

# Prediction of Aqueous Solubility Based on Large Datasets Using Several QSPR Models Utilizing Topological Structure Representation

Joseph R. Votano\*, Marc Parham  
ChemSilico LLC, 48 Baldwin Street, Tewksbury, MA 01876

Lowell H. Hall  
Department of Chemistry  
Eastern Nazarene College, Quincy, MA 02170

Lemont B. Kier  
Department of Medicinal Chemistry  
School of Pharmacy  
Virginia Commonwealth University, Richmond, VA 23298

L. Mark Hall  
Hall Associates Consulting  
2 Davis Street  
Quincy, MA 02170

## ABSTRACT

Several QSPR models were developed for predicting aqueous solubility,  $S_o$ . A dataset of 5,964 compounds was subdivided into two classes, aromatic ring containing and non-aromatic compounds. Three models were created with different methods on both data sets: two regression models (multiple linear regression and partial least squares) and an artificial neural network model. These models were based on 3343 aromatic and 1674 non-aromatic compounds with 938 compounds used in external validation testing. The range in  $-\log S_o$  was -2 to 10. Topological structure descriptors were used with all models. A genetic algorithm was used for descriptor selection for regression models. For the ANN model, descriptor selection was done with a standard backward elimination process. All models performed well with  $r^2$  values ranging 0.72 to 0.84 in external validation testing. The mean absolute errors in validation ranged from 0.44 to 0.80 for both classes of compounds for the models. These statistical results indicate a sound model. Furthermore, in a comparison with eight other available models, based on predictions on a validation test set (442 compounds), the artificial neural network model presented in this work (CSLogWS) was clearly superior based on both the mean absolute error and the percentage of residuals less than one log unit. In the ANN model both E-State and hydrogen E-State descriptors were found to be important.

## Introduction and Background

Aqueous solubility of an oral drug is an important factor in the bioavailability of a drug. Poor solubility usually translates into a higher dosage level to achieve the desired therapeutic outcome. This, in turn, can lead to eventual toxicity problems. Today, there are few benign options to circumvent poor aqueous solubility of a drug taken orally. There has been a trend to build combinatorial libraries involving higher molecular weight compounds with the likelihood of both an increase in their lipophilicity and decrease in their aqueous solubility. Hence, the need for reliable aqueous solubility prediction becomes even more crucial, especially in the early drug discovery stage. Ultimately, over a pH range from 2 to

7.4, it is essential for an orally administered drug to have adequate aqueous solubility and low first pass metabolism so that it will reach its targeted site of action in sufficient quantity to be of therapeutic value.

Several different approaches have been presented for predicting intrinsic aqueous solubility ( $S_o$ ), defined as the solubility (mol/L) in an unbuffered solution for the uncharged form of the compound. Compound descriptor sets that have been used in these previous modeling approaches can be considered in three principal categories: bulk properties [1-3] i.e., melting point and  $\log P$ ; atom/group contributions [4-6], and those dependent on structural and electronic properties of compounds [7-15] with or without a bulk property included.

---

\* Author to whom correspondence should be address:  
JVotano@ChemSilico.com

All models [11,12,15] for estimating aqueous solubility using moderately-sized databases ranging from approximately 1,300 to 2,400 compounds have employed topological structure representation developed by Kier and Hall, including molecular connectivity and atom-type E-state indices which have also been used in a wide variety of models [16-30]. These structure descriptors were used either alone or in conjunction with numerous others; e.g., partial charged surface areas, polarizability, and partial atomic charges. For these published datasets, construction of the quantitative structure-property relationships (QSPR) models used either artificial neural network techniques (ANN) or partial least squares (PLS) employing a genetic algorithm (GA) for descriptor selection. The PLS-GA model [15] yielded a root mean square error (RMSE) of approximately 1 log unit on 1,665 validation compounds. The model utilized topological descriptors and logP, charged partial surface parameters, and projected volume descriptors [31]. The artificial neural network (ANN) models [11,12] had smaller standard errors (RMSE = 0.53 and 0.60) respectively on 413 and 412 compounds in external validation testing. One ANN model [12] used only atom-type E-state descriptors while the other model used many types of topological descriptors.

