Comparison of Predictive Ability of Water Solubility QSPR Models Generated by MLR, PLS and ANN Methods

Dániel Erös[a], György Kéri[a,b], István Kövesdi[b], Csaba Szántai-Kis[c], György Mészáros[d] and László Örfi[c,b,*]

[a] *Semmelweis University, Department of Medical Chemistry, Peptide Biochemistry Research Group of the Hungarian Academy of Sciences, Puskin u. 9., Budapest, 1088, Hungary*

[b] *Vichem Chemie Ltd., Herman Ottó u. 15., Budapest, 1022, Hungary*

[c] *Semmelweis University, Cooperative Research Center/Department of Pharmaceutical Chemistry, Högyes E. u. 9., Budapest, 1092, Hungary*

[d] *Semmelweis University, Cooperative Research Center/Department of Medical Chemistry, Puskin u. 9., Budapest, 1088, Hungary*

Abstract: ADME/Tox computational screening is one of the most hot topics of modern drug research. About one half of the potential drug candidates fail because of poor ADME/Tox properties. Since the experimental determination of water solubility is time-consuming also, reliable computational predictions are needed for the pre-selection of acceptable "drug-like" compounds from diverse combinatorial libraries. Recently many successful attempts were made for predicting water solubility of compounds. A comprehensive review of previously developed water solubility calculation methods is presented here, followed by the description of the solubility prediction method designed and used in our laboratory. We have selected carefully 1381 compounds from scientific publications in a unified database and used this dataset in the calculations. The externally validated models were based on calculated descriptors only. The aim of model optimization was to improve repeated evaluations statistics of the predictions and effective descriptor scoring functions were used to facilitate quick generation of multiple linear regression analysis (MLR), partial least squares method (PLS) and artificial neural network (ANN) models with optimal predicting ability. Standard error of prediction of the best model generated with ANN (with 39-7-1 network structure) was 0.72 in logS units while the cross validated squared correlation coefficient (Q^2) was better than 0.85. These values give a good chance for successful pre-selection of screening compounds from virtual libraries, based on the predicted water solubility.

Keywords: Virtual screening, water solubility, ADME, QSPR, MLR, PLS, ANN, external validation.

INTRODUCTION

Acceptable drug absorption highly depends on aqueous solubility [1]. ADME/Tox computational screening is one of the most hot topics of modern drug research because even a rough experimental determination of water solubility requires one magnitude greater amounts of substances than the biological assays. Nowadays 400 μg material is enough for about 30 HTS assays while water solubility determination in a single experiment needs milligrams of the given compound. Since the experimental procedure is time-consuming also, reliable computational predictions are needed for the preselection of acceptable "drug-like" compounds from diverse combinatorial libraries.

Aqueous solubility (S, [mol/l]) represents the maximum amount of solute in moles that dissolves in 1 liter of water to make a saturated solution at a given temperature. The model presented below is based on the solubility data of the unionized molecular species.

Recently many successful attempts were made for predicting water solubility of compounds (Tables 1-3). Numerous, different approaches used for the prediction of solubility have been summarized by Yalkowsky and Banerjee [2] and divided into three groups:

a) correlations with experimentally determined physico-chemical properties such as logP, melting point, boiling point, chromatographic retention data, molar volume, etc.;

b) correlations based on group contributions;

c) correlations with parameters calculated solely from the molecular structure (QSPR approaches).

Representative examples for group a), b) and c) are presented in Tables 1, 2 and 3, respectively.

The "correlations with experimentally determined physico-chemical properties"- type methods give the best results (especially for a series of molecules) because they are based on experimentally determined data. But, because of the same reasons they can not be used for in-silico screening. The group-contribution methods are rather empirical and less accurate. These methods may have difficulties with "unknown" fragments that are not implemented in their database. QSPR approaches are the best tools for in-silico

*Address correspondence to this author at the Semmelweis University, Department of Pharmaceutical Chemistry, Högyes E. u. 9., Budapest, 1092, Hungary; Tel: +361-217-0891; E-mail: orlasz@hogyes.sote.hu

1389-5575/04 $45.00+.00

Table 1. Correlations with Experimentally Determined Physico-Chemical Properties

Authors	Compound data	logS range in dataset	Descriptors	Model statistics
Yalkowsky and Valvani [3]	167 organic compounds	9 orders of magnitude	melting point, logP, entropy of fusion	aae[a]=0.5

Ran *et al.* [4]	21 organic compounds	-8.08->0.39	melting point, logP	aae=0.53
				rmse[b]=0.72

Ran *et al.* [5]	1026 organic compounds	-12.95->1.58	melting point, logP	aae=0.38
				rmse=0.53

Isnard and Lambert [6]	300 structurally diverse compounds	-	melting point, logP	sd[c]=0.466 for liquids
				sd=0.582 for solids

Warne *et al.* [7]	16 compounds	-5.1->-1.6	melting point, ASED[d]	aae=64%
Miller *et al.* [8]	12 chlorobenzenes	-	boiling point	aae=7.16%

Yaw *et al.* [9]	26 liquid alkanes	-	boiling point	aae=0.05 (1.04%)
				sd=0.05 (1.1%)

Ruelle and Kesselring [10]	531 heterogeneous compounds	-12.79->0.51	melting point, molar volume, additional term accounting for solvation effects	aae=0.371

[a] average absolute error, [b] root-mean-square error, [c] standard deviation, [d] approximate sigma electron density term

prediction of huge databases because QSPR methods are based on descriptors calculated from the molecular structure only. The accuracy of any type of prediction can never be expected higher than the accuracy of the experimental determination.

