solubility have been published. They have good performances on the data sets which have been used for training the models, but usually these data sets do not contain many structures similar to the structures of interest to the drug research and their applicability in drug hunting is questionable. A very diverse data set has been gathered with compounds issued from literature reports and proprietary compounds. These compounds have been grouped in a so-called literature data set I, a proprietary data set II, and a mixed data set III formed by I and II. About 100 descriptors emphasizing surface properties were calculated for every compound. Bayesian learning of neural nets which cumulates the advantages of neural nets without having their weaknesses was used to select the most parsimonious models and train them, from I, II, and III. The models were established by either selecting the most efficient descriptors one by one using a modified Gram-Schmidt procedure (GS) or by simplifying a most complete model using automatic relevance procedure (ARD). The predictive ability of the models was accessed using validation data sets as much unrelated to the training sets as possible, using two new parameters: $NDD_{x,ref}$ the normalized smallest descriptor distance of a compound x to a reference data set and $CD_{x,mod}$ the combination of $NDD_{x,ref}$ with the dispersion of the Bayesian neural nets calculations. The results show that it is possible to obtain a generic predictive model from database I but that the diversity of database II is too restricted to give a model with good generalization ability and that the ARD method applied to the mixed database III gives the best predictive model.

INTRODUCTION

A key parameter which ultimately governs the in-vivo activity of a drug is the free concentration in the immediate vicinity of its biological target. This depends on physicochemical properties of the molecule like solubility, dissolution rate, partition coefficient, pK_a, membrane permeability, and affinity to plasma proteins to name just a few. These properties are more "generic" and are more likely to be predictable from the structure than the binding affinity for the target protein which controls the intrinsic activity. These physicochemical properties and in particular solubility have been the subject of many studies aimed at finding a way to predict them from the structure alone.

Many different attempts, which were reviewed up to the early 1990s by Yalkowsky and Banerjee[1] and recently by Morris and Bruneau,[2] have been made to predict aqueous solubility. One of the most successful methods is now known as Yalkowsky's equation[3] which in its simplified form relates the logarithm of the solubility in water expressed in mole.L^{-1} ($\log S_w$) to the logarithm of the partition coefficient between octanol and water ($\log P_{oct}$) and the melting point (mp) of the compound as shown in eq 1:

$$\log S_w = -1.00 \log P_{oct} - 0.012(\mathrm{mp} - 25) + 0.87 \quad (1)$$

Although Yalkowsky's equation has been established from a theoretical and structurally independent basis, its usage as such on a structurally diverse data set shows its limitations,[4] and it is best used after optimization of the coefficients A, B, and C of eq 2 on different compounds series.

$$\log S_w = A(\mathrm{mp} - 25) + B \log P_{oct} + C \quad (2)$$

With this latter approach the Yalkowsky model is very useful for estimating the solubility of members of series where $\log S_w$, $\log P_{oct}$, and mp have been measured on a few examples.

Since solubility in water has a strong relationship with the partition coefficient between water and a lipophilic phase, the same methods can be applied to predict them both. One of these methods is the fragment contribution approach which defines the solubility as an additive property of the atoms or the fragments of the molecule. Irmann[5] pioneered the approach by proposing eq 3

$$\log S_w = a + \sum n_i b_i + \sum n_j c_j + 0.0095(\mathrm{mp} - 25) \quad (3)$$

in which a is the contribution of the compound type, b_i is the contribution of the ith atom type which occur n_i times, and c_j is the contribution of the jth fragment which occurs n_j times. Equation 3 summarizes in a generalized formulation all the atomic or fragmental approaches which may be series dependent and may require a correction term for the solid state. Several successful studies on this approach have been published.[6−12]

But the fragmental approaches suffer from two main weaknesses: first there is no reason that a fragment has the same influence in different structural environments, and second substructures which are peculiar to an area of drug research are likely to remain undetermined because of their rarity in the public domain.

[†] Corresponding author phone: (+33)326616852; fax: (+33)326616842; e-mail: Pierre.Bruneau@astrazeneca.com.

10.1021/ci010363y CCC: $20.00
Published on Web 09/13/2001

Table 1. Comparative Characteristics of Some Published Databases Used for Training Solubility Models

database[a]	n[b]	MW (sd)[c]	Clog P_{oct} (sd)[d]	log S_w (sd)[e]	r^2 (nb descr)[f]	r^2 with Clog P_{oct}[g]
Bodor[13]	331	124 (59)	2.39 (1.29)	−2.05 (1.59)	0.96 (19)	0.89
Katritzky[15]	411	111 (42)	2.00 (1.3)	−1.60 (1.53)	0.88 (6)	0.86
Abraham[8]	648	147 (66)	2.64 (1.60)	−2.50 (2.00)	0.92 (5)	0.78
Jurs[14]	320	179 (87)	2.81 (2.47)	−3.30 (2.35)	0.97 (9)	0.73
Huuskonnen[16]	235	255 (78)	2.01 (1.76)	−3.06 (1.50)	0.90 (23)	0.56
AstraZeneca[h]	522	379 (96)	3.55 (1.49)	−4.75 (1.11)		0.42

[a] Database published in the corresponding paper. [b] Number of compounds in the database. These numbers may differ from the ones reported in the corresponding papers. The discrepancies are due to the failure of Drone (see methods) to calculate some structures. [c] Mean molecular weight (MW) and standard deviation (sd). [d] Mean Clog P_{oct} and standard deviation (sd). [e] Mean log S_w and standard deviation (sd). [f] Published squared coefficient of correlation of the calculated log S_w with the measured log S_w and number of descriptors involved in the model (nb descr.). [g] Squared coefficient of correlation of Clog P_{oct}[17] with measured log S_w. [h] This work.

Thus, the approach which involves calculation of a wide variety of 2D and 3D descriptors and which tries to find the most relevant model from them seems to be more attractive. The most comprehensive studies in this approach are the ones published by Bodor and Huang,[13] Mitchell and Jurs,[14] Katritzky et al.,[15] and Huuskonen et al.[16] Some data obtained from the databases published in these papers along with that of Abraham and Le[8] and in-house data are summarized in Table 1 for comparison purposes.

In Table 1, the AstraZeneca database is taken as an example of a set of compounds which are of interest to pharmaceutical research. Obviously the compounds which make up the data sets of Bodor, Katritzky, and Abraham are smaller, less lipophilic, and more soluble than that of AstraZeneca. It is striking that the simple linear correlation with Clog P_{oct}[17] ($r^2 = 0.86$) performs almost as well as the six descriptor model reported by Katritzky et al.[15] ($r^2 = 0.88$). This is probably due to the fact that for small molecules, mostly liquids, the Yalkowsky's model works quite well, and also that for small molecules Clog P_{oct} is an accurate prediction of log P_{oct}. Although still populated with smaller molecules, the database of Jurs overlaps the "drug-like" database in terms of lipophilicity and solubility. The database of Huuskonen is the one which incorporates the largest number of drugs, and this is reflected by a higher mean molecular weight and a poorer correlation with Clog P_{oct}.

The conclusion of this comparison is that the successful models published in the literature should be applied with great care in drug hunting programs, since they have been trained using structures which are quite different from those of interest to drug research.[18]

METHODS

Overall. The method used in this work to establish predictive models was first to calculate a relevant set of descriptors and second to use a suitable mathematical tool to select the best model using a dataset of measured properties. It was then straightforward to train the model and validate it with new measured data.

Descriptors. The calculated descriptors can be classified into three classes, namely topological (2D dependent), geometrical (3D dependent), and electronic descriptors (charge dependent). Emphasis has been put on surface dependent descriptors since it is intuitively felt that molecules interact with their environment via their surface properties. The topological class includes molecular weight, calculated log P_{oct} (CLOGP), calculated molecular refractivity (CMR), various counts of atoms and rings, and a series of descriptors which are the counts of the number of hydrogen bond donor or acceptor substructures present in the structure. Flags to indicate that the molecule is likely to be positively or negatively ionized at physiological pH are also determined.