In this present study, a large database of 5,964 highly diverse compounds was built, consisting of 1,849 non-aromatic and 4,115 aromatic compounds. QSPR models were developed using multiple linear regression (MLR), PLS, and ANN methods. All models employed only topological descriptors. These structure descriptors encode whole molecule structure information as well as atom level descriptors that encode both the topological environment of each atom and also the electronic influence of all other atoms.

**Data Sets Description.** Datasets were constructed from six sources: the Aquasol database [32], PhysProps database [33], PDR [34], and datasets kindly provided by Taskinen [10], Tetko [12], and Lobell [35]. The datasets were supplied in SMILES format and converted to Mol files with careful checking of the resultant structures to insure correct assignment of aromaticity. Experimental aqueous solubility values were obtained at or between 20 and 25 °C and expressed as the common logarithm of that value,  $-\log(S_0)$  (= pS). A survey of 75 compounds selected randomly from the Aquasol database was made to assess DelpS, the difference in pS at 20 and 25°C, reported by the same source. The average DelpS was found to be 0.11 log units. Since solubility values were generally not available with their associated experimental errors, an additional survey was done on another 75 compounds measured at same the temperature, 20 or 25 °C, but from different sources. A value of  $0.37 \pm 0.40$  was found for DelpS (and its standard deviation). Such a large a variation is not unexpected due the variability in the purity of compounds, experimental protocols, and analytical methods employed from laboratory to laboratory in conducting solubility determinations. However, these DelpS values do provide some indication for the expected lower bound for the error in predicting pS for compounds not included in the model development.

A database was constructed from all these sources and examined for duplicate molecular structures and for multiple experimental values with the same molecular structure. Compound duplications were determined by comparing both the sum of all computed descriptor values and molecular weight values. Agreement of the sum values to within 0.001 was considered an indication of compound duplication; one of the structures was removed from the data set. If duplicate pS values did not agree within 0.4 in pS units, both were disregarded; otherwise the lower pS value was retained. The final dataset of 5,964 compounds was sub-divided into two classes: those containing aromatic ring systems and those that do not. To indicate structure diversity in the data sets, Table 1 provides a list of compound structure attributes for the training sets for both classes. Approximately 25% of the aromatic class contained nitrogen-heterocyclic aromatic rings; approximately 33% contained fused ring systems. About 33% of the non-aromatic class contained at least one ring.

**Table 1.** Attributes of Compound Training Sets<sup>a</sup>

Compound Attribute	Non-Aromatic (No. of Cpds)	Aromatic (No. of Cpds)
Ring(s)	579	3343
Fused ring(s)	199	1127
N-heteroaromatic ring(s)	0	833
Aromatic Rings(only)	0	2510
<Rotbonds>	6.3	6.2
<numHBd>	1.1	1
<numHBa>	3.6	4.6
Halogen(s)	289	1145
Amine(s)	578	1094
-O-	486	1193
-C=O	766	1677
-OH	360	768
-CO2H	299	373
-SH	160	11
<MW>	178.0	262.8
Therapeutic Drugs	47	146
Total Compounds	1674	3343

<sup>a</sup><Rotbonds>: average number of rotatable bonds; <numHBd,a>: average number of H-bond donors or acceptors; Halogen(s): compound contains one or more F, Cl, Br, or I atoms; Amine(s): compound contains primary, secondary, or tertiary amines; OH and -C=O independent of CO2H.

## Description of Methods

**Selection of Validation Sets.** The term validation set means compounds not used in any manner in the model building or in the descriptor selection process. The validation set compounds may be called new chemical entities (NCEs). Selection of an external validation set for the non-aromatic class was performed randomly to give 166 compounds, equal to 10% of the train/test set (1674). The external validation set for the aromatic class is composed of 772 compounds; 420 were provided by Lobell [35] and 352 selected randomly from the aromatic class of compounds prior to model building. The remainder, 3343 aromatic compounds, composed the train/test set.

**Descriptor Selection Process.** An initial set of 542 computed structure descriptors (indices) was reduced to 128 for the non-aromatic class and 160 for the aromatic class using the criterion that at least 5% of the descriptor values must be non-constant (usually non-zero). These two descriptor sets were used as the initial sets in the modeling process. Final descriptor selection for PLS and MLR was performed using a genetic algorithm in the QsarIS software [36], resulting in 67 descriptors for the aromatic and 52 for non-aromatic training sets. The genetic algorithm (GA) was driven by  $r^2$  optimization and qualified by the reciprocal of the Friedman's lack-of-fitness function [37]. For the ANN model, descriptor selection was accomplished by the standard backward elimination method [38], resulting in selection of 47 descriptors for the aromatic set and 35 for the non-aromatic set.

**Analysis of Data.** PLS, MLR, and cluster analysis was accomplished with the QsarIS software [36]. For the principal component analysis (PCA) employed JMP [39]. In the MLR modeling process, both forward and backward regression was used to assess any substantial changes in the statistical outcomes for the addition or removal of a descriptor. None were found. Goodness of fit was determined by  $r^2$ ,  $q^2$ , and the F statistic with all parameters accepted at the 95% confidence level. A 100-fold randomization of pS values was performed an  $r^2$  computed for each case, (standard method in QsarIS), yielding an average  $r^2$  less than 0.02 in all MLR-GA models. The results of this randomization method indicate that the model is different from an equation based on random numbers, indicating that significant information is contained in the model. Cross-validation using the leave-one-out method (LOO) gave less than a 3% decrease in  $r^2$  for both the PLS and MLR models for two training sets, aromatic and non-aromatic compounds. In PLS modeling, the number of latent variables (LV) was determined with the criterion that adding a latent variable must improve the sum of residual squared error (RSS) with at least a 0.25% increase.

ANN analysis was performed on 90% of the dataset (train/test) with 10% set aside for external validation. The train/test set,

designated the principal set, was randomly split into 85% for train and 15% as a selection set for early stopping of the learning process to avoid over fitting. The train set was selected randomly 10 times in 10 folds of data using 75% of the train set and a mutually exclusive 10% withheld set where each compound appears only once in each withheld set. This multiple selection process gives a set of 10 models derived from the principal set using 10 mutually exclusive withholding sets. Using this approach the non-contributory variables are pruned to give an optimal subset of significant variables. The relative importance of each eliminated variable is based on its contribution across the entire train/withholding sets by calculation of  $r^2$  in each instance when the row (compound) appears in the withholding set. This value is designated  $q^2$ , that is, the  $r^2$  value for all instances when the data was withheld from the modeling process. Since  $q^2$  is used to select the variables, it does not provide a completely accurate assessment of the predictive accuracy of the overall algorithm. This task is reserved for a validation set. The standard back propagation network is used with no more than 9 hidden neurons, using the backward elimination approach [38] adapted from traditional linear regression approaches. The 10 fold cross-validation algorithm is used as a consensus model in which the average value of 10 neural nets gives the predicted pS value for a compound.

## Results and Discussion

**Statistical Evaluation of Models.** Table 2 presents a summary of results from the six QSPR models, three each for aromatic and non-aromatic data sets. All models used 3343 compounds as train/test set for aromatic and 1674 for the non-aromatics and 938 compounds for the external validation test sets, 166 compounds and 772 compounds in the two sets, as described above. The MLR and PLS models used the same GA selected descriptors except that highly correlated variables ( $r^2 > 0.8$ ) were removed for the MLR-GA models, leaving 54 and 42 descriptors respectively for aromatic and non-aromatic classes. All models performed well with training sets. Overall, the best results are obtained from the ANN model as indicated in Table 2.

**Table 2.** Statistical Parameters for Aromatic and Non-Aromatic Datasets<sup>a</sup>

Model	Nv	LV	N	$r^2$	MAE	RMSE	%Cpds (AE<0.5)	%Cpds (AE<1)	N	$r^2$	MAE	RMSE	%Cpds* (AE<0.5)	%Cpds (AE<1)
Aromatic Compounds(Train Set)									Aromatic Compounds(Validation Set)					
ANN	47		3343	0.88	0.51	0.74	63	88	772	0.77	0.62	0.89	58	82
PLS-GA <sup>b</sup>	67	51	3343	0.79	0.71	0.97	49	76	772	0.72	0.78	1.04	42	72
MLR-GA <sup>b,c</sup>	54		3343	0.77	0.75	1.01	44	75	772	0.72	0.76	1.01	43	73
Non-Aromatic Compounds(Train Set)									Non-Aromatic Compounds(Validation Set)					
ANN	35		1674	0.88	0.44	0.61	67	93	166	0.84	0.56	0.75	55	86
PLS-GA <sup>b</sup>	52	35	1674	0.79	0.61	0.81	52	82	166	0.78	0.68	0.87	43	78
MLR-GA <sup>b,c</sup>	42		1674	0.76	0.63	0.83	52	83		0.76	0.66	0.88	49	79

\* Percentage of compounds with predicted absolute error (AE) less than specified amount.

<sup>a</sup> ANN: artificial neural net; PLS-GA and MLR-GA: partial least squares and multiple linear regression; Nv = number of variables; LV: number latent variables; N: number of compounds;  $r^2$ : Square of correlation coefficient;

MAE: Mean absolute error  $\Sigma(|\log S_{\text{calc}} - \log S_{\text{exp}}|) / N$ ; RMSE:  $\text{RMSE} = \sqrt{\Sigma(|\log S_{\text{calc}} - \log S_{\text{exp}}|)^2 / N}$ .

<sup>b</sup> Genetic algorithm selected variables. <sup>c</sup> MLR model with no pairwise correlated variables with  $r^2 > 0.80$ .

For validation of the aromatic class model, the decrease in  $r^2$  for the three models is well within accepted limits as shown in Table 2 (comparing the left side of the table to the right). Furthermore, the ANN model gave the best performance when considering the additional criterion based on assessment of the size of residuals. Table 2 shows the percent compounds predicted within 0.5 log units of the experimental value. The ANN model yielded a 38% larger percentage when compared to the regression models for the aromatic group and 28% for the non-aromatic group. A two-tailed significance test for  $r^2$  gave p values  $< 0.002$  (95% confidence interval) for the ANN model. On the basis of these comparisons, the ANN model is statistically more significant than either PLS-GA or MLR-GA for aromatics.

For the non-aromatic dataset, all models (ANN, PLS-GA, and MLR-GA) did well in fitting the training set of 1,674 compounds. The PLS and MLR statistical parameters (Table 2) indicate both models are statistically similar; the MLR-GA yielded a slightly better performance in the external test set. Statistically, the ANN model is more significant than the PLS-GA model at the 90% confidence interval and more significant at the 95% level with respect to the MLR-GA model, based on a two-tailed significance test. Overall, all models performed better with this validation set when compared to its aromatic counterpart.

Figure 1a

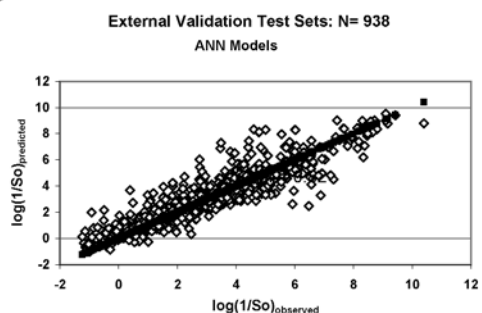
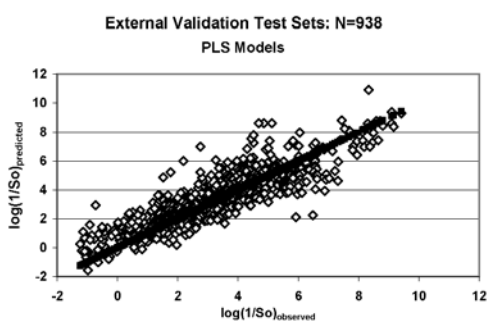


Figure 1b



**Figure 1.** Correlation of  $\log(1/S_o)$  for predicted versus observed intrinsic aqueous solubilities for 938 aromatic and non-aromatic validation compounds. In Fig.1a are the results from the ANN analysis using 47 and 35 variables respectively. In Fig.1b are similar results from PLS using 51 latent variables for the aromatic class and 35 latent variables for the non-aromatic set.

Figure 1 contains a plot of the combined external validation sets for aromatic and non-aromatic compounds for the ANN model (1a) and for the PLS-GA model (1b). The ANN plot shows a much tighter clustering of predicted value along the diagonal (observed values), reflecting the higher percentage of compounds predicted within 0.5 log units of the experimental value as compared to PLS-GA model.

**Structure Features in Models.** The structure features found to be important in the ANN model yield some useful information about the relation between molecular structure and intrinsic aqueous solubility. Although the MLR, PLS and ANN models all shared several structure descriptors in common, the ANN model allows for non-linear relationship between structure and solubility.

Structure features found to be important in the ANN model include the electron accessibility as encoded in the E-State indices [16,22,27,30]. Especially significant is the electron accessibility on the electronegative atoms N, O, and S as represented by their E-State indices. Their presence tends to yield higher predicted solubility. Also the participation of hydride groups with nitrogen and oxygen atoms in hydrogen bonding is indicated by the presence of descriptors of hydrogen bond donor and acceptor strength. Further, internal hydrogen bonding leads to decreased predicted solubility. Specific E-State descriptors for amines indicate their individual importance for both aromatic and non-aromatic amines. The importance of polar regions in the molecules is further indicated for atoms with the largest E-State value (also largest hydrogen E-State value). The strength of organic acids is also an important structure feature based on hydrogen E-State descriptors for organic acid strength. Non-polar regions of molecules also play an important role based on E-State and hydrogen E-State descriptors for alkyl groups, leading to lower predicted solubility.

In addition to an emphasis on electron accessibility (E-State indices), skeletal ramification is found to be important. Molecular connectivity chi indices and kappa shape indices are included in the model. The low order chi indices in these models encode the degree of branching in the skeleton. The model indicates that an increase in molecular size generally leads to lower solubility. Difference chi indices present in the model represent skeletal variation independent of molecular size whereas the molecular connectivity chi indices include size. For this model, adjacency of branching also plays an important role, indicating that tightly branched compounds are predicted to be less soluble. Furthermore, the kappa shape indices in the model indicate the importance of taking overall molecular shape into account for predicting solubility.

**Table 3.** Hierarchical Clusters of Datasets<sup>a</sup>

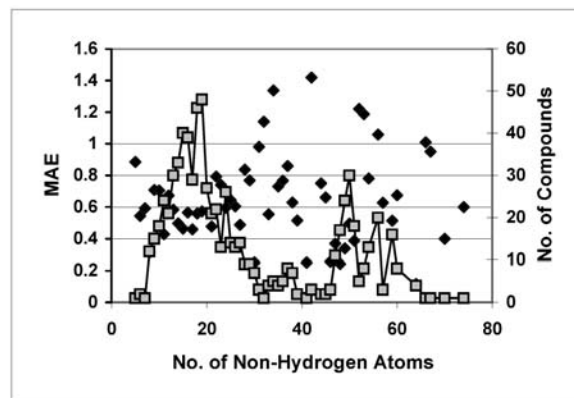
Cluster No.	Non-Aromatic		Aromatic	
	No. of cpds. in cluster	% of NCEs in cluster	No. of cpds. in cluster	% of NCEs in cluster
1	59(93)	5%(9%)	323(282)	18%(18%)
2	9(21)	22%(10%)	740(1246)	11%(18%)
3	205(175)	11%(10%)	381(305)	8%(9%)
4	44(9)	11%(22%)	382(408)	11%(14%)
5	167(49)	11%(16%)	866(389)	20%(11%)
6	103(258)	14%(10%)	160(847)	26%(10%)
7	182(31)	6%(3%)	346(74)	14%(12%)
8	422(348)	9%(9%)	270(161)	24%(23%)
9	446(62)	8%(13%)	385(146)	8%(26%)
10	212(803)	8%(8%)	262(257)	75%(76%)
<b>Total</b>	<b>1849</b>		<b>4115</b>	

<sup>a</sup> Hierarchical clustering used 35 and 47 molecular descriptors from the neural analysis and 42 and 54 for MLR-GA for the non-aromatic and aromatic training sets. Ward's clustering algorithm was used with three (3) initial starting clusters. Aromatic and non-aromatic classes contained 772 and 166 validation compounds respectively. Numbers in parentheses are for MLR results. NCE: New Chemical Entities, compounds in the external validation test set, not used in model development.

It is of some interest to determine how the structure descriptors obtained in the modeling can perform in clustering the whole data set. In particular, a comparison can be made between the structure space provided by the MLR model compared to the ANN model. Hierarchical clustering was carried out using Ward's method in QsarIS [36]. Table 3 presents the results, based on a model for ten-clusters, for both ANN and MLR (in parenthesis). The cluster sizes for the two QSPR models differ as expected since the ANN variables differ from those found for MLR with only about 45% of the descriptors being the same. Nonetheless, the compounds in the external validation set (NCE) do occur in all of the clusters. The NCEs are more evenly distributed in the non-aromatic set. In the aromatic set, cluster ten for both ANN and MLR-GA models has a specific interest. It contains approximately 75% NCEs. These particular NCEs are highly diverse, high molecular weight compounds [35]. The lower value of  $r^2$  with aromatics in all models may be due to contribution of the 420 NCEs from Lobell [41]. Their average molecular weight is 397 compared to 263.3 for the training set.

In the principal component analysis, the first four PCA scores accounted for 71.9% and 62.9 % of the variance of ANN descriptors for non-aromatic and aromatic classes respectively whereas the GA selected descriptors accounted for 53.8% and 49.9% of the variance respectively. Several additional cluster models, using 15 and 20 clusters, showed no significant differences in trends of percentages of NCEs in clusters when compared with the ten-cluster model.

Although cluster ten seems to contain a significant portion of high molecular weight compounds, no trend is found between molecular weight and error in the predictions from the models. Tetko reported, however, a dependency on number of non-hydrogen atoms versus RMSE values for his ANN model [12]. Examination of our results for all models (ANN, PLS, MLR; aromatic, non-aromatic) does not indicate any meaningful correlation with MAE or RMSE; all  $r^2$  values were  $< 0.05$ .



**Figure 2.** Mean absolute error (MAE),  $<|\log S_{obs} - \log S_{pred}|>$ , versus number of non-H atoms for 772 aromatic compounds in the validation set. The grayed boxes are molecules with same number of non-H atoms. Black diamonds are the MAE values for groups of with same number of non-H-atoms. MW ranged from 68.1 to 1035.3 with  $<MW> = 397.2 \pm 217.4$ .

Figure 2 shows the variation of MAE as function of the number non-hydrogen atoms in the 772 compound aromatic validation set (grayed squares). The correlation between MAE and count of non-hydrogen atoms is  $r^2 = 0.01$ . Also shown is the number of molecules for each count of non-hydrogen atoms (black diamonds).

The rationale for sub-dividing the 5,964 dataset into aromatic and non-aromatic subsets arises from two primary considerations. First, in the aromatic systems conformational constraints due to the presence of bulky and rigid planar ring systems can diminish the opportunity for intra and intermolecular hydrogen bonding interactions among substituent groups both in crystals and in solution. On the other hand, resonance assisted hydrogen bonding can be very strong between aromatic systems containing hydrogen donors and acceptors [40]. Indeed, conformational constraint effects exist in 33% of non-aromatic compounds; however it is a substantially smaller factor than with the aromatics. A second factor was the anticipation that the selected descriptors would be reasonably different for the aromatic and non-aromatic classes regardless of which QSPR approach used.

**Comparison with Other Models.** It is most difficult to make a direct comparison of the model presented here with other ANN or regression models for at least three reasons: (1) significantly larger size of our dataset, (2) subdivision of our data into two classes, and (3) differences in topological indices employed. However, a strong indication of the high quality of our ANN based QSPR model can be obtained from a comparison study [41]. Lobell compared results from nine published or commercial models (including the ANN model presented here, CSLogWS) using a 442 member validation set of observed  $S_o$  values, known not to be included in our

**Table 4.** Aqueous Solubility [-log(So)] Predictive Results for 442 Predominately Uncharged Compounds<sup>a</sup> Based on Available Models, Including CSLogWS, as Described in this Work

Method to Predict -log(S <sub>o</sub> )	Errors <sup>b</sup>	r <sup>2</sup>	MAE	MRE <sup>c</sup>	AE<1	AE<2
CSLogWS (ChemSilico) [42]	0	0.58	0.70	0.00	79%	93%
ws2 (Novartis) [43]	0	0.46	0.91	0.17	64%	93%
ABB [41]	0	0.5	1.04	0.03	55%	87%
ACD 6.0 (ACD Labs) [44]	85	0.65	1.06	-0.84	57%	84%
Tetko LogS [12]	8	0.46	1.17	-0.85	50%	83%
QikProp 2.0 (Shrodinger) [45]	8	0.33	1.36	-0.01	48%	73%
C2-ADME (Accelrys) [46]	3	0.19	1.35	0.84	46%	76%
PreADME [47]	0	0.68	1.49	-1.37	42%	69%
Syracuse (SyracuseResearch) [33]	0	0.58	1.85	-1.55	38%	63%

<sup>a</sup> Statistical results supplied by Mario Lobell, OSI Pharmaceuticals, Oxford, UK.

<sup>b</sup> Errors: Number of compounds that failed to yield prediction; results based on remaining compounds.

<sup>c</sup> MRE: mean relative error,  $\{ \sum [ \log(1/S_{o,exp}) - \log(1/S_{o,pred}) ] \} / N$ , where N is number of compounds.

train/test data set. A summary of these comparisons is given in Table 4. The ANN model presented here clearly gives the best statistical results, including the smallest mean absolute error, MAE = 0.70. The model with the next best result yielded MAE = 0.91; the remaining six models had an average MAE = 1.33. Furthermore, the ANN model presented here performed best when considering the percentage of predicted residuals below a certain cut-off value. The ANN model presented here predicted 79% of 442 compounds within one log unit of their experimental value, significantly more than any of the other models (Table 2). On the basis of this significant test and all the other information presented here, it seems clear that the ANN model described here is very sound for predicting aqueous solubility. These differences in the results of predictions done using the various models are a function of the modeling approach employed, the descriptors used, and the data set of compounds.

## Conclusions

The approaches to modeling aqueous solubility described here have been shown to lead to a model that produces high quality predictions. A carefully developed database of experimental solubility data was created along with computed topological descriptors of molecular structure. Parallel development led to models based on multiple linear regression, partial least squares, and artificial neural network methods. Statistical analysis clearly indicates that the superiority of the ANN model. Comparison of predictions based on this ANN model with other available models also clearly shows that this model is superior. This ANN model is now commercially available as CSLogWS [42].

In our development of ADME models, we are further investigating the three components of our approach that we think are most responsible for the high quality of the aqueous solubility model. Our search for high quality data on diverse molecular structures continues. Although the descriptors used to represent molecular structure have served us well in the development of the model, research continues for additional descriptors as well as improved descriptors. Finally,

development of modeling techniques for artificial neural networks also continues. At this point in time, it appears to us that our combination of qualified data, topological descriptors, and modeling techniques provides the basis for the high quality model described in this work.

## References

- [1] Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE), *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 354-357.
- [2] Peterson, D.L.; Yalkowsky, S. H. Comparison of Two Methods for Predicting Aqueous Solubility, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1531-1534
- [3] Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Octanol/water Partition Coefficient, *Environ. Toxicol. Chem.*, **1996**, 15, 100-106.
- [4] Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 439-445.
- [5] Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals, *Chemosphere*, **1995**, 30, 2061-2077.
- [6] Lee, Y-H.; Myrdal, P. B.; Yalkowsky, S. H. Aqueous Functional Group Activity Coefficients [AQUAFAC] 4: Application to Complex Organic Compounds, *Chemosphere*, **1996**, 33, 2129-2144.
- [7] Abraham, M. 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, 868-880.
- [8] Mitchel, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 489-496.
- [9] 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.
- [10] 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

- [11] J. Huuskonen, Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology, *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 773-777.
- [12] Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1488-1493.
- [13] McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1237-1247.
- [14] McFarland J. W.; Avdeef, A.; Berger, C. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structure Alone, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1355-1359.
- [15] Cheng, A.; Merz Jr, K. M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure-Property Relationships, *J. Med. Chem.*, **2003**, 46, 3572-3580
- [16] Kier, L. B.; Hall, L.H. *Molecular Structure Description*, Academic Press, **1999**.
- [17] a. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press Inc., New York, **1976**.  
b. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press Ltd., Hertfordshire, England and John Wiley and Sons, New York, **1986**.
- [18] Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, *J. Chem. Inf. Comput. Sci.*, **1995**, 35, 1039-1045.
- [19] Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules, *Pharm. Res.*, **1990**, 7, 801-807.
- [20] Hall, L. H.; Mohny, B. M.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR, *Quant. Struc.-Act. Relat.*, **1991**, 10, 43-51.
- [21] Hall, L. H.; Mohny, B. M.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs, *J. Chem. Inf. Comput. Sci.*, **1991**, 31, 76-82.
- [22] Kier, L. B.; Hall, L. H. in 'Advances in Drug Design' Ed. Bernard Testa, Academic Press, London, **1992**, Vol. 22, Chapter 1, pp 2-38.
- [23] Gough, J. D.; Hall, L.H. Modeling the Toxicity of Amide Herbicides using the Electrotopological State, *Environ. Tox. Chem.*, **1999**, 18, 1069-1075.
- [24] Maw, H. H.; Hall, L. H. E-State Modeling of Dopamine Transporter Binding: Validation of Model for Small Data Set, *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 1270-1275.
- [25] Maw, H. H.; Hall, L. H. E-State Modeling of Corticosteroids Binding Affinity: Validation of Model for Small Data Set, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1248-1254.
- [26] Kier, L. B.; Hall, L.H. Inhibition of Salicylamide Binding: An Electrotopological State Analysis, *Med. Chem. Res.*, **1992**, 2, 497-502.
- [27] Hall, L. M.; Hall, L. H.; Kier, L. B. QSAR E-State Modeling of Beta-Lactam Binding to Human Serum Proteins, *J. Comp.-Aid Molec. Des.*, **2003**, 17, 103-118.
- [28] Hall, L. M.; Hall, L. H.; Kier, L. B. Modeling of Drug Albumin Binding Affinity with E-State Topological Structure Representation, *J. Chem. Inf. Comput. Sci.*, **2003**, 43, 2120-2128.
- [29] Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information, *J. Chem. Inf. Comput. Sci.*, **1995**, 35, 1039-1045.
- [30] Rose, K.; Hall, L. H. Modeling Blood-Brain Barrier Penetration Using the Electrotopological State, *J. Chem. Inf. Comput. Sci.*, **2002**, 42, 651-666.
- [31] Rohrbaugh R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/activity and Structure/property Relationships, *Anal. Chim. Acta*, **1987**, 199, 99-109.
- [32] Yalkowsky S.H.; Dannelfelser, R.M. The Arizona Database of Aqueous Solubility, College of Pharmacy, University of Arizona, Tucson, AZ, **1997**.
- [33] Physical/Chemical Property Database (PHYSOPROP), Syracuse Research Corporation, SRC Environmental Research Center, Syracuse, NY, **1999**.
- [34] PDR Electronic Library, Version 6.0, Volume 2003
- [35] Personal communication. Mario Lobell, OSI Pharmaceutical, Watlington Road, Oxford OX4 6LT, UK.
- [36] QsarIS, v2, MDL Information Systems, San Leandro, CA.
- [37] Rogers, D. in *Genetic Algorithms in Molecular Modeling*; Devillers, J.; Ed.; Academic Press, San Diego, **1996**, pp 87-106.
- [38] Miller, A. in *"Subset Selection in Regression"*, 2nd Edition Chapman & Hall/CRC Press, **2002**.
- [39] JMP ver. 5.01, SAS Institute, Cary, NC.
- [40] Bertolasi, V.; Gilli, P.; Feretti, V.; Gilli, G. Intermolecular N-H-O Hydrogen Bonds Assisted by Resonance Heteroconjugated Systems as Hydrogen-bond Strengthening Functional Groups, *Acta Cryst.*, **1995**, B51, 1004-1015.
- [41] Lobell, M.; Sivarajah, V. In Silico Prediction of Aqueous Solubility, Human Plasma Protein Binding, and Volume Distribution From Calculated pKa and AlogP98 Values, *J. Molec. Diversity*, **2003**, 7, 69-87
- [42] Model from ChemSilico, 48 Baldwin Street, Tewksbury, MA 01876; <http://www.chemsilico.com>
- [43] Model from Novartis, Novartis International AG, CH-4002 Basel, Switzerland; <http://www.novartis.com>
- [44] Model from ACD Labs, 90 Adelaide Street West, Suite 600, Toronto, Ontario M5H 3V9, Canada; <http://www.acdlabs.com>
- [45] Model from Schrödinger, 120 West Forty-Fifth Street, 32nd Floor, Tower 45, New York, NY 10036 4041; <http://www.schrodinger.com>
- [46] Model from Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752; <http://www.accelrys.com>
- [47] Model from PreADME; <http://preadme.bmdrc.org>