Estimation of Aqueous Solubility of Organic Compounds with QSPR Approach

Hua Gao,[1,2] Veerabahu Shanmugasundaram,[1] and Pil Lee[1]

Received November 12, 2001; accepted December 6, 2001

Purpose. To derive a QSPR model for estimation of aqueous solubility of organic compounds.

Methods. Solubility data for 930 diverse compounds was investigated with principal component regression analysis. This set of compounds consists of pharmaceuticals, pollutants, nutrients, herbicides, and pesticides. The diversity of this collection was analyzed using MACCS fingerprint and BCUT chemistry space.

Results. The training set of the solubility data is as diverse as the Available Chemicals Directory, and more diverse than the MDL Drug Data Report. Forty-six molecular descriptors were screened using a genetic algorithm. A QSPR model with a squared correlation coefficient (r^2) of 0.92, a root mean square error of 0.53 log molar solubility (log S_w), an average absolute estimation error of 0.36 log S_w, and a cross-validated q^2 of 0.91 was derived. The QSPR model was validated with a test set of 249 compounds not included in the training set. The absolute estimation error for the test set of compounds was 0.39 log S_w.

Conclusions. A highly predictive QSPR model for estimating aqueous solubility was derived and validated. This model can be used to estimate aqueous solubility for virtual screening and combinatorial library design.

KEY WORDS: solubility; Genetic algorithm; diversity; principal component regression; AquaSol database.

INTRODUCTION

Virtual screening and combinatorial chemistry are firmly established powerful techniques in drug discovery efforts, particularly in lead discovery and optimization. Incorporation of medicinal chemistry knowledge (drug-likeness, SAR, and QSAR information) and biopharmaceutical properties into library design is a prerequisite to rational drug design (1). In recent years, various computational methods have been developed to filter and select sublibraries with relevant physicochemical and biopharmaceutical properties such as lipophilicity, polar surface area, hydrogen bond numbers, and aqueous solubility (1–3).

Since the pioneering work of Hansch and co-workers (4) correlating aqueous solubility with octanol/water partition coefficient, numerous methods for estimating aqueous solubility have been developed (5–11). These can be grouped into three classes based on the type of physicochemical properties and molecular descriptors used in the analysis: (i) based on experimentally determined physicochemical properties such as partition coefficient, melting point, etc (6); (ii) based on group-contribution theory (9); (iii) based on calculated molecular descriptors such as cLog P, molecular surface area, topological indices, etc (10,11). We think that the model(s) based on calculated molecular descriptors is most suitable for general virtual screening and library design purpose. Most of the earlier models were derived from simple, monofunctional, environmentally interesting compounds. These models performed poorly in estimating the solubility of drug-like molecules with multifunctional molecular structures (2). Other models derived from structural analogs do not have broad applicability for virtual screening.

The structural diversity of compounds in a training set used to derive a model determines whether the derived model can be broadly applicable to different classes of compounds and thus suitable for general virtual screening and library design applications. In our ongoing work to develop a more general predictive solubility model(s) for virtual screening and library design we have started with a structurally diverse set of compounds, mostly drug-like molecules, and a set of calculated molecular descriptors to derive a QSPR model for the estimation of aqueous solubility of organic compounds. In this report, the solubility data for a diverse set of compounds gathered from the literature and in-house sources was analyzed using principal component regression based on a set of calculated molecular descriptors.

METHODS

Aqueous Solubility

The aqueous solubility data consist of a subset from AquaSol database (12) and the literature (13–22), and in-house measured solubility data. The AquaSol database contains large amounts of solubility records extracted from a number of scientific references. It covers a variety of compounds including pharmaceuticals, pollutants, nutrients, herbicides, pesticides, agricultural and industrial compounds (12). Most of the compounds in AquaSol database are simple, nonfunctional chemicals or environmentally interesting compounds. To increase the representation of drug-like molecules in our analysis, we included our in-house measured solubility and solubility data collected from the literature for drugs and drug-like molecules. All solubility data were converted to log(molarity), log S_w. The aqueous solubility (log S_w) used in this study ranges from –11.62 to 4.75. A set of 930 compounds was selected as a training set and a set of 249 compounds as a test set.