In this study we set as an aim to measure the predictive ability of models obtained by multiple linear regression (MLR) analysis, partial least squares (PLS) method and artificial neural network (ANN). We have selected carefully 1381 compounds from scientific publications [4, 23, 25, 27-29, 34], stored in a unified database and used the same dataset in the three types of calculations. In these publications the authors presented good quality predictions of aqueous solubility, all of them used reliable, validated data. The database contains aqueous solubility data for a range of –10.80 to 2.06 logS. All the data found was used in the calculations including numerous outliers published in the literature.

The standard error of prediction of the best model, generated with ANN, was 0.72 in logS units while the cross validated squared correlation coefficient (Q^2) was better than 0.85. These values give a good chance for successful pre-selection of compounds from virtual libraries, based on the predicted water solubility.

Since about one half of the potential drug candidates fail because of poor ADME/Tox properties, it is advantageous to use property based design in the very early stage of drug research, which can tend to decrease development costs. It has been shown that QSPR models, generated by our statistical methods, can reliably estimate the inherent error of the physico-chemical data and in this way they can be used as automatic quality assurance tools for controlling the experimental procedure as well.

MATERIALS AND METHODS

The database of 1381 molecules was built in ISISBase [43] on an IBM Pentium PC. We calculated the 3D structures of the molecules by the CONCORD algorithm of the Tripos software package [44]. A *structure definition* file (sdf) containing the chemical structures and experimental logS values of the molecules was created and imported into

Table 2. Correlations Based on Group Contributions

Authors	Compound data	logS range in dataset	Fragments	Model statistics

Wakita *et al.* [11]	314 aliphatic and aromatic liquids	-5.24->0.68	40 fragment terms derived from liquids, 2 fragment terms derived from aliphatic solids, 5 fragment terms derived from aromatic solids and melting point correction factors	aae=0.25
				sd=0.20

	aliphatic solids	-0.01->0.17		sd=0.151

	134 aromatic solids	-10.49->0.34		aae=0.58
				sd=0.65

Suzuki [12]	497 compounds	-10.49->1.96	10 individual atom type fragments	aae=0.39
				sd=0.505

Klopman *et al.* [13]	496 organic compounds	-	45 fragments and 1 constant	sd=1.43
Klopman and Zhu [14]	1288 organic compounds	-	118 parameters	sd=0.79
Kühne *et al.* [15]	694 organic compounds	-11.62->1.81	55 fragments, 2 melting point terms	aae=0.38

an in-house developed computer program, 3DNET [45]. The software can calculate various, user defined 1D, 2D and 3D molecular descriptors. The externally validated models were based on calculated descriptors only. The starting descriptor pool is listed in Table 4.

The descriptor calculation resulted a data file, which contained the descriptors and the experimental logS data for each molecule. This data file was fed into the statistical program 3DNET4W [64]. This program was designed for the automatic selection of the descriptors needed for the optimal structure-property (or structure-biological activity) model.

The database was divided into three main parts:

1. ***Work set*** of 1050 molecules was used in model building.

Table 3. Correlations with Parameters Calculated Solely from the Molecular Structure

Authors	Compound data	logS range in dataset	Descriptors	Model data	Model statistics
Medir and Giralt [16]	84 hydrocarbons	-	zero-order molecular connectivity index, dipole moment, number of carbon atoms, acentric factor	MLR (inverse of solubility was calculated)	sd=0.18-1.26
Nirmalakhandan and Speece [17]	145 compounds	-	zero-order valence molecular connectivity index, polarizability parameter	MLR	sd=0.311
Nirmalakhandan and Speece [18]	325 compounds	-9.32->-3.03	zero-order valence molecular connectivity index, modified polarizability parameter	MLR	sd=0.33
Patil [19]	71 PCBs	-	first-order valence molecular connectivity index	MLR	aae=0.45 (9.4%)
Makino [20]	136 PCB congener	-10.32->-5.33	6 descriptors	MLR	aae=0.1681 sd=0.225
Huibers and Katritzky [21]	109 hydrocarbons and 132 halogenated hydrocarbons	-10.41->0.51	topological and charge descriptors	MLR	se[a]=0.386
Katritzky *et al.* [22]	411 compounds	-6.44->1.57	6 descriptors	MLR	se=0.573
Huuskonen *et al.* [23]	211 drugs (51 in test set)	-5.82->0.54	topological indices	ANN (23-5-1)	r^2=0.86 s=0.53, for the test set
Sutter and Jurs [24]	diverse set of 140 compounds	-10.83->0.28	electrical, topological, geometrical descriptors	ANN (9-3-1)	rmse=0.222 for the prediction set
Mitchell and Jurs [25]	332 organic compounds	-12.8->1.57	topological, geometric, electronic descriptors	MLR	rmse=0.556 for the prediction set
				ANN (9-6-1)	rmse=0.343 for the prediction set
Engkvist and Wrede [26]	3658 molecules (307 in independent validation test)	-	1D+2D descriptors	ANN (63-5-1)	r^2=0.86 sd=0.80 for the independent validation set
McElroy and Jurs [27]	176 compounds (22 in prediction set)	-7.41->0.96	topological, geometric, electronic, "hybrid" descriptors	MLR	r^2=0.80 rmse=0.661 for the prediction set
				ANN	r^2=0.57 rmse=1.555 for the prediction set
	223 compounds (28 in prediction set)	-8.77->1.57		MLR	r^2=0.62 rmse=1.233 for the prediction set
				ANN	r^2=0.79 rmse=0.644 for the prediction set
	399 compounds (51 in prediction set)	-8.77->1.57		MLR	r^2=0.56 rmse=1.490 for the prediction set
				ANN	r^2=0.56 rmse=1.234 for the prediction set

(Table 3). contd.....

Authors	Compound data	logS range in dataset	Descriptors	Model data	Model statistics
McFarland [28]	22 drugs or drug-like molecules	-6.17->-1.38	partial atomic charges, hydrogen bond factors + calculated logP	MLR	r^2=0.82 Q^2=0.64 s=0.70
			partial atomic charges, hydrogen bond factors + measured logP	MLR	r^2=0.88 Q^2=0.83 s=0.58
Yaffe *et al.* [29]	515 organic compounds	-11.62->4.31	PM3 and topological descriptors	ANN	sd=0.26
				Fuzzy ARTMAP method	sd=0.16
Liu and So [30]	1312 organic compounds (21 in prediction set)	-11.62->1.58	1D+2D descriptors	ANN	r=0.89 s=0.91
Yin *et al.* [31]	71 aromatic sulfur-containing carboxylates	-6.24->-0.70	quantum chemical semi-empirical descriptors	MLR	r^2_{CV}=0.9095 PRESS[b]=13.1768
Jorgensen and Duffy [32]	150 organic solutes	-10.8->2.06	obtained from Monte Carlo simulations	MLR	r^2=0.88 Q^2=0.87 rmse=0.72
Collette [33]	40 diverse organic esters	-	infrared spectral data and interferogram based desc.	PLS	r^2=0.928 rmse=0.395
Huuskonen [34]	1297 organic compounds (413 in test set)	-11.62->1.58	Molecular connectivity, shape, and atom-type E-state indices	ANN (30-12-1)	r^2=0.92 s=0.60
				MLR	r^2=0.88 s=0.71
Chen *et al.* [35]	321 drugs or related compounds (54 in testing set)	-8.80->1.70	Electronic, geometric, topological descriptors	MLR	r=0.84 rmse=0.86
Bergström *et al.* [36]	17 structurally diverse drugs	~9 orders of magnitude	lipophilicity, partitioned molecular surface areas	MLR	r^2(tr)=0.91 rmse(tr)=0.61
Delgado [37]	50 chlorinated hydrocarbons	-10->-1	2 theoretical molecular descriptors	MLR	r^2=0.96 se=0.45
Jorgensen and Duffy [38]	337 compounds (20 in test set)	-	obtained from Monte Carlo simulations	MLR	r^2=0.95 rms=0.70 for the test set
Huuskonen *et al.* [39]	734 organic compounds (21 in test set)	~-12->~2.5	atom-type E-state indices	MLR	r^2=0.80 s=0.87 for the test set
				ANN (34-5-1)	r^2=0.84 s=0.75 for the test set
Wanchana *et al.* [40]	211 drugs or drug-like compounds	-5.82->0.55	topological indices	PLS	q^2=0.785 sep[c]=0.676
Abraham and Joelle [41]	659 compounds (65 in test set)	-9.02->1.97	6 descriptors	MLR	sd=0.50
Bruneau [42]	2494 compounds (934 in test set)	-	topological, geometrical and electronic descriptors	ANN (16-8-1)	rmse=0.81

[a] standard error, [b] prediction residual error sum of squares, [c] standard error of prediction

Table 4. Molecular Descriptors Used for Logs Calculation

Descriptors	No. of descriptors	Reference
Molecular mass	1	
Molecular volume, solvent extended volume	2	[46, 47, 48]
Molecular surface, solvent accessible surface, solvent extended surface	3	[47, 48, 49]
Globularity	1	[50]
WHIM descriptors of atomic mass, position, electronegativity, localized charge, atomic polarizability contribution, atomic electro topological index, pi functionality. Moments and T A V K combinations were used	7x7=49	[51]
Polarizability	1	[52, 53]
Dipole moment	1	[54]
Hildebrand solubility parameter	1	[55]
Unsaturation number	1	
Degree of chemical bond rotational freedom	1	[56]
Wiener index	1	[57]
Randic index	1	[58]
HDSA1, HDSA2, HASA1, HASA2 hydrogen bond (HB) descriptors	4	[59]
Gravitational index	1	[59]
Topological electronic index	1	[59]
QN, QO, QNO, QTOT Bodor charge descriptors for logP	4	[60]
Min, max and average of electrostatic potential (ESP) on the vdw surface	3	[52]
Min, max and average of molecular lipophilicity potential (MLP)	3	[61]
Number of specified atom types	38	[53]
Electrostatic HB basicity and acidity, max. plus summed values	4	[62]
Calculated logP by 3DNET4W	1	[63]

2. ***External validation set*** of 250 molecules validated the optimized models at the end of the optimization processes.

Table 5. Descriptors that Proved to be Important in Each Model for Explaining Experimental Solubility Data

Calculated logP by 3DNET4W [63]
Globularity [50]
Degree of chemical bond rotational freedom [56]
Unsaturation number
Hildebrand solubility parameter [55]
Electrostatic HB (hydrogen bond) basicity, max. plus summed values [62]
Electrostatic HB acidity (hydrogen bond) [62]
Topological electronic index [59]
QN, QTOT Bodor charge descriptors for logP [60]
WHIM descriptor of atomic mass, A combination [51]
Number of OH groups, number of oxygen and number of aromatic oxygen atoms [53]

3. ***Final external validation set*** of 81 molecules used as a further control of the reliability of the best obtained models.

We used the *uniform distribution method* for the creation of *work set* and *external validation set* in order to get homogenously distributed subsets in the *n*-dimensional space of the *n,* user defined, molecular descriptors [65].

In the cyclic-iterative model optimization process, the ***work set (1)*** has been randomly, repeatedly split into two halves:

- a) ***training set***: containing molecules used in the actual model building,

- b) ***monitoring set***: was used to control the predictive ability of the actual model.

According to our previous studies [63], applying the split-half method five times, gave reliable models and optimal calculation speed.

Variable subset selection (VSS) was performed with genetic algorithm (GA) or sequential selection algorithms (SSA). The aim of the optimization was to improve repeated evaluations statistics (Q^2) of the predictions and effective descriptor scoring functions were used to facilitate quick

Table 6. Average Q^2 and SEP Values of the Optimized Models for the *External Validation Sets*

Model type	MLR		PLS		ANN			
Optimized by:	SSA	GA	SSA	GA	SSA(MLR)	**GA(MLR)**	SSA(PLS)	GA(PLS)
Q2	0.89	0.91	0.88	0.91	0.93	**0.94**	0.93	0.93
SEP	0.92	0.78	0.95	0.80	0.71	**0.69**	0.71	0.73
Number of descriptors	55	39	54	50	55	**39**	54	50
Model data	-	-	40 components	38 components	5 hidden neurons	**7 hidden neurons**	4 hidden neurons	2 hidden neurons

generation of MLR, PLS models with optimal predictive ability. The model optimizations were stopped when any change in the descriptor set of the given model decreased the average Q^2 of the five cross validations for that model. The obtained models were recalculated with ANN using the ***work set (1)*** as a learning set, and the ***external validation set (2)*** as a test set. We used feed-forward networks with back-propagation learning method in the calculations. Continuously decreasing learning rate was used during the training, momentum term was not used. Network architectures with 1 to 8 hidden neurons were checked. Finally the models were “trained” on the 1300 ***work set*** and ***external validation set*** molecules ***(1+2)*** and were applied for the prediction of the logS values for the 81, "never seen before", ***final external validation set (3)*** molecules.

RESULTS

The descriptors that turned to be present in each final model are listed in Table 5. Each optimal model contains other descriptors as well, but the former are the most important ones in each model. The Q^2 and the standard error of prediction (SEP) values for the cross validations are listed in Table 6, the best model is highlighted. The Q^2 and the SEP values for the ***final external validation set*** molecules are listed in Table 7. The ***work set*** fit and the ***final external***

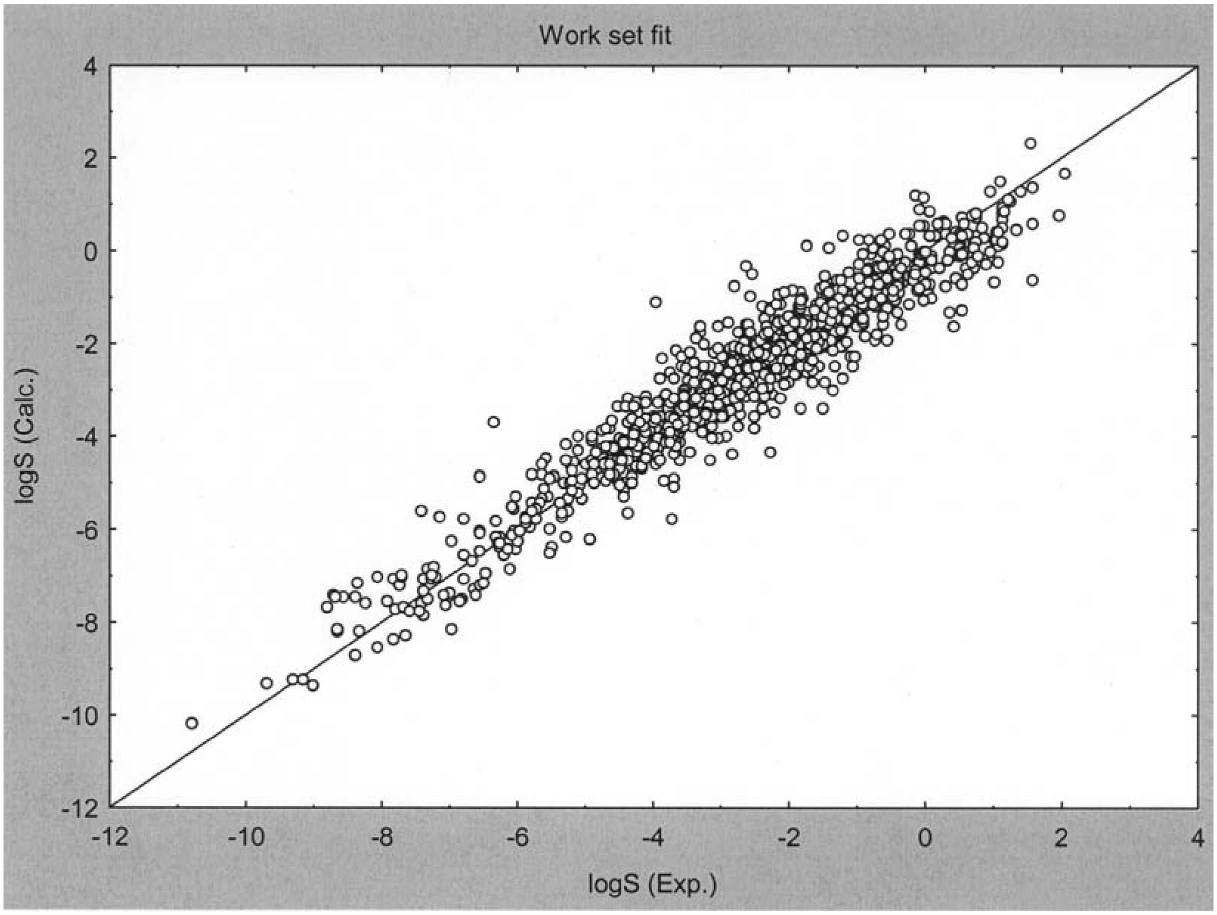

Fig. (1). ***Work set*** fit of the best ANN model.

Table 7. Validation of the Optimized Models with the 81 Molecules of the *Final External Validation Set*

Model type	MLR		PLS		ANN			
Optimized by	SSA	GA	SSA	GA	SSA(MLR)	GA(MLR)	SSA(PLS)	GA(PLS)
Q2	0.80	0.83	0.81	0.82	0.86	0.86	0.85	0.85
SEP	0.83	0.78	0.81	0.81	0.70	0.72	0.73	0.71

validation set prediction of the best model (where the starting descriptor set of the ANN model was pre-selected by the MLR method based model, applying GA) are plotted in Fig. (**1**) and Fig. (**2**).

The final ANN model used 7 hidden neurons. To compare the best ANN model to another logS prediction method we calculated the logS values of the same 81 molecules in the ***final external validation set*** with a reliable computer program ALOGPS 2.1, accessible via internet [66] (see Appendix for details). ALOGPS 2.1 uses the molecular weight and electrotopological E-state indices to estimate aqueous solubility by Artificial Neural Networks. This neural network with 33-4-1 neurons provided results with Q^2=0.83 and SEP=0.83. The results show that 3DNET is slightly better than ALOGPS 2.1. It should be mentioned that ALOGPS 2.1 uses only 1D and 2D descriptors to predict logS.

SUMMARY

Due to the importance of solubility data in drug design, various prediction methods have been developed:

a) correlations with experimentally determined physico-chemical properties,

b) correlations based on group contributions,

c) correlations with parameters calculated solely from the molecular structure.

Almost all of these methods meet the requirements they were developed for. Many of the predictions can be successfully applied for a particular compound family only. A frequently asked question is the salt formation, because none of the presently used programs can cope with this. Really we can ignore this problem in the majority of cases because salt formation, in general, increases water solubility dramatically.

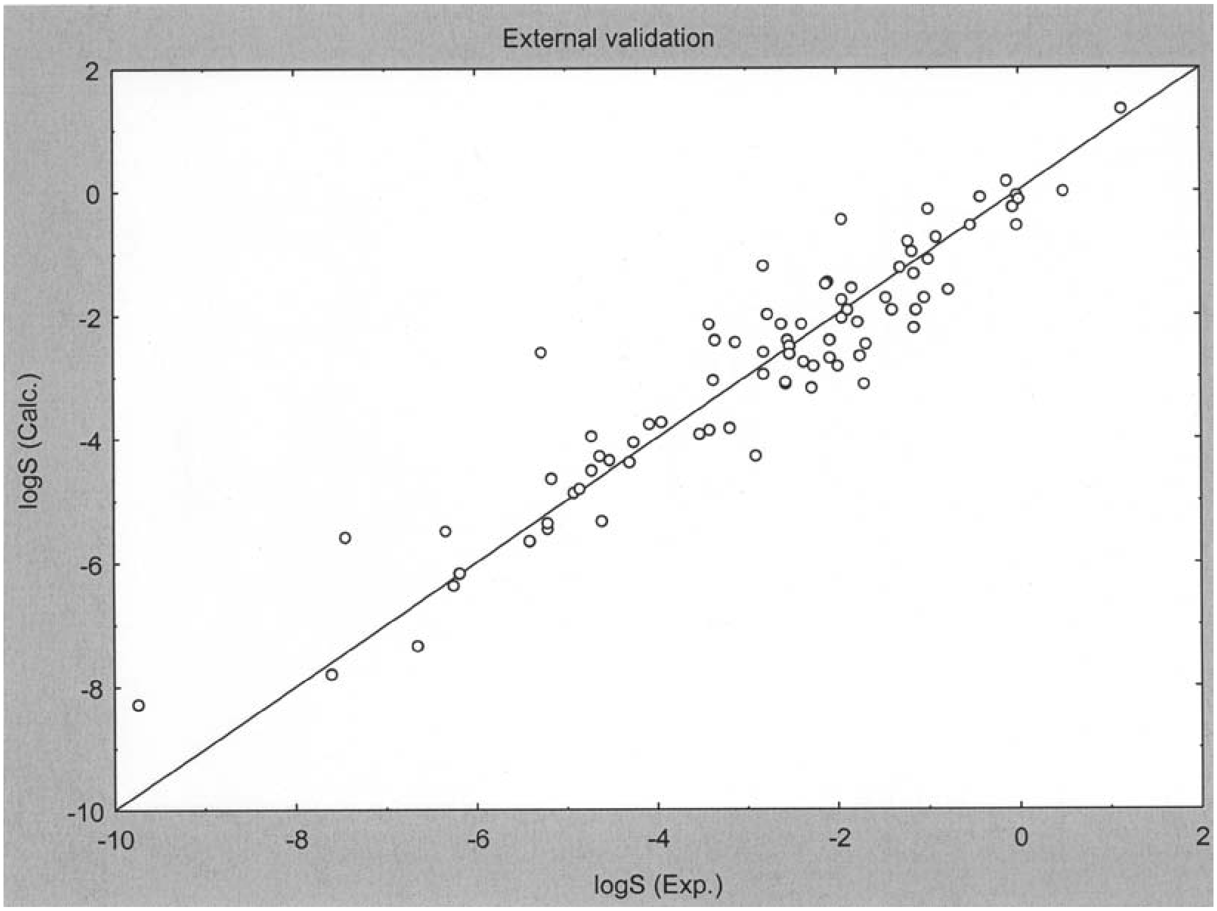

Fig. (2). Results of the ***final external validation set*** prediction.

Appendix Predicted and Experimental Water Solubility Data of the Final External Validation Set

No.	Name	$logS_{exp}$	$logS_{calc}$ (ANN)	ΔlogS (ANN)	$logS_{calc}$ (ALOGPS)	ΔlogS (ALOGPS)
1	m-Nitrobenzoic acid	-1.68	-2.13	-0.45	-2.37	-0.69
2	Indole	-1.21	-0.64	0.57	-1.35	-0.14
3	Phenobarbital	-2.29	-2.99	-0.70	-2.91	-0.62
4	Tetrahydrofuran	0.49	-0.10	-0.59	-0.02	-0.51
5	Procaine	-1.78	-2.55	-0.77	-1.53	0.25
6	Benzonitrile	-1.00	-1.11	-0.11	-1.05	-0.05
7	Pyrene	-6.18	-6.17	0.01	-6.95	-0.77
8	2-Amino-1-naphtalenesulfonic acid	-1.70	-3.14	-1.44	-2.27	-0.57
9	9H-Carbazole	-5.27	-2.84	2.43	-3.40	1.87
10	2,4-Dinitrotoluene	-2.82	-2.20	0.62	-3.27	-0.45
11	4-Heptanone	-1.30	-1.14	0.16	-1.40	-0.10
12	1,2-Diethoxy-ethane	-0.77	-1.38	-0.61	-0.60	0.17
13	3,4-Dichloro-biphenyl	-7.44	-5.58	1.86	-5.58	1.86
14	2,3,4'-PCB	-6.26	-6.30	-0.04	-6.31	-0.05
15	1,1,2,2-Tetrachloroethane	-1.74	-2.70	-0.96	-2.19	-0.45
16	1,3-Dimethylnaphthalene	-4.29	-4.66	-0.37	-4.62	-0.33
17	1-Hexyne	-2.36	-2.62	-0.26	-2.71	-0.35
18	1-Methylbenz[a]anthracene	-6.64	-7.28	-0.64	-7.96	-1.32
19	2,3-Dimethylnaphthalene	-4.72	-4.72	0.00	-4.58	0.14
20	2,4-Dimethylpentane	-4.26	-3.63	0.63	-3.60	0.66
21	2,6-Dichloro-1,1'-biphenyl	-5.21	-5.41	-0.20	-5.66	-0.45
22	2-Chloro-1,1,1-trifluoroethane	-1.15	-2.23	-1.08	-0.83	0.32
23	2-Chloro-1-nitrobenzene	-2.55	-2.24	0.31	-2.83	-0.28
24	2-Iodopropane	-2.09	-2.71	-0.62	-1.97	0.12
25	2-Nonanone	-2.57	-3.17	-0.60	-2.89	-0.32
26	3-Hexyne	-1.99	-2.72	-0.73	-2.20	-0.21
27	3-Methylheptane	-5.16	-4.10	1.06	-4.40	0.76
28	3-Methylthiophene	-2.39	-2.19	0.20	-1.97	0.42
29	Anthracene	-6.35	-5.57	0.78	-5.57	0.78
30	Benzo[e]pyrene	-7.60	-7.84	-0.24	-8.36	-0.76
31	Butyraldehyde	-0.01	-0.10	-0.09	-0.13	-0.12
32	Decanal	-3.41	-4.05	-0.64	-4.29	-0.88
33	Hexabromobenzene	-9.74	-7.51	2.23	-6.31	3.43
34	Hexachloro-1,3-butadiene	-4.92	-4.90	0.02	-5.61	-0.69
35	Isobutyl formate	-1.00	-0.38	0.62	-0.61	0.39
36	Isobutyl methyl ether	-0.90	-1.04	-0.14	-0.76	0.14
37	n-Butyl propionate	-1.94	-1.64	0.30	-1.61	0.33
38	n-Heptane	-4.53	-3.85	0.68	-3.98	0.55

(Appendix). contd.....

No.	Name	$logS_{exp}$	$logS_{calc}$ (ANN)	ΔlogS (ANN)	$logS_{calc}$ (ALOGPS)	ΔlogS (ALOGPS)
39	n-Hexylbenzene	-5.21	-5.48	-0.27	-5.22	-0.01
40	N,N-Dimethyl formamide	1.14	1.19	0.05	1.02	-0.12
41	p-Xylene	-2.82	-3.19	-0.37	-2.73	0.09
42	Trichlorofluoromethane	-2.10	-1.58	0.52	-1.75	0.35
43	Pentobarbital	-2.52	-2.73	-0.21	-2.42	0.10
44	1H-Isoindole-1,3(2H)-dione	-2.61	-1.91	0.70	-1.58	1.03
45	2,5-Pyridinedicarboxylicacid	-2.13	-1.37	0.76	-1.70	0.43
46	2,2'-Biquinoline	-5.40	-6.51	-1.11	-5.02	0.38
47	2-Methyl-2-propenenitrile	-0.42	-0.29	0.13	-0.47	-0.05
48	2-Phenyl-4-carboxyquinoline	-3.19	-3.81	-0.62	-3.85	-0.66
49	N'-(3,4-dichlorophenyl)-N-methoxy-N-methyl urea	-3.52	-3.70	-0.18	-3.41	0.11
50	2,4-Pyridinedicarboxylic acid	-1.83	-1.45	0.38	-1.67	0.16
51	1-Nitropentane	-1.95	-1.33	0.62	-1.66	0.29
52	(3-i-Propyl-5-methyl)-phenyl-N-methyl carbamate	-3.35	-2.63	0.72	-3.13	0.22
53	3-Hydroxy-5-methylisoxazole	-0.07	0.19	0.26	-0.36	-0.29
54	1-Methyl-1-phenylethyl hydroperoxide	-1.04	-2.27	-1.23	-2.04	-1.00
55	Tetrahydro-2H-pyran	-0.03	-0.56	-0.53	-0.34	-0.31
56	2-Butenoic acid	0.00	0.13	0.13	0.12	0.12
57	Malathion	-3.36	-4.25	-0.89	-3.30	0.06
58	Phenolphthalein	-2.90	-4.11	-1.21	-4.48	-1.58
59	1-Anthranol	-4.73	-3.52	1.21	-4.45	0.28
60	Bibenzyl	-4.62	-5.62	-1.00	-5.17	-0.55
61	Gallic acid	-1.16	-1.02	0.14	-1.54	-0.38
62	Isopentanol	-0.52	-0.30	0.22	-0.34	0.18
63	Monolinuron	-2.57	-3.06	-0.49	-2.68	-0.11
64	o-Chlorobenzoic acid	-1.89	-1.80	0.09	-2.45	-0.56
65	Phthalic anhydride	-1.39	-1.02	0.37	-1.47	-0.08
66	Sulfamethazine	-2.27	-2.90	-0.63	-3.07	-0.8
67	Tubercidin	-1.95	-2.01	-0.06	-1.20	0.75
68	1,1,3-Trimethyl cyclohexane	-4.85	-4.66	0.19	-4.57	0.28
69	Diallate	-4.08	-4.34	-0.26	-4.00	0.08
70	Ethiofencarb	-2.09	-2.65	-0.56	-3.09	-1.00
71	6-Chlorpteridine	-1.12	-1.72	-0.60	-0.90	0.22
72	4-Hydroxypteridine	-1.47	-2.16	-0.69	-1.19	0.28
73	Alclofenac	-3.13	-2.89	0.24	-2.89	0.24
74	Deoxycorticosterone acetate	-4.63	-5.12	-0.49	-4.88	-0.25
75	5,5-Di-isopropyl barbiturate	-2.77	-2.15	0.62	-2.22	0.55
76	5-Ethyl-5-octyl barbiturate	-3.94	-3.80	0.14	-3.71	0.23

(Appendix). contd.....

No.	Name	$\log S_{exp}$	$\log S_{calc}$ (ANN)	$\Delta\log S$ (ANN)	$\log S_{calc}$ (ALOGPS)	$\Delta\log S$ (ALOGPS)
77	2,6-Dimethylaniline	-1.17	-1.01	0.16	-1.36	-0.19
78	2-Chlorotoluene	-2.53	-2.63	-0.10	-2.91	-0.38
79	Methane	-2.82	-1.79	1.03	0.85	3.67
80	Triethylamine	-0.14	0.08	0.22	0.11	0.25
81	Isoguanine	-3.40	-2.97	0.43	-1.35	2.05

In this article we have summarized the most important methods used to predict the aqueous solubility of drug-like compounds followed by a presentation of our solubility calculating method(s). We have built a model showing satisfactory predictive power, comparable to the ones published in the literature. We have used a chemically diverse set of compounds in model generation, which resulted in good generalization ability. This feature is extremely important in virtual screening of large, structurally diverse compound libraries.

The calculation of all descriptors for 1000 molecules takes about 30 minutes while the prediction of solubility data of the compounds takes only seconds with our method.

ABBREVIATIONS

QSPR = Quantitative Structure-Property Relationships

MLR = Multiple Linear Regression

PLS = Partial Least Squares method

ANN = Artificial Neural Network

ADME/Tox = Absorption, Distribution, Metabolism, Excretion, Toxicity related properties

HTS = High Throughput Screening

VSS = Variable Subset Selection

GA = Genetic Algorithm

SSA = Sequential Selection Algorithm

Q^2 = Cross validated squared correlation coefficient

SEP = Standard Error of Prediction.

ACKNOWLEDGEMENT

This work was supported by the research grants of the Hungarian Science Foundation, OTKA No. T-32415 and T-029598, NKFP-1/041 and NKFP 1/A/0020/2002.

REFERENCES

[1] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rew.* **1997**, *23*, 3.

[2] Yalkowsky, S. H.; Banerjee, S. Aqueous solubility: methods of estimation for organic compounds. Dekker: New York, **1992**.

[3] Yalkowsky, S. H.; Valvani, S. C. *J. Pharm. Sci.* **1980**, *69*. 912.

[4] Ran, Y.; Jain, N.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 1208.

[5] Ran, Y.; He, Y; Yang, G.; Johnson, J. L. H.; Yalkowsky, S. H. *Chemosphere* **2002**, *48*, 487.

[6] Isnard, P.; Lambert, S. *Chemosphere* **1989**, *18*, 1837.

[7] Warne, M.; Connel, D. W.; Hawker, D. W.; Schhhrmann, G. *Chemosphere* **1990**, *21*, 877.

[8] Miller, M.; Ghogbane, S.; Waski, S. P.; Tewari, T. B.; Nartire, D. E. *J. Chem. Eng. Data* **1984**, *29*, 184.

[9] Yaws, C.; Pan, X.; Lin, X. *Chem. Eng.* **1993**, *Feb*, 108.

[10] Ruelle. P.; Kesselring, U. *Chemosphere* **1997**, *2*, 275.

[11] Wakita, K.; Yoshimota; Miyamoto; Watanable *Chem. Pharm. Bull.* **1986**, *34*, 4663.

[12] Suzuki, T. *J. Comput.-Aid. Mol. Design* **1991**, *5*, 149.

[13] Klopman, G.; Wang, S.; Balthasar, D. M. *J. Chem Inf. Conput. Sci.* **1992**, *32*, 474.

[14] Klopman, G.; Zhu, H. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 439.

[15] Kühne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schüürmann, G. *Chemosphere* **1995**, *30*, 2061.

[16] Medir, M.; Giralt, F. *AIChE J.* **1982**, *28*, 341.

[17] Nirmalakhandan, N. N.; Speece, R. E. *Environ. Sci. Technol.* **1988**, *22*, 328.

[18] Nirmalakhandan, N. N.; Speece, R. E. *Environ. Sci. Technol.* **1989**, *23*, 708.

[19] Patil, G. S. *Chemosphere* **1991**, *22*, 723.

[20] Makino, M. *Environ. Intl.* **1998**, *24*, 653.

[21] Huibers, P. D. T.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283.

[22] Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720.

[23] Huuskonen, J.; Marja Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 450.

[24] Sutter, J. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100.

[25] Mitchell, B. E.; Jurs, P. C.; *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489.

[26] Engkvist, O.; Wrede, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247.

[27] McElroy, N. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237.

[28] McFarland, J. W. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355.

[29] Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177.

[30] Liu, R.; So. S.-S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633.

[31] Yin, C.; Liu, X.; Guo, W.; Lin, T.; Wang, X.; Wang L. *Water Research* **2002**, *36*, 2975.

[32] Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**. *10*, 1155.

[33] Collette, T. W. *Vibrational Spectroscopy* **1997**, *15*, 113.

[34] Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 773.

[35] Chen, X. Q.; Cho, S. J.; Li, Y.; Venkatesh, S. *J. Pharm. Sc.* **2002**, *91*, 1838.

[36] Bergstrom, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. *Pharm. Res.* **2002**, *19*, 182.

[37] Delgado, E. J. *Fluid Phase Equilibria*, **2002**, *199*, 101.

[38] Jorgensen, W. L.; Duffy, E. M. *Adv. Drug Del. Rev.* **2002**, *54*, 355.

[39] Huuskonen, J.; Rantanen, J.; Livingstone, D. *Eur. J. Med. Chem.* **2000**, *35*, 1081.

[40] Wanchana, S.; Yamashita, F.; Hashida, M. *Pharmazine* **2002**, *2*, 127.

[41] Abraham, M. H.; Joelle, J. *J. Pharm. Sci.* **1999**, *88*, 868.

[42] Bruneau, P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605.

[43] Isis^{TM}/Base, Ver. 2.3, MDL Information Systems, Inc., **1990-2000**.

[44] Ruskino, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. Univ. of Texas, Austin, TX and Tripos, St. Louis, MO USA

1988, *CONCORD 6.0, 1992, TRIPOS Associates Inc., St. Louis, Missouri.*

[45] 3DNET, Vichem Ltd., Budapest, Hungary **1999**.

[46] Connolly, M. L. *J. Am. Chem. Soc*. **1985,** *107*, 1118.

[47] De Bruijn, J; Hermkens, J. *J. Quant. Struct-Act. Relat*. **1990**, *9*, 11.

[48] Bodor, N.; Buchwald, P. *J. Phys. Chem.* **1997**, *101*, 3404.

[49] Iwase, K; Komatau, K.; Hirono, S.; Nakagawa, S.; Moriguchi, I. *Chem. Pharm. Bull*. **1985,** *33*, 2114.

[50] Meyer, A. Y. *J. Chem. Soc. Rev*. **1986,** *15*, 449.

[51] Todeschini, R.; Grammatica, P. *Quant. Struct.-Act. Relat*. **1997**, *16*, 120.

[52] Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model*. **1997**, *3*, 142.

[53] Miller, K. J. *J. Am. Chem. Soc*. **1990**, *112*, 8533.

[54] Mortier, W. J.; Genechten, K.; Gasteiger, J. *J. Am. Chem. Soc*. **1985**, *107*, 829.

[55] Fedors, R. F.; Van Krevelen, D. V.; Hoftyzer, P. J. *CRC Handbook of Solubility Parameters and Other Cohesion Parameters*, CRC Press: New York, **1986**.

[56] Andrews, P. R.; Craik, D. J., Martin, J. L. *J. Med. Chem*. **1984**, *27*, 1648.

[57] Wiener, H. *J. Am. Chem. Soc*. **1947**, *69*, 2636.

[58] Randic, M. *J. Am.Chem. Soc*. **1975**, *97*, 6609.

[59] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. J. *Chem. Inf. Comput. Sci*. **1998**, *38*, 28.

[60] Bodor, N.; Huang, M. J.; Harget, A. *J. Mol. Struct. (Theochem.)* **1994**, *309*, 259.

[61] Gaillard, P.; Carrupt, P.; Testa, B.; Boudon, A. *J. Comput.-Aided Mol. Des.* **1994**, *8,* 83.

[62] Cronce, T. D.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. *J. Chem. Soc. Perkin Trans*. **1998**, *2*, 1293.

[63] Eros, D.; Kövesdi, I.; Orfi, L.; Takács-Novák, K.; Acsády, Gy.; Kéri, Gy. *Curr. Med. Chem.* **2002**, *9*, 1819.

[64] 3DNET4W, Vichem Ltd., Budapest, Hungary **2002**.

[65] Snee, R. D. *Technometrics* **1997**, 415.

[66] http://146.107.217.178/lab/alogps/start.html.