The 3D coordinates are used to compute the geometrical descriptors, like the van der Waals and solvent accessible surface area of atoms which have been identified as hydrogen bond donor or acceptor or polar.[19]

In addition to the 3D coordinates, the atomic charges are used to compute a series of charge dependent descriptors, for example statistics on the distribution of charges and areas of positive or negative electrostatic potential on the van der Waals or solvent accessible surfaces.[20,21] They are also used to calculate a series of descriptors reported by Katritzky et al.[15]

Overall, more than 100 descriptors which are described in Table 2a−c can be calculated

Mathematical Tool. There are many mathematical tools available to find the best correlation between descriptors and the variable to be explained. They all have advantages and disadvantages. Numerous methods based on linear regressions, like multiple linear regression (MLR), principal component regression (PCR), and partial least-squares (PLS) methods, are fast, and they deliver a more or less interpretable model.[28] On the contrary, methods based on neural networks are often slow to train and do not allow easy model interpretation of what should be changed in a structure to improve the predicted property. Nevertheless neural nets have the advantage of being able to model unforeseen nonlinear relationships, including cross terms, between the dependent and the independent variables.[29] However classical neural nets show other difficulties which make their use computationally demanding. For example, the results depend on the initial random distribution of the parameters (weights) which define the neural nets. This can be solved by repeating the training of a neural network starting every time from a different distribution of weights. In addition the neural networks are also prone to overfitting and overtraining.[30,31] Overfitting occurs when too many weights are used compared to the number of compounds available in the training set. The overfitting risk can be reduced by using simple rules to choose the architecture of the network. For a three layered fully connected network a rule of thumb is to keep the number of weights under half the number of examples of the training set.[32,33] In most of the models generated in this work the ratio of the number of weights over the number of examples has been kept at a conservative ratio of 0.1 by adjusting the architecture of the neural nets.

Overtraining occurs when the number of optimizing cycles is too high, in this case the neural net is able to fit the data themselves including the noise. Obviously when this situation occurs the predictability of the model obtained is degraded. To detect the moment when the overtraining effect appears one must reserve some data which are not used in the training set (test set) to evaluate the predictability of the model as it is trained in order to select the model at its optimum predictability.

Another way to solve the problems of overfitting and of overtraining is to use Bayesian neural nets (BNN).[34] Very few publications about the use of BNN in the QSAR field have been made so far.[35−38] BNN differs from classical neural nets in that every weight is replaced by a distribution of weights. This method leads to the exploration of a large number of combination of weights (therefore of networks), and it is less likely to end in a local minimum. Furthermore there is no single set of weights (one network) which best fits the training dataset, but a set of different solutions that equally well fit the training dataset but do not have the same predictive ability. BNN allows evaluation of the likely uncertainty of a prediction.[39] These advantages of BNN plus a few others like the ability to give less influence to the outliers, or to automatically simplify the model as shown in this work and in a recent publication,[38] may well establish it as the most robust method for training QSAR models.

Model Selection. Having calculated as many relevant descriptors as required and having chosen the most convenient mathematical tool to establish the model we are faced with the problem of selecting the most parsimonious model, that is the one which uses the least number of parameters according to Ockham's razor principle.[40] According to this theory one must favor the model involving the smallest number of descriptors and consequently the most relevant ones. This induces the use of a simpler architecture for the final neural network, and it is in fact another facet of the overfitting problem which has been already discussed.

The problem of selecting the smallest set of relevant descriptors is recurrent in QSAR/QSPR fields, and many methods have been used with some success. Without doing an exhaustive review we may cite generalized simulated annealing (GSA),[41] multiregression (MR),[42] fast random elimination of descriptors (FRED),[43] genetic algorithm and neural network (GNN),[44] principal component analysis (PCA),[45] PCA and neural network,[46] genetic function approximation (GFA),[47] and GOLPE.[48]

We have initially chosen to use a method adapted from the Gram-Schmidt (GS) method.[49,50] The initial step of the GS procedure is to find the most correlated descriptor. The initial data matrix is then replaced with the residuals obtained after having computed the correlation between the remaining descriptors or the dependent variable with the selected descriptor. The process is repeated until no significant correlation can be extracted from the residuals. In this work all correlations are computed using BNN.

The GS method selects the descriptors one by one, and although the final model will, via the final training step, incorporate cross terms of the individually selected descriptors, the overall procedure will not consider the descriptors which are important only through their cross terms. For this reason we have explored the use of a facility built in Bayesian flexible models (BFM)[51] called automatic relevance determination (ARD). ARD, which was developed by D. McKay and R. Neal,[39] automatically determines which of many inputs to a neural network are relevant to the prediction of targets. It is done by adding to each input unit a hyperparameter which controls the magnitudes of the weights of the connections of that input unit. As the training proceeds the weights associated to irrelevant inputs are forced to small values, while the weights associated to important inputs are allowed to take high values. Starting with the most complete model (which includes all the available descriptors) the network is trained, then the distribution of the weights associated with every descriptor is analyzed, and only the descriptors with non small values are retained to the next step. The process is looped until it is impossible to remove any more descriptors without degrading the performance of the resulting model.

DATA

The databases which have been used to develop the methodology were either from the literature (I), were proprietary (II), or were a combination of both (III). The literature database consists of 1038 compounds taken from the papers of Abraham and Le,[8] Mitchell and Jurs,[14] Bodor and Huang,[13] Katritzky et al.,[15] and Huuskonen et al.[16] The proprietary database is the solubility of 522 research compounds non-ionized at pH 7.4 measured at a single site of AstraZeneca (Alderley Park) at pH 7.4 after 1 or 3 days in equilibrium.

The validation databases were obtained from a recent publication of Huuskonen.[52] It included 1308 compounds, 635 of which were already included in I leaving 673 new compounds to form database Ival. Proprietary validation database IIval was formed with new data obtained since the beginning of this work.

Table 3 summarizes the characteristics of the different databases. As already seen in the comparison of the different published databases (Table 1), I and Ival have a smaller mean molecular weight, smaller mean Clog P_{oct}, and higher mean aqueous solubility than II and IIval.

EXPERIMENTAL SECTION

Measurements. Compound solubility was determined by incubating an excess amount of solute in phosphate buffer (0.01 M, pH 7.4) at 25 °C for 3 days under stirring. After equilibration the stirrer was stopped, and solution was allowed to settle for 2 h. The supernatant liquid was collected and centrifuged at high speed (ca. 4000 rpm) under thermostated conditions (25 °C). After 15 min of centrifugation the supernatant liquid was transferred to a clean vial and centrifugated a further 15 min. The solute was then quantified by gradient HPLC and using a single external standard.

Recently a higher throughput assay has been adopted. This assay was identical to the one just described except that the buffer is 0.1 M and the equilibration phase lasts only 1 day. Numerous validation tests indicate that there are no significant difference between the methods, and the results were pooled in the same database.

Databases. In-house databases comprising solubility data and structures in SMILES[53,54] codes were collected from the central AstraZeneca database.

Table 2. 2D, 3D, and Charge Dependent Descriptors

	a. 2D Dependent Descriptors
MW	molecular weight
CLOGP	daylight Clog P_{oct}[17]
CMR	calculated molecular refractivity[17]
FLEX_BND	number of rotatable bonds
MOLE_FLEX	molecular flexibility[23]
SPEC_FLEX_BND	FLEX_BND/HEAVIES
HB_DON(ACC)	number of potential H bond donor(acceptor) bonds
HBTOT	HB_DON + HB_ACC
AT_TOT	total number of atoms
HEAVIES	number of heavy atoms
POS(NEG)_CHARGES	number of potential positive (negative)charges
POS(NEG)_CHARGED	= 1 if POS(NEG)_CHARGES > 0
CHARGES	POS_CHARGES + NEG_CHARGES
CHARGED	= 1 if CHARGES > 0
NBN(O)(C)(P)(S)(X)	number of nitrogen (oxygen) (carbon) (phosphorus) (sulfur)(halogen) atoms
NEL	number of electrons
SIC	structural information content of 0 order[15]
NB_RINGS	number of rings
Pat	number of polar atoms (O, N, S, P)
NPat	number of nonpolar atoms (NBC + NBX − Pat)
MWPat	MW * Pat/AT_TOT
MWNPat	MW * NPat/AT_TOT
MWSHDA	MW * HB_TOT/AT_TOT
QUATER	number of quaternary nitrogen
PIAT	number of pi atoms (number of double bonds + number of halogen atoms)
HAROM	number of hydrogens linked to an aromatic atom
AROM	number of aromatic atoms
	b. 3D Dependent Descriptors
M1(2)(3)M	moment of inertia along the first (second) (third) principal axe of the molecule
GAUS_VOL	Gaussian volume
OVAL	area/area of a sphere with the same volume as the molecule[24]
VDW_POL(NONPOL)_AREA	van der Waals polar(non polar) area
SAS_POL(NONPOL)_AREA	solvent accessible surface polar(non polar) area
VDW_HB_A(D)_AREA	van der Waals H bond acceptor(donor) area
SAS_HB_A(D)_AREA	solvent accessible surface H bond acceptor(donor) area
VDW_TOT_AREA	van der Waals (VDW) total area
SAS_TOT_AREA	solvent accessible surface (SAS) total area
SPEC_(vdw area)	any vdw area/VDW_TOT_AREA
SPEC_(sas area)	any sas area/SAS_TOT_AREA
SPEC_HB_TOT	HB_TOT/VDW_TOT_AREA
	c. 3D and Charge Dependent Descriptors[a]
MM_AsumQ	sum of absolute atomic charges[25]
MM_Qneg(pos)Mean	mean of negative (positive) charges[26]
MM_Qneg(pos)Var	variance of negative (positive) charges[26]
MM_MAXPOS(NEG)	maximum positive (negative) atomic charge
MM_QON	sum of atomic charges on O and N[13]
MM_QO(N)(C)(H)	sum of atomic charges on O (N) (C) (H)[13]
MM_ZAP_SOLVNRG	ZAP solvation energy[27]
MM_ZAP_SAREA	ZAP surface area[27]
MM_ZAP_PCR1(2)(3)	ZAP % region 1 (2) (3) ranges = −0.10, +0 × 10[27]
MM_ZAP_SAR1(2)(3)	ZAP surface area of region 1 (2) (3) ranges = −0.10, +0 × 10[27]
MM_VDW_EP_P(N)_AREA	area of van der Waals surface with positive(negative) electrostatic potential
MM_SAS_EP_P(N)_AREA	area of solvent accessible surface with positive(negative) electrostatic potential
MM_VDW_EP_P(N)_SUM	sum of positive(negative) electrostatic potentials on van der Waals surface
MM_VDW_EP_P(N)_MEAN	mean of positive(negative) electrostatic potentials on van der Waals surface
MM_VDW_EP_P(N)_VAR	standard deviation of positive(negative) electrostatic potentials on VDW
MM_SAS_EP_P(N)_SUM	sum of positive(negative) electrostatic potentials on solvent accessible surface
MM_SAS_EP_P(N)_MEAN	mean of positive(negative) electrostatic potentials on solvent accessible surface
MM_SAS_EP_P(N)_VAR	standard deviation of positive(negative) electrostatic potentials on SAS
MM_SPEC_SAS_EP_P(N)_AREA	MM_SAS_EP_P(N)_AREA/SAS_TOT_AREA
MM_SPEC_VDW_EP_P(N)_AREA	MM_VDW_EP_P(N)_AREA/VDW_TOT_AREA
MM_HDSA	sum of [(charge on H bond donor atom)*(sqrt area of the atom)]/(total area)[15]
MM_HDCA	sum of [(charge on H bond donor atom)*(sqrt area of the atom)]/(sqrt total area)[22]
MM_HASA	sum of [(charge on H bond acceptor atom)*(sqrt area of the atom)]/(total area)[15]
MM_HACA	sum of [(charge on H bond acceptor atom)*(sqrt area of the atom)]/(sqrt total area)[22]
MM_HADSA	sum of [(charge on H bond acceptor or donor atom)*(sqrt area of the atom)]/(total area)[15]
MM_HADCA	sum of [(charge on H bond acceptor or donor atom)*(sqrt area of the atom)]/(sqrt total area)[22]
MM_FHDSA	MM_HDSA/(sqrt total area)[15]
MM_FHADSA	MM_HADSA/(sqrt total area)[15]
MM_FHASA	MM_HASA/(sqrt total area)[15]
MM_RNCS	relative negative charge surface area[15]
MM_PCWT	most negative partial charge weighted topological index[22]

[a] A prefix (MM_) is added to all these descriptors names to indicate that the charges have been obtained in this work by a molecular mechanic method.

Table 3. Comparative Characteristics of the Databases Used for Training and Validating Solubility Models

database[a]	content	n^b	MW (sd)[c]	Clog P_{oct} (sd)[d]	log S_w (sd)[e]
I	literature	1038	173 (87)	2.41 (1.85)	−2.51 (2.11)
II	AstraZeneca	522	379 (96)	3.35 (1.49)	−4.75 (1.11)
III	**I** + **II**	1560	242 (133)	2.72 (1.79)	−3.26 (2.10)
Ival	literature	673	226 (93)	2.16 (2.26)	−2.79 (2.11)
IIval	AstraZeneca	261	384 (79)	2.97 (1.45)	−4.95 (1.08)
IIIval	**Ival** + **IIval**	934	270 (114)	2.39 (2.10)	−3.39 (2.18)

[a] Name of the database (see text). [b] Number of compounds in the database. [c] Mean molecular weight (MW) and standard deviation (sd). [d] Mean Clog P_{oct} and standard deviation (sd). [e] Mean log S_w and standard deviation (sd) in mole.L^{-1}.

Literature data were either scanned from the printed paper or obtained as Supporting Information or directly from the authors.[55] When the SMILES codes were not directly available, the reported names were used to screen an ASCII version of ACD database,[56] and the structures were extracted in the SD file format.[57] A Daylight[58] routine was then used to convert the SD files in uniquified SMILES codes.

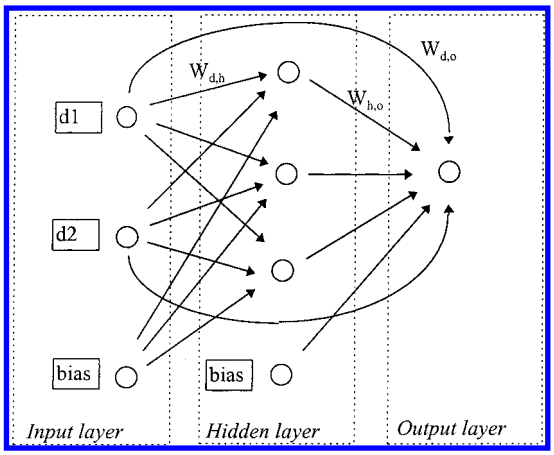
Structures and Descriptors. An in-house program called Drone has been written in Perl language to automatically generate the 3D structures and calculate the descriptors. This program described in Chart 1 simultaneously calculates the 2D dependent descriptors and generates a 3D structure using CONCORD[59] followed by a minimization in the local minimum within Sybyl[60] using the built-in MMFF94 force field.[60]

Meanwhile Drone calculates various atom and substructural features with an in-house routine called StructAnal. In StructAnal, substructures able to give or receive hydrogen bonds or likely to be ionized at physiological pH are predefined using SMARTS codes.[61] The occurrence of these properties are then counted for each molecule.

The descriptors prefixed by MM_ZAP in Table 2c refer to Poisson−Boltzmann calculations, computed using a smooth permittivity function.[27] The descriptor MM_ZAP_SOLVNRG is the molecular solvation energy, whereas the quantities MM_ZAP_PCRn and MM_ZAP_SARn (n = 1, 2, 3) arise from an analysis of the electrostatic potential at the solvent-accessible surface of the molecule. In this analysis the solvent-accessible surface is divided into three sections depending upon the surface values of the total potential (including the polarization contribution from arising from solvent, modeled as linear dielectric). The ranges are more negative than −0.1 (n = 1), between −0.1, +0.1 (n = 2), and more positive than 0.1 (n = 3) where the units are kT/e. The percent fraction of the surface area in each of these ranges is the parameter MM_ZAP_PCRn, where n = 1, 2, 3. These fractional areas are converted to absolute areas in the parameters MM_ZAP_SARn, using the descriptor MM_ZAP_SAREA, which is the total solvent accessible surface area.

The different surface area descriptors shown in Table 2b are computed by evenly distributing an equal number of points on the spheres formed by each atom. A small area depending on the radius of the van der Waals or solvent accessible sphere of each atom is associated with each point. After having removed the points situated inside the sphere of a neighbor atom, the total or partial surface areas are calculated by summing the area associated with all remaining points.

Scheme 1. Architecture of the Neural Nets



The MMFF94 force field is also used to define the atomic point charges. An electrostatic potential created by the atomic point charges is associated with every surface point already defined. Theses potentials and the atomic charges are used to calculate the charge dependent descriptors shown in Table 2c.

Statistics. Statistics on the data and on the results were calculated using JMP.[62] The retained criteria for the goodness of fit is the root-mean-square error (rmse) calculated according to eq 4, where n is either the number of measured examples in the case of the training of the models or the number of calculated examples in the case of the validation of the models.

$$\mathrm{rmse} = \sqrt{\frac{\sum_{i=1}^{n}(\log Sw_{i,\mathrm{measured}} - \log Sw_{i,\mathrm{calculated}})^2}{n}} \quad (4)$$

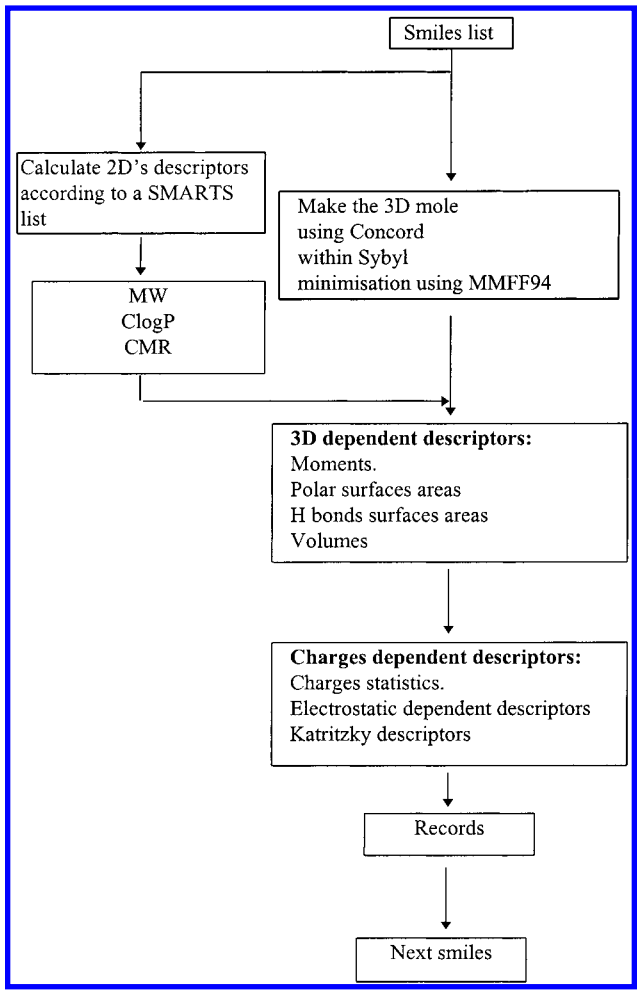
In the case of model training, r^2 was also calculated according to eq 5[62] where $\sigma_{\mathrm{measured}}$ is the standard deviation of the measured data.

$$r^2 = 1 - \frac{n}{n-1}\frac{\mathrm{rmse}^2}{\sigma_{\mathrm{measured}}{}^2} \quad (5)$$

Bayesian Neural Nets. The Bayesian training of the neural nets was performed using programs in Perl language incorporating routines included in Flexible Bayesian Models suite of programs written by R. Neal.[51] To use a data matrix comprising n examples, d descriptors, and one target value per example, the initial step was to build a neural net with d nodes in the input layer, h nodes in the hidden layer, and one node in the output layer. All the nodes of the input layer were connected to the output node and to all the nodes of the hidden layer which are in turn all connected to the output node. In addition a bias node linked to all hidden nodes is added to the input layer and a bias node linked to the output node is added to the hidden layer.

Scheme 1 shows an example of a neural net with 2 input nodes for two descriptors, 3 hidden nodes, and 1 output node. The number of nodes in the hidden layer was calculated to control ρ, the ratio of n, the number of examples to w, the

Chart 1. Drone



total number of connections in the network. w, ρ, and h were calculated according to eqs 6, 7, and 8, respectively.

$$w = (h + 1)(d + 1) + h \tag{6}$$

$$\rho = \frac{n}{(h + 1)(d + 1) + h} \tag{7}$$

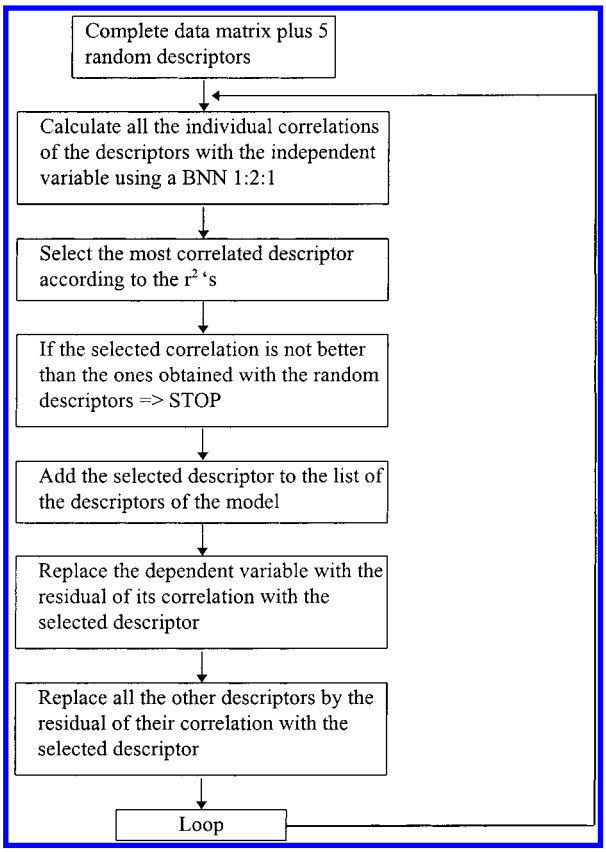
$$h = \frac{\frac{n}{\rho} - d - 1}{d + 2} \tag{8}$$

Model Selection and Training. The descriptor vectors and the dependent variable were scaled to give a mean equal to 0 and a standard deviation equal to 1 prior to being fed into the neural net.

The networks were then trained for a few thousand cycles, and the 200 networks issued from the last 200 cycles were taken as a sample of the final distribution of the weights in the resulting BNN. The results of the BNN ($\text{result}_{\text{BNN}}$) were calculated as the mean of the values given by this sampling associated with their standard deviation (sd_{pred}). The obtained mean was unscaled according to eq 9 where $\log Sw_{\text{calculated}}$ is the final result, $\overline{\log Sw_{\text{measured}}}$ is the mean of the data used to train the model, and σ_{measured} is their standard deviation.

$$\log Sw_{\text{calculated}} = \text{result}_{\text{BNN}} \times \sigma_{\text{measured}} + \overline{\log Sw_{\text{measured}}} \tag{9}$$

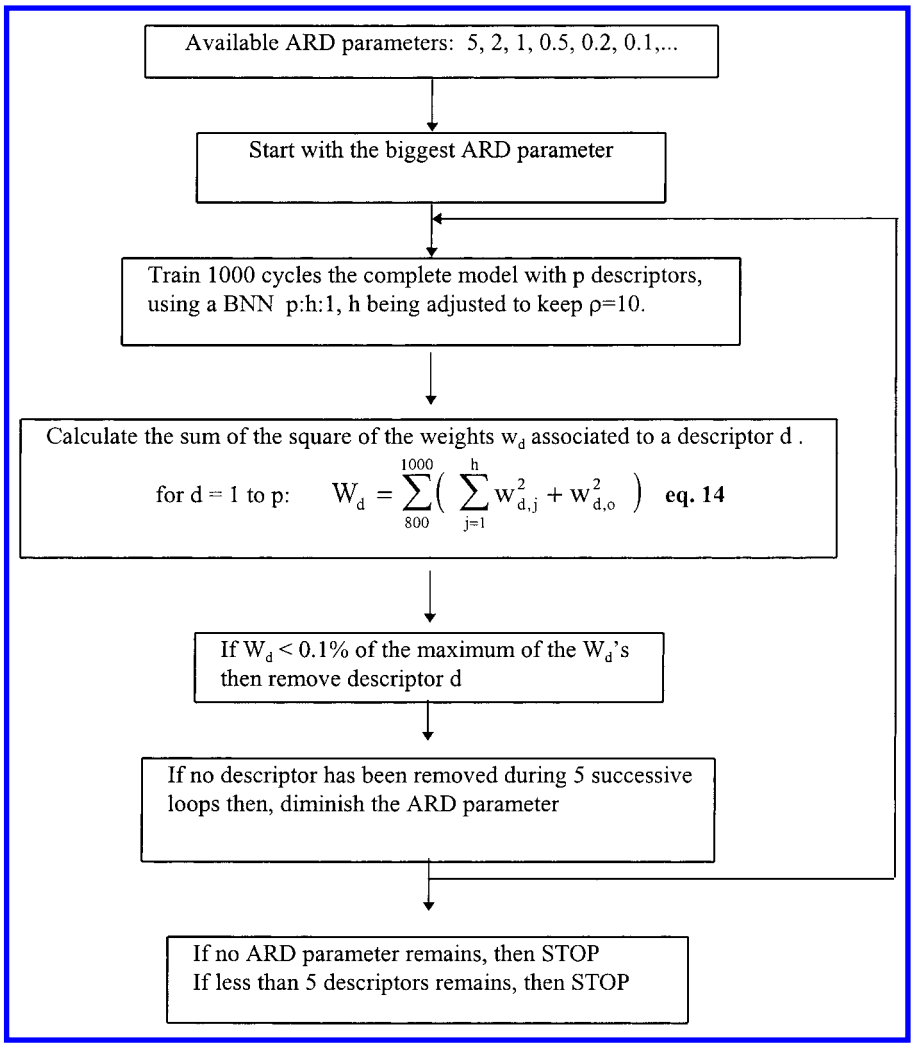
Chart 2. GS Method



The process of the GS method is described in Chart 2. To the initial data matrix is added five pseudodescriptors filled up with random numbers. The process starts by establishing all the correlations of the individual descriptors with the dependent variable, using a BNN with 1 input node, 2 hidden nodes, and 1 output node, trained for 2000 cycles. The r^2's are calculated, and the descriptor with the highest r^2 is retained. The second step is to replace the initial data matrix minus the retained descriptor, with the residuals of the correlation of every descriptor with the selected descriptor (this step is called Gram-Schmidt orthogonalization[50]). The dependent variable is also replaced with its residuals of the correlation with the retained descriptor. These correlations are performed using a BNN with 1 input node, 2 hidden nodes, and 1 output node, trained for 2000 cycles. The process is repeated until the correlations of the individual "real" descriptors are not different from the correlations obtained with the "pseudo" random descriptors. This means that no further information can be extracted from the initial data matrix.

Contrary to the GS method, the ARD method starts with training the most complete model using the full initial data matrix as described in Chart 3. The ARD as implemented in BFM is controlled by a parameter which hardens the process as it becomes smaller. The BNN was trained for 1000 cycles with as many input nodes as the number of descriptors, a number of hidden nodes calculated to keep ρ whenever is possible equal to 10 and 1 output node. When the number of descriptors and the number of examples do not allow to keep ρ equal to 10 without having less than two hidden nodes (see eq 8), the number of hidden nodes is locked at 2. After training, the weights related to each descriptor are examined i.e., for each descriptor the sum of the last 200 networks of

Chart 3. ARD Method



the square of the weights associated to the links between the descriptor and the hidden and output nodes is calculated according to eq 14 shown in Chart 3, where $w_{d,j}$ and $w_{d,o}$ are defined in Scheme 1. All the descriptors with an associated sum of weights less than 0.1% of the maximum sum of weights are discarded from the initial data matrix. If no descriptor is removed for five successive loops the ARD parameter is decreased to the next one. The process is repeated until the performance of the obtained simplified model, controlled by its r^2, is significantly degraded compared to the r^2 of the most complete model.

The final models were obtained by training for 2000 cycles a BNN with as many input nodes as the number of descriptors selected with the GS or ARD methods, a number of hidden nodes calculated to obtain $\rho = 10$ according to eq 8 and a single output node. As recommended in the BFM help manual,[51] a representative sampling of the Bayesian neural nets is given by selecting the last 200 cycles of the training phase instead of effectively sampling the distribution of the weights.

Descriptors Distance. A descriptors distance between a compound x and a compound j belonging to a reference set is defined by eq 10 where $Dist_{xj}$ is the calculated distance between compound x and the jth compound of the considered reference set, D_{dj} is the value of the *d*th descriptor of the jth compound of the reference set, D_{dx} is the value of the *d*th descriptor of the compound x, and σ_d is the standard deviation of the *d*th descriptor in the reference set.

$$Dist_{xj} = \sqrt{\sum_{d=1}^{p} \frac{(D_{dj} - D_{dx})^2}{\sigma_d^2}} \qquad (10)$$

For a compound x, there are as many distances as there are compounds in the reference set. This set of distances forms a vector called $D_{x,ref}$. For a compound its characteristic distance to a reference set, called descriptors distance ($DD_{x,ref}$) is defined as the minimum value of the elements of $D_{x,ref}$ as shown is eq 11.

$$DD_{x,ref} = \min(D_{x,ref}) \qquad (11)$$

To compare different models with different numbers of descriptors a normalized value $NDD_{x,ref}$ has been defined as in eq 12, where *d* is the number of descriptors in the reference set.

$$NDD_{x,ref} = \sqrt{\frac{15}{d}} * DD_{x,ref} \qquad (12)$$

Another indicator of the proximity of a compound x to be predicted to a training set used to make the model mod is

Table 4. Selection of Models: Descriptors Selected by the Gram-Schmidt and ARD Methods

databases[a]	GS [b]	ARD [b]
I	**CLOGP** MM_PCWT MM_AsumQ **CHARGES** HEAVIES **VDW_TOT_AREA (6)**	AROM **CHARGES CLOGP** FLEX_BD HB_DON M1M MM_FHADSA MM_HADCA MM QMIN MM_SPEC SAS_EP_N_AREA NBN NB_RINGS NEGCH **VDW_TOT_AREA (14)**
II	**CLOGP** PIAT VDW_HB_D_AREA MM ASumQ MM_FHASA HB_DON **(6)**	AT_TOT **CLOGP** MM_HADCA MM_HDSA MM_ZAP_SAREA NEL **(6)**
III	Npat **CLOGP** PIAT MM_PCWT CHARGES **NB_RINGS** MM_RNCS **(7)**	AROM AT_TOT **CLOGP** FLEX_BD HB_DON M1M MM_HDSA MM_MAXNEG MM_QposMean MM SOLVNRG MM_VDW_EP_P_SUM **NB_RINGS** NEL Pat POSCHARGED SPEC_VDW AREA **(16)**

[a] Databases used for the selection of the models. [b] Method used for the selection of the models (see text), the common descriptors between corresponding GS and ARD models are shown in bold, number of descriptors are shown between brackets.

Table 5. Selection of Models[g]

		GS[c]			ARD[c]		
databases[a]	sub-db[b]	nb hidden[d]	r^2 [e]	rmse[f]	nb hidden[d]	r^2 [e]	rmse[f]
I		12	0.94	0.50	6	0.95	0.45
II		6	0.64	0.67	6	0.64	0.67
III		17	0.93	0.56	8	0.94	0.53
	I		0.94	0.50		0.95	0.45
	II		0.65	0.66		0.65	0.66

[a] Databases used for the selection of the models. [b] Subdatabases in database **III**. [c] Method used for the selection of the models. [d] Number of hidden nodes in the neural network. [e] Squared correlation coefficient of the linear correlation between calculated and measured values of solubility. [f] Root-mean-squared error of the calculated values of solubility in log unit (mole.L^{-1}). [g] Final training using $\rho = 10$.

given by the combined distance $CD_{x,mod}$ defined by eq 13 where sd_{pred} is the standard deviation of the distribution of the predictions given by the BNN.

$$CD_{x,mod} = \sqrt{sd_{pred} * NDD_{x,ref}} \quad (13)$$

RESULTS AND DISCUSSION

Selection of Models. The three training databases were submitted to model selection using the GS method and the ARD method. Apart from the in-house database II, the ARD method selects more descriptors than the GS method (Table 4). This may indicate that combinations of descriptors are more important for optimizing the models than the individual descriptors alone. There are few descriptors in common between the two methods: 3 for database I (CLOGP, CHARGES and VDW_TOT_AREA), 1 for database II (CLOGP), and 2 for database III (CLOGP and NB_RINGS). This is probably due to the fact that the descriptors are highly intercorrelated; therefore, the variation of the dependent variable can be explained equally well by different combinations of descriptors. It is also remarkable that only CLOGP is a common descriptor for all six models, thus confirming the importance of log P_{oct} in the modelization of the solubility as reported by Yalkowsky et al.[3]

Training. When the best models were selected, they were trained to obtain the final models. Table 5 summarizes the results obtained for the three databases after the training phases. The statistics obtained on the two subdatabases I and II with the models trained on the complete database III are also shown. The ARD method gives at least equal or slightly better results on this training phase than the GS method. The performances obtained with the literature database I are good with rmse = 0.50 and 0.45 log units for GS and ARD, respectively. The results obtained on proprietary database II are less satisfactory (rmse = 0.67 in both methods). This may be explained by the fact that the samples used in the drug hunting phase are not always well crystallized nor pure thus leading to a larger experimental error for this category of compounds compared to the ones reported in the literature. Furthermore, as CLOGP appears to be an important descriptor in the models, the error in CLOGP might have a higher impact on the results obtained on the proprietary database II (and IIval) compared to the results obtained with simpler compounds of the literature database I (and Ival). This is due to the fact that the fragmental method used to calculate CLOGP is more likely able to give better estimations on simple compounds compared to the more unusual discovery compounds.

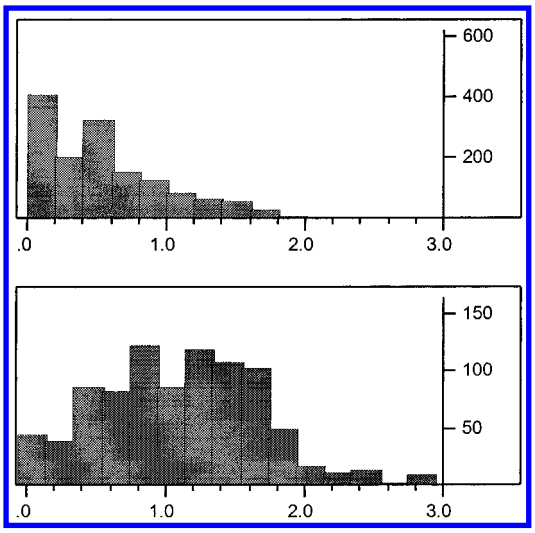


Figure 1. (a) Distribution of the descriptors distances of model IIIARD for a simulation of a leave-one-out procedure. (b) Distribution of the descriptors distances of IIIval in the case of model IIIARD.

Table 6. rmse Obtained on the Different Validation Databases by the Different Models

validation db[a]	I GS[b]	I ARD[b]	II GS[b]	II ARD[b]	III GS[b]	III ARD[b]
Ival	**1.08**	**0.84**	1.43	1.88	**0.91**	**0.82**
IIval	1.42	1.00	**0.81**	**0.78**	**0.93**	**0.79**
IIIval	1.17	0.88	1.29	1.65	**0.91**	**0.81**

[a] Databases used for the validation of the models. [b] Database and method used for the selection and training of the models. The reported figures are the rmse in log unit (mole.L^{-1}). Results shown in bold are the results obtained using same category of compounds for the training and the validation sets.

Table 7. Validations: Distributions of rmse According to $NDD_{x,ref}$

	IARD[a]			II ARD[a]			III ARD[a]		
bins[b]	Ival[c]	IIval[c]	IIIval[c]	Ival[c]	IIval[c]	IIIval[c]	Ival[c]	IIval[c]	IIIval[c]
<0.5	0.52 (82)		0.55 (83)	1.20 (15)	0.62 (49)	0.78 (64)	0.56 (114)	0.62 (22)	0.57 (136)
>0.5 < 1.0	0.60 (122)		0.60 (124)	1.08 (142)	0.83 (120)	0.97 (262)	0.64 (187)	0.68 (88)	0.65 (275)
>1.0 < 1.5	0.78 (138)	0.73 (10)	0.78 (148)	1.40 (193)	0.79 (61)	1.28 (254)	0.87 (187)	0.85 (95)	0.86 (282)
>1.5 < 2.0	0.90 (142)	0.88 (36)	0.90 (178)	1.61 (114)	0.93 (15)	1.54 (129)	0.97 (135)	0.81 (47)	0.93 (182)
>2.0	1.03 (189)	1.03 (212)	1.03 (401)	2.71 (209)	0.64 (16)	2.61 (225)	1.24 (50)		1.24 (59)

[a] Database and method used for the selection and training of the models. [b] Bins of $NDD_{x,ref}$. [c] Validation databases. The reported figures are the means of the rmse's in log units (mole.L^{-1}); the number of examples in each bin is reported in brackets. No rmse is reported when there are less than 10 examples in the corresponding bin.

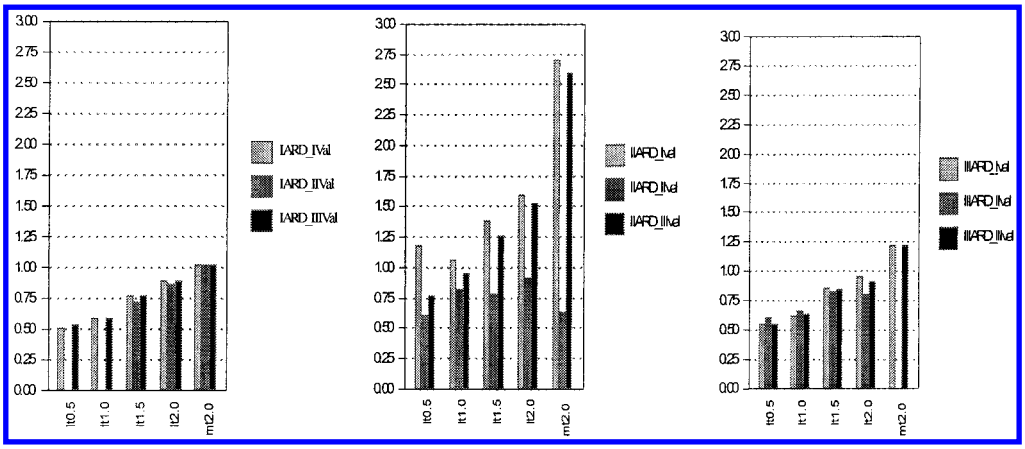


Figure 2. Rmse of each bin of $ndd_{x,ref}$ when (a) IARD, (b) IIARD, and (c) IIIARD are applied to Ival, IIval, and IIIval, respectively.

It is remarkable that when the complete database III was trained, the results obtained on the two subdatabases I and II are identical to the ones obtained when databases I and II are individually trained. This is seen as an effect of the stability and robustness of the Bayesian method of neural networks learning.

These results must be compared to others reported in the literature. Abraham and Le[8] made a review on the published results comparing the standard deviation (SD)[8,66] of the error of the different models calculated as log S_w in mole.L^{-1}. Although the comparison of models is difficult due to the fact that outliers are sometimes removed, Abraham and Le reported that the SDs span from 0.23 log units for a 123 compound data set with no complicated structures to 0.56 log units for a much more diverse 873 compound data set. Several authors have discussed the influence of the experimental error on the performance of a predictive model. Myrdal et al.[9] reported examples where the SD of repeated solubility measurements can be up to 0.38 log units. Katritzky et al.[15] calculated that the average SD for aqueous solubilities is 0.58 log units. A study of our in-house experimental database showed that the average SD of repeated measurements on different batches of the same compound is 0.49 log units.[63] To solve this problem, an expert system aimed at evaluating the quality of the reported solubility values has been proposed by Heller et al.[64] Abraham and Le [8] concluded that for a large diverse data set the SD of any predictive model cannot be lower than 0.50 log units.

Validation Problems. The main criteria of the quality of a model is given by its ability to predict the solubility of compound unrelated to those included in the training set. Often a dataset used to train, cross-validate, or test a predictive model includes numerous closely related series. In this situation any compounds included in the validation process have one or more closely related compounds in the corresponding training set. Even when a randomly selected subset is initially reserved to test the model, it is likely that the compounds in it have a close analogue in the training set. To demonstrate this concept we have looked at the distribution (Figure 1a) of descriptors distances in a leave-one-out (loo) simulation of the model IIIARD and at the distribution (Figure 1b) of the descriptors distances of compounds of the dataset IIIval when predicted with model IIIARD. By comparing the two distributions we deduce that there are many more chances to have a closely related compound in the training set(s) in the loo procedure than if the validation set is selected without any relationship with the training set. Therefore the validation results obtained in various cross-validation conditions indicate only the ability of the model to predict the property within the initial set. On another hand it is intuitively felt that no model trained on a finite set can predict a property of the entire chemical universe. We thus have tried to establish a method which would evaluate the confidence of the prediction of any structure given the prior knowledge of the diversity of the training set. $NDD_{x,ref}$ and $CD_{x,mod}$ were designed for this purpose.

Validation. As reported in Table 6 all models are best at predicting their own category of data, i.e. models trained on literature data (IGS and IARD) give better predictions on literature compounds (Ival) than on in-house compounds (IIval), thus confirming the inherent differences between literature and drug-like compounds. On this aspect, models trained with in-house data (IIGS and IIARD) are even poorer at predicting the literature data. This is due in part to the structural differences between the two data sets (see Table 3) but also mainly to the fact that the in-house dataset (II) contains only nonionizable compounds and thus is lacking

Table 8. Validations: Distributions of rmse According to $CD_{x,mod}$

	I ARD[a]			II ARD[a]			III ARD[a]		
bins[b]	Ival[c]	IIval[c]	IIIval[c]	Ival[c]	IIval[c]	IIIval[c]	Ival[c]	IIval[c]	IIIval[c]
<0.2	0.38 (47)	− (1)	0.37 (48)	− (1)	0.58 (15)	0.57 (16)	0.45 (75)	− (7)	0.45 (82)
> 0.2 < 0.4	0.64 (184)	− (7)	0.64 (191)	0.83 (52)	0.75 (131)	0.78 (183)	0.63 (181)	0.65 (94)	0.64 (275)
> 0.4 < 0.6	0.79 (207)	0.94 (39)	0.82 (246)	1.20 (220)	0.86 (85)	1.11 (305)	0.91 (238)	0.86 (129)	0.89 (367)
> 0.6 < 0.8	0.93 (137)	0.86 (60)	0.91 (197)	1.63 (182)	0.92 (17)	1.58 (199)	0.87 (125)	0.90 (22)	0.88 (147)
> 0.8	1.20 (98)	1.08 (154)	1.13 (252)	2.67 (218)	0.61 (13)	2.60 (231)	1.19 (54)	- (9)	1.16 (63)

[a] Database and method used for the selection and training of the models. [b] Bins of $CD_{x,mod}$. [c] Validation databases. The reported figures are the rmse in log units (mole.L^{-1}); the number of examples in each bin is reported in brackets. No rmse is reported when there are less than 10 examples in the corresponding bin.

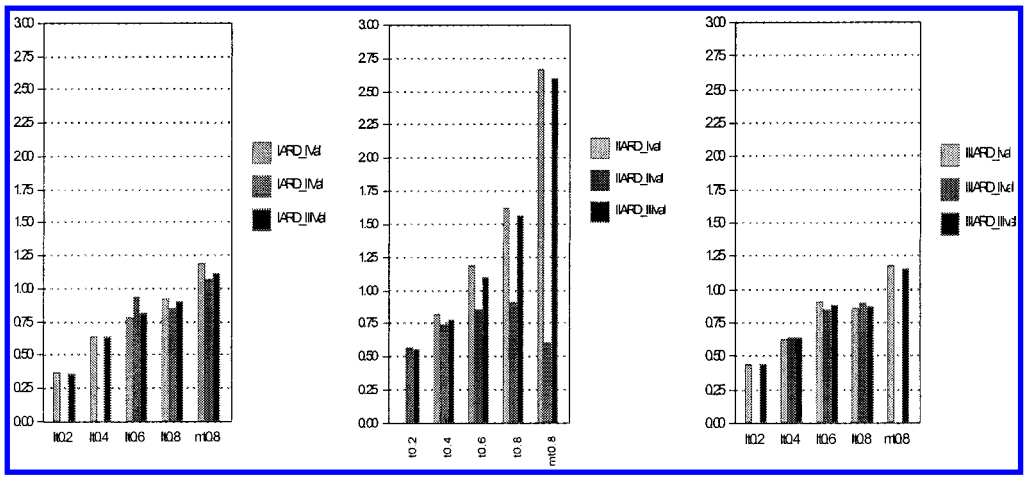


Figure 3. Rmse of each bin of $CD_{x,mod}$ when (a) IARD, (b) IIARD, and (c) IIIARD are applied to Ival, IIval, and IIIval, respectively.

important information to predict ionizable literature compounds. Using a mixed database (III) solves the problem since it gives models with similar performances on either literature or in-house validation databases indicating that the BNN has learned the different characteristics of the databases. Except for the models obtained with the in-house data (IIGS and IIARD), the ARD method has an advantage over the GS method in its ability to predict solubility, and the best model is clearly IIIARD. It is again remarkable that the results obtained by the model IIIARD on the subdatabases Ival and IIval (rmse = 0.82 and 0.79, respectively) are almost identical to the results obtained by the models I ARD applied to Ival and IIARD applied to IIval (rmse = 0.84 and 0.78, respectively).

The overall predicting performances of ARD models (rmse around 0.8) are however disappointing compared to their training results (rmse around 0.5 if IIARD is excepted). Since the validation datasets are a priori unrelated to the training sets and since it is Utopian to expect that a model can predict any structure, it is useful to evaluate how similar the various validation sets are to the training sets. One way to do this is to calculate the normalized smallest descriptors distances ($NDD_{x,ref}$) from the validation sets to the training sets. These distances were hashed into 5 bin and the rmse's of each bin are reported in Table 7 and shown in graphical format in Figure 2 (parts a−c).

When model IARD is applied to Ival, if $NDD_{x,ref} < 0.5$ the result is similar (rmse = 0.52) to the one obtained in the training phase (rmse = 0.45), then as the compounds get more dissimilar the rmse's increase smoothly up to 1.03 for $NDD_{x,ref} > 2.0$. When IARD is applied to the in-house dataset IIval, the results are surprisingly identical to the ones obtained when IARD is applied to Ival, providing there are enough compounds in the bins to calculate the rmse's. This indicates that (i) IARD is a model able to predict drug-like compounds with the same performance as it is able to predict compounds of its own family and (ii) the in-house solubility measurements are not intrinsically different from the data reported in the literature thus contradicting what has been observed in the training phase. This overall picture is the same for model IIIARD when applied to Ival, IIval, and IIIval, although the rmse's are in these cases correspondingly higher for all bins. Model IIARD is able to predict IIval with comparable performance as the other models but not Ival which gives significantly higher rmse's even for compounds which are the more similar to the ones in the training set. Two reasons can be put forward to explain this behavior. First, as already evoked, the model cannot learn the properties of ionizable compounds which are missing in the training set, but this should be reflected in the descriptors distances. Second, although the set of six descriptors of model IIARD is sufficient to describe and predict the reduced chemical space of the in-house compounds, it is not large enough to have good generalization properties.

The consistent higher rmse's of IIIARD compared to those of IARD is puzzling. It is intuitively felt that IIIARD, being trained on a more chemically diverse database, should be at least equal or better at predicting solubility than IARD. On another hand the descriptors distances $NDD_{x,ref}$ take into account only the descriptor space of the model, not the characteristics of the BNN. Since the BNN are made by a set of individual networks which all, after training, fit the data, the distribution of these networks is different on various parts of the descriptors space.[36] This latter property is quantified in sd_{pred}, the standard deviation of the prediction. To combine the BNN properties and the descriptors proper-

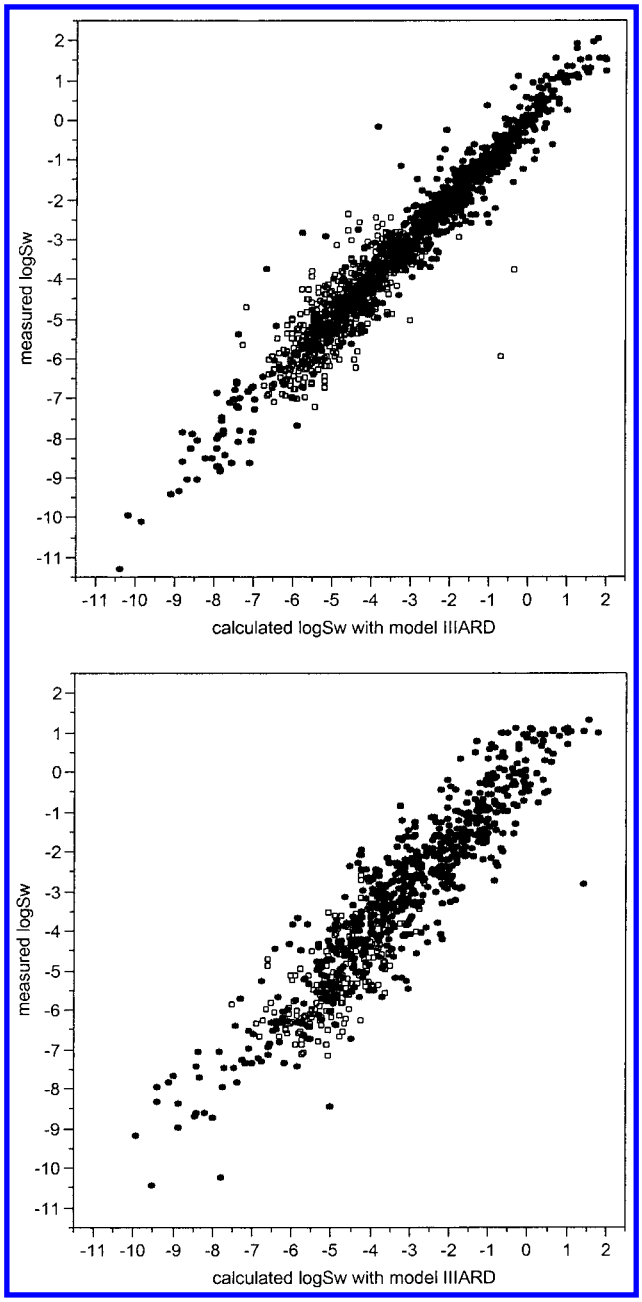


Figure 4. (a) Training of model IIIARD with database III. Solid circles are part of subdatabase I. Open squares are part of subdatabase II. (b) Validation of model IIIARD with database IIIval. Solid circles are part of subdatabase Ival. Open squares are part of subdatabase IIval.

ties $CD_{x,mod}$ has been calculated for all compounds of the validation sets and similarly to $NDD_{x,ref}$ was hashed into bins as shown in Table 8 and in Figure 3(a−c). When using the $CD_{x,mod}$ criteria to evaluate the distance between validation and training datasets, the performances of models IIIARD and IARD are very similar for all the validation datasets and thus $CD_{x,mod}$ is a good parameter to evaluate the ability of a model to predict the solubility of an unknown compound.

CONCLUSIONS

In this work, it has been shown that (i) the automatic relevance determination is a reliable and a better method than GS, to select a good model with as few descriptors as possible from a large data matrix, (ii) that the Bayesian learning of neural nets is a very powerful tool to train the models without the limitations usually found with classical neural nets, (iii) that the usual cross-validation methods used to judge the ability of a model to predict unknown compounds are not well adapted to predicting compounds outside the descriptor space of the training set, (iv) that the descriptors distance and, even better, its combination with the standard deviation of the Bayesian predictions are excellent parameters to evaluate the confidence we may have in a prediction of the solubility, and (v) that the value of a predictive model depends on the quality of the data used in the training set and on chemical diversity of the training set.

The methodology, described in this work, which has been largely automatized has been successfully applied to other QSAR/QSPR problems, in particular in the domain of predictions of ADME properties of candidate drugs such as plasma protein binding and various permeabilities.[65]

ACKNOWLEDGMENT

The author wants to thank Drs. D. Leahy, A. Wilkinson, and P. Siret for their constant support of this work, Drs. J. J. Morris and A. Davis for reviewing the manuscript and for their helpful comments, and especially Dr. D. Cosgrove for his invaluable help in setting up numerous useful routines and for solving intractable programming problems.

REFERENCES AND NOTES

(1) Yalkowsky, S.; Banerjee, S. *Aqueous Solubility. Methods of Estimatiom for Organic Compounds*; Marcel Dekker: New York, 1992; 264pp.
(2) Morris, J.; Bruneau, P. Prediction of Physicochemical Properties. In *Virtual Screening for Bioactive Molecules*; Böhm, H., Schneider, G., Eds.; Wiley-VCH: 2000; pp 33−58.
(3) Yalkowsky, S.; Valvani, S. Solubility and Partitioning I: Solubility of Nonelectrolytes in Water. *J. Pharm. Sci.* **1980**, *69*, 912−922.
(4) Unpublished results.
(5) Irmann F. Eine eeinfache Korrelation zwischen Wasserlöslichkeit und Struktur von Kohlenwasserstoffen und Hologenkohlenwasserstoffen. *Chem. Ing. Tech*. **1965**, *37*, 789−798.
(6) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. A Method for Calculation of the Aqueous Solubility of Organic Compounds by Using New Fragment Solubility Constants. *Chem. Pharm. Bull*. **1986**, *34*(11), 4663−4681.
(7) Suzuky, T. Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comput.-Aided Mol. Design* **1991**, *5*, 149−166.
(8) Abraham, H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci*. **1999**, *88*(9), 868−880.
(9) Myrdal, P.; Manka, A.; Yalkowsky, S. Aquafac 3: Aqueous functional group activity coefficients; Application to the estimation of aqueous solubility. *Chemosphere* **1995**, *30*(9), 1619−1637.
(10) Lee, Y.; Myrdal, P.; Yalkowsky, S. Aqueous functional group activity coefficients (Aquafac) 4: Applications to complex organic compounds. *Chemosphere* **1996**, *11*, 2129−2144.
(11) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 474−482.
(12) Kühne, R.; Ebert, R.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*(11), 2061−2077.
(13) Bodor, N.; Huang, M.-J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci*. **1992**, *81*, 954−960.
(14) Mitchell, B.; Jurs, P. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 489−496.
(15) Katritzky, A.; Wang, Y.; Sild, S.; Tamm, T.; Kalrelson M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 720−725.

(16) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural network Modeling. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 450−456.
(17) CLOGP V4.62 and CMR Daylight Chemical Information Software. Daylight Chemical Information Inc.: 27401 Los Altos, Suite #370 Mission Viejo, CA 92691.
(18) Taskinen, J. Prediction of aqueous solubility in drug design. *Current Opinion Drug Discovery Devel.* **2000**, *3*(1), 102−107.
(19) Clark, D. Rapid Calculation of Polar Molecular Surface Area and its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci*. **1999**, *88*(8), 807−814.
(20) Stanton, D.; Jurs, P. Development and Use of Charge Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem*. **1990**, *62*, 2323−2329.
(21) Stanton, D.; Jurs, P. Development and use of charged Partial Surface Area Structural descriptors in computer-assisted quantitative structure−property relationship studies. *Anal. Chem*. **1990**, *62*, 2323−2329.
(22) Katritzky A.; Mu L.; Kalrelson M. A QSPR Study of the Solubility of Gases and Vapors in Water. J. *Chem. Inf. Comput. Sci*. **1996**, *36*, 1162−1168.
(23) Dannenfelser, R.-M.; Yalkowsky, S. Predicting the Total Entropy of Melting: Application to Pharmaceuticals and Environmentally Relevant Compounds. *J. Pharm. Sci*. **1999**, *88*(7), 722−724.
(24) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc*. **1991**, *113*, 9480−9483.
(25) Collantes, E.; William, D. Amino Acid Side Chain Descriptors for Quantitative Structure−Activity Relationship Studies of Peptide Analogues. *J. Med. Chem*. **1995**, *38*, 2705−2713.
(26) Wilson, L.; Farmini, G. Modeling Molecular Interactions. *Book of Abstracts;* 217th ACS National Meeting, Anaheim, CA, March 21−25, 1999.
(27) Grant, J.; Pickup, B.; Nicholls, A. A Smooth Permittivity Function for Poisson−Boltzmann Solvation Methods. *J. Comput. Chem*. **2001**, *22*, 608−640.
(28) Otto, M. *Chemometrics. Statistics and Computer Application in Analytical Chemistry*; Wiley-Vch: 1999; 313pp.
(29) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophysics Mol. Biol*. **1998**, *70*, 175−222.
(30) Tetko, I.; Livingstone, D.; Luik, A. Neural Network Studies. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 826−833.
(31) Livingstone, D.; Manallack, D.; Tetko, I. Data modelling with neural networks: Advantages and limitations. *J. Comput.-Aided Mol. Design* **1997**, *11*, 135−142.
(32) Andea, T.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem*. **1991**, *34*, 2824−2836.
(33) Manallack, D.; Livingstone, D. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 293−318.
(34) Sarle, W.: ftp://ftp.sas.com/pub/neural/FAQ3.html.
(35) Ajay, W.; Murcko, M. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem*. **1998**, *41*, 3314−3324.
(36) Burden, F.; Winkler, D. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem*. **1999**, *42* (16), 3183−3187.
(37) Burden, F.; Winkler, D. New QSAR Methods Applied to Structure−Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 236−242.
(38) Burden, F.; Ford, M.; Whitley, D.; Winkler, D. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1423−1430.
(39) Neal, R. *Bayesian Learning for Neural Networks*; Springer Pub.: 1996.
(40) Hoffmann, R.; Minkin, V.; Carpenter, B. Ockham's Razor and chemistry. *Bull. Soc. Chim. Fr*. **1996**, *133*, 117−130.
(41) Sutter, J.; Dixon, S.; Jurs, P. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.
(42) Lučić, B.; Trinajstić, N.; Sild, S.; Kalrelson, M.; Katritzky, A. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 610−621.
(43) Waller, C.; Bradley, M. Development and Validation of a Novel Variable Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 345−355.
(44) So, S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem*. **1996**, *39*, 1521−1530.
(45) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 699−704.
(46) Viswanadhan, V.; Mueller, G.; Basak, C. A New QSAR Algorithm Combining Component Analysis with a Neural Network: Application to Calcium Channel Antagonist. Network Sci. [Electronic Publication: http://www.netsci.org/Science/Compchem/feature07.html] 1996.
(47) Kovar, T. Genetic Function Approximation Experimental Design (GFAXD): A New Method for Experimental Design. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 858−866.
(48) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optirnal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat*. **1993**, *12*, 7550.
(49) Araujo, O.; Morales, A. Properties of New Orthogonal Graph Theoretical Invariants in Structure−Property Correlations. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 1031−1037.
(50) Duprat, A.; Huynh, T.; Dreyfus, D. Toward a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR. Application to the Prediction of LogP. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 586−594.
(51) Radford, N. Software for Flexible Bayesian Modeling, version of 06−12−1999. http://www.cs.utoronto.ca/∼radford.
(52) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 773−777.
(53) SMILES tutorial, Daylight Chemical Services, http://www.daylight-.com/smiles.
(54) Weininger, D. SMILES, a Chemical Language and Information System 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 31.
(55) Thanks to Jarmo Huuskonen for providing an electronic version of his database.
(56) Available Chemicals Directory. Distributed by MDL Information Systems Inc. 2132 Farallon Drive, San Leandro, CA 94577. http://www.mdli.com.
(57) SD a 2D or 3D molecular structure format including data, from MDL Information Systems Inc 2132 Farallon Drive, San Leandro, CA 94577. http://www.mdli.com.
(58) Daylight Chemical Information Software. Daylight Chemical Information Inc.: 27401 Los Altos, Suite #370 Mission Viejo, CA 92691.
(59) CONCORD, a Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures; The University of Texas at Austin and Tripos Associates: St. Louis, MO.
(60) Sybyl V6.5 distributed by Tripos Inc. St. Louis, MO, http://www.tripos.com.
(61) SMARTS, Daylight Chemical Services, http://www.daylight.com/products/smarts_kit.html.
(62) JMP version 3.2.6. Distributed by SAS Institute Inc. Http://www.JMPdiscovery.com.
(63) Unpublished results. The standard deviation of repeated measurements using the same batch of a compound is much lower.
(64) Heller, S.; Bigwood, D.; May, W. Expert Systems for Evaluating Physicochemical Property Values. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 627−636.
(65) H2X a package of predictive models, accessible on the AstraZeneca intranet.
(66) When the number of examples is high and the numbers of descriptors are comparable, SD is almost identical to rmse.

CI010363Y