Molecular Descriptors

All molecular descriptors used in this investigation were implemented and calculated with MOE software (23).

Diversity Analysis

Using MACCS Fingerprint

Molecular diversity was analyzed using average pairwise *Tanimoto* coefficient (T_c) for the training set of the solubility data as reported in our previous publication (24). The average T_c values were calculated using the MACCS fingerprints

[1] Computer-Aided Drug Discovery, Pharmacia, 301 Henrietta Street, Kalamazoo, Michigan 49007

[2] To whom correspondence should be addressed. (e-mail: hua.gao@pharmacia.com)

implemented in MOE (23). The MACCS keys are bit string representations of structures, where each bit refers to the presence or absence of a unique substructural pattern. The *Tanimoto* coefficient for two molecules 1 and 2 was calculated as $T_c = B_c/(B_1 + B_2 - B_c)$, where B_c is the number of common bits set, B_1 and B_2 are the bits set in the fingerprints of molecules 1 and 2, respectively.

The Available Chemicals Directory (ACD) (25) and MDL Drug Data Report (MDDR) (26) were used as reference databases in molecular diversity analysis. A total of 85,949 compounds with molecular weight less than 700 and containing no metals, Si, and B elements were extracted from the MDDR. A subset of 18,683 representative compounds was selected from the ACD using a cell-based selection algorithm implemented in DiverseSolutions (27). A set of 138 estradiol analogs was collected from the literature as a reference of a congeneric set of compounds (28).

Using BCUT Chemistry Space

Pearlman and Smith have developed a novel cell-based, low-dimensional chemistry space representation algorithm (partitioning chemistry space into hypercubic cells) that enables reference to both inter-compound distances and absolute position of compounds in chemistry space using BCUT descriptors (27). Using the chi-squared algorithm implemented in DiverseSolutions a six-dimensional chemistry space (see Table I), which best represents the diversity contained in the ACD database was defined. Pearlman and Smith recommend a resolution (number of bins per axis), which yields roughly 12% to 16% occupancy (fraction of non-empty cells). Hence, the chemistry space was equally partitioned into 6 bins along each co-ordinate. The partitioned chemistry space thus contains $6^6 = 46{,}656$ cells.

Two cell-based measures can be used to characterize *the intrinsic diversity of a compound collection in chemistry space*: 1) P_{cell} (Percentage of cells occupied), $P_{cell} = (N_{occ.cells}/N_{cells}) \times 100$, where $N_{occ.cells}$ is the number of occupied cells, N_{cells} is the total number of cells in the chemistry space; 2) P_{eff} (Effectiveness of coverage of chemistry space by occupied cells), $P_{eff} = (N_{occ.cells}/N_{cpds}) \times 100$, where N_{cpds} is the total number of compounds in the collection. P_{cell} gives a measure of the coverage of chemistry space, and P_{eff} yields a measure of the effectiveness of this coverage.

QSPR Analysis

Principal Component Regression (PCR) Analysis

The PCR algorithm implemented in MOE was used for the QSPR analysis. In this study, variable selection was achieved using a genetic algorithm developed in our group (24).

Table I. BCUT Descriptors Used to Define the ACD Chemistry Space

Symbol	Definition
BCUT-CHRG1	BCUT_gastchrg_S_invdist6_0.60_R_H
BCUT-CHRG2	BCUT_gastchrg_S_invdist_1.50_R_L
BCUT-HA	BCUT_haccept_S_invdist_0.60_R_H
BCUT-HD	BCUT_hdonor_S_invdist_0.45_R_H
BCUT-POL1	BCUT_tabpolar_S_invdist6_1.25_R_L
BCUT-POL2	BCUT_tabpolar_S_invdist_0.50_R_H

Neural Network Analysis

The neural network analysis was carried out using NeuralWare software (29). A three-layered, fully connected neural network was trained by the standard back-propagation learning algorithm with a sigmoid activation function for hidden nodes. The input and output values were scaled between 0.1 and 0.9, and adjustable weights between neurons were given random values between –0.5 and 0.5. A number of 50,000 learning cycles was used.

RESULTS AND DISCUSSION

Diversity Analysis

The calculated average *T*c values for ACD, MDDR, and the 138 estradiol analogs are 0.25 ± 0.12 (average ± SD), 0.39 ± 0.11, and 0.78 ± 0.12, respectively, although the training set in our study has an average *T*c value of 0.21 ± 0.17. Based on the average *T*c values, the solubility dataset is as diverse as the ACD, and more diverse than the MDDR. Because the MACCS type of fingerprint represents the presence and absence of different substructural and functional groups in a given molecule, the relatively low average *T*c value of the solubility dataset indicates it contains very diverse structural and functional features.

In the BCUT chemistry space analysis, the ACD (total of 144,684 compounds), which is considered as a good representation of diverse compounds occupies 5,498 of 46,656 cells of the chemistry space. Thus, the percentage of cells occupied by the ACD compounds is 11.8%. The effectiveness of coverage is 3.8%. The solubility dataset of compounds occupies 382 of the 46,656 cells. The P_{cell} for the solubility set is 0.8% and the P_{eff} is 41.1%. In other words, the solubility dataset covers 0.8% of the chemistry space with a much higher effectiveness of 41.1%. The higher effectiveness of the solubility dataset shows that it is 'intrinsically diverse'. The coverage of chemistry space is understandably low because of the small number of compounds (930) attempting to cover a chemistry space binned into 46,656 cells. Fig 1. shows that the solubility dataset spans the entire ACD chemistry space. The overall results from the diversity analysis indicate that the solubility dataset has a high structural diversity and covers a very broad range of chemistry space.

QSPR Analysis

The genetic algorithm used in the variable selection had an initial population of 120 chromosomes, a good population of 20 chromosomes, a uniform crossover rate of 0.5, and a mutation rate of 0.05. With this configuration of the genetic algorithm, a set of 24 out of 46 molecular descriptors was selected as a preferred set of descriptors for the training set of compounds. The 24 molecular descriptors used in the model and their definitions are listed in Table II. Fig. 2 and Fig. 3 show the squared correlation coefficient increases and the root mean square error (RMSE) decreases with the increase in the number of principal components used.

Equation (1) is the estimated normalized linear equation using 24 principal components. Ninety-two percent (r^2 =

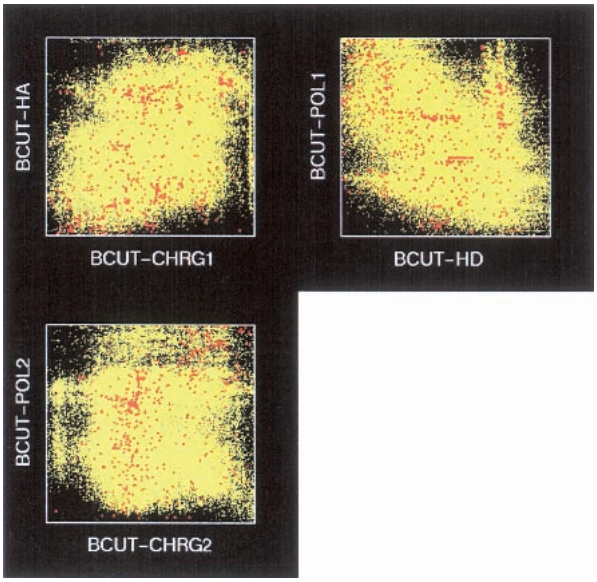


Fig. 1. Two-dimensional projections of the six-dimensional ACD chemistry space (yellow dots: ACD compounds; red dots: training set of compounds).

0.92) of the variance can be explained with Eq. (1). In this equation, n is the number of data points, r the correlation coefficient, RMSE, q the cross-validated correlation coefficient, and AAEE the average absolute estimation error.

Table II. Molecular Descriptors Used in the QSPR Model

Symbol	Definition[a]
b_ar	Number of aromatic bonds
log P	Log partition coefficient
SV_6	bin 5 Slog P
SV_8	bin 7 Slog P
SV_9	bin 8 Slog P
VDM	vertex distance magnitude index
vsa_hyd	VDW hydrophobic surface area
$^0\chi^v c$	zero order carbon valence connectivity index
$^1\chi c$	first order carbon valence connectivity index
$^1\kappa^\alpha$	first alpha modified shape index
$^2\kappa^\alpha$	second alpha modified shape index
k_{aaCH}	Kier E-state for ---CH--
k_{aasC}	Kier E-state for ---C-- (|)
k_{aaaC}	Kier E-state for ---C-- (|)
k_{sNH2}	Kier E-state for $-NH_2$
k_{ssNH}	Kier E-state for —N(H)—
k_{aaN}	Kier E-state for ---N---
k_{sssN}	Kier E-state for N(|)
k_{ddsN}	Kier E-state for N(=)
k_{sOH}	Kier E-state for –OH
k_{dO}	Kier E-state for =O
k_{ssO}	Kier E-state for —O—
k_{sF}	Kier E-state for –F
k_{dssS}	Kier E-state for —S(=)—

[a] ---, aromatic bond; -, single bond; =, double bond.

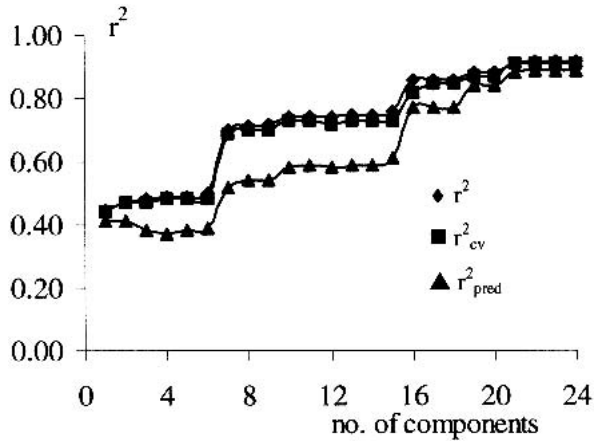


Fig. 2. Plot of squared correlation coefficients vs. number of principal components (r^2: non-cross-validated; r^2_{cv}: cross-validated; r^2_{pred}: test set).

$$\log S_w = 2.34 - 0.41 \log P - 0.24\ b\_ar + 0.35\ {}^0\chi^v_C - 0.31\ {}^1\chi_C - 0.83\ {}^1\kappa^\alpha + 0.34\ {}^2\kappa^\alpha \quad \text{Eq. (1)}$$

$$+ 0.12\ SV_6 - 0.10\ SV_8 - 0.17\ SV_9 - 0.79\ VDM + 0.44\ vsa\_h + 0.12\ k_{aaCH}$$

$$+ 0.05\ k_{aasC} - 0.05\ k_{aaaC} + 0.03\ k_{sNH2} + 0.06\ k_{ssNH} + 0.10\ k_{aaN} - 0.06\ k_{sssN}$$

$$+ 0.07\ k_{ddsN} + 0.23\ k_{sOH} + 0.27\ k_{dO} + 0.08\ k_{ssO} + 0.07\ k_{sF} + 0.02\ k_{ssS}$$

$n = 930,\ r^2 = 0.92,\ q^2 = 0.91,\ RMSE = 0.53,\ AAEE = 0.36$

Further analysis indicates that the 24 molecular descriptors used are orthogonal except some covariance between $^0\chi^v_C$ and $^1\chi_C$ ($r^2 = 0.80$), $^0\chi^v_C$ and VDM ($r^2 = 0.70$), $^0\chi^v_C$ and vsa_h ($r^2 = 0.70$), $^1\kappa^\alpha$ and $^2\kappa^\alpha$ ($r^2 = 0.70$), $^1\kappa^\alpha$ and vsa_h ($r^2 = 0.70$), b_ar and k_{aaCH} ($r^2 = 0.70$). The model with 24 principal components was selected as the solubility estimation model. The observed and calculated solubility data of the training set are plotted in Fig. 4. The orthogonality of the 24 molecular descriptors explains why a high number of principal components are needed in Eq. (1). Due to the intrinsic diversity of the training set, to capture the underlying struc-

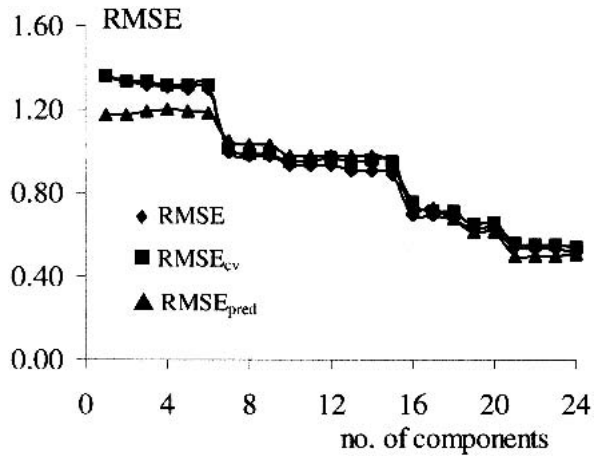


Fig. 3. Plot of RMSE vs. number of principal components (RMSE: non-cross-validated; $RMSE_{cv}$: cross-validated; $RMSE_{pred}$: test set).

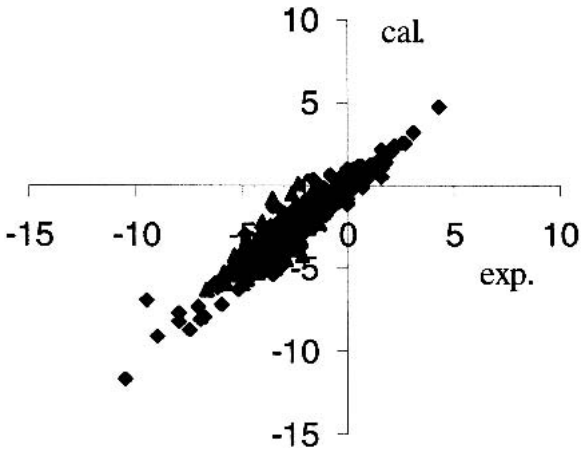


Fig. 4. Plot of experimental (exp.) and calculated (cal.) solubility data of the training set of compounds.

tural features of structure-solubility relationship, many molecular descriptors including Kier's indices had to be used in the correlation. The aqueous solubility of a compound depends on three factors: 1) the entropy of mixing; 2) the difference between the solute-water adhesive interactions and the sum of the solute-solute and water-water cohesive interactions; and 3) the solute-solute interactions in the crystal lattice of crystalline solutes (6). Increase in hydrogen bonding and polarity of solutes generally enhances aqueous solubility. However, for solid compounds, the increase in hydrogen bonding and polarity could also contribute to crystal lattice stability, thus decreasing aqueous solubility. Considering the complicated nature, interpreting the contributions of molecular descriptors to both solute-water interaction in the aqueous phase and solute-solute interaction in the crystal lattice is difficult. Nevertheless, it is apparent from Eq. (1) that solubility decreases with increasing hydrophobicity (negative correlation with log P). The effects of molecular shape on solubility are contained in the connectivity and shape indices. All E-state indices for heteroatoms capable of forming hydrogen bond positively contribute to the aqueous solubility. Our objective of the study is to develop a model for estimating solubility. A detailed interpretation of structure-solubility relationship, and the derived PCR model is beyond the scope of this report.

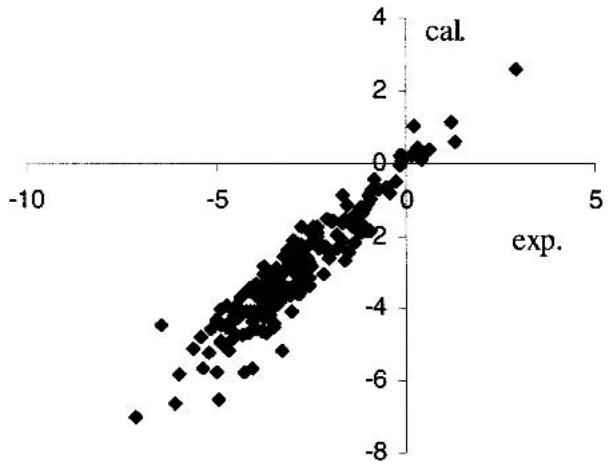


Fig. 5. Plot of experimental (exp.) and calculated (cal.) solubility data of the test set of compounds.

Table III. Summary of Solubility Prediction

Compounds		Range of AAEE 0 ~ 0.5	0.5 ~ 1	1 ~ 2	2 ~ 3
Training set	No. of compounds	684	183	56	7
	% of total	74	20	6	0.7
Test set	No. of compounds	143	59	7	
	% of total	68	28	3	

It has been pointed out that there are some nonlinear dependencies between some molecular descriptors and the aqueous solubility (log S_w) (10). Therefore, the solubility data of the training set was also analyzed with neural network to detect the possible nonlinear dependencies. It turned out that the neural network model (data not shown) was not as good as the derived PCR model. Also, the results suggest that there are no significant nonlinear dependencies between the molecular descriptors used and the aqueous solubility analyzed.

Validation of the QSPR Model

To evaluate the predictive ability of the derived QSPR model, the solubilities for a set of 249 compounds not included in the training set were calculated from the model. The estimation has a r^2_{pred} of 0.91, a RMSE of 0.49 log S_w, and an AAEE of 0.39 log S_w. The experimental and calculated values for the set are plotted in Fig. 5. The solubility data from the literature is summarized in APPENDIX I. The validation result is consistent with the QSPR model. The test set of compounds covers very diverse structures, ranging from simple chemicals to pharmaceuticals with complex structures such as HIV protease inhibitors. The AAEE for 7 HIV protease inhibitors is 0.41 log S_w. The results indicate that the derived QSPR model is highly stable and predictive, thus may have broad applicability of solubility estimation for many classes of compounds.

The prediction results of the training set and the test set of compounds are summarized in Table III. For the training set, 74% of solubility data was estimated within an error of 0.5 log S_w, 94% within an error of 1 log S_w. In the case of the test set, 69% of data was predicted within an error of 0.5 log S_w, and 96% within an error of 1 log S_w. It has been pointed out that the experimental solubility can differ by 1.0 log S_w, especially for compounds with very low solubility (19). The plots in Figs. 4 and 5 also show that the compounds with lower solubilities have larger variance. In our analysis, only two-dimensional (2D) descriptors and descriptors containing implicit three-dimensional (3D) information such as Kier's shape indices were used. Explicit 3D descriptors were not used to avoid bias of the analysis due to predicted conformational effects. In our view, *a priori* solubility estimation using method(s) with a multi-parameter equation like ours should be very useful for 'rank-ordering' of virtual library compounds and prioritization of their synthesis.

CONCLUSIONS

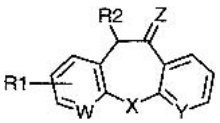
A highly predictive QSPR model was derived based on the calculated molecular descriptors for a diverse set of com-

APPENDIX I. Solubility Data of Compounds from the Literature in the Test Set

compound	log S_w* exp.	log S_w* cal.	compound	log S_w exp.	log S_w cal.
atropine	–2.12	–2.99	nopinen	–3.91	–3.52
acetaminophen	–2.62	–2.79	β-myrcene	–3.58	–2.86
anethole trithione	–5.80	–4.25	N2-acetylacyclovir	–1.92	–1.21
dithiolethione	–2.48	–1.51	O-acetylacyclovir	–0.86	–1.66
didanosine	–0.90	–0.98	o-phenanthroline	–3.34	–3.50
delavirdine	–4.76	–4.34	fenchyl alcohol	–2.27	–2.77
efavirenz	–4.57	–5.17	α-pinene oxide	–2.59	–2.89
indinavir	–3.94	–3.48	α-ionone	–3.06	–3.38
ritonavir	–5.16	–4.70	carveol	–1.88	–2.53
amprenavir	–4.00	–3.81	α-terpineol	–1.91	–2.32
saquinavir	–4.33	–4.98	5-Et-5-iPr-BA	–2.15	–2.36
3-POM-5,5-DH[a]	–4.68	–4.15	5-Et-5-allyl-BA	–1.61	–1.98
3-OOM-5,5-DH[b]	–6.52	–4.95	5-Et-5-(3MBE)-BA	–2.25	–2.41
5-phenyldithiolethione	–5.64	–4.08	5,5-diphenyl-BA	–4.20	–3.24
dimethylthiolethione	–3.42	–3.05	5,5-di-iPr-BA	–2.77	–2.74
5,5-Me_2-BA[c]	–1.74	–1.10	5-iPr-5-(3MBE)-BA	–2.59	–2.77
5-Me-5-allyl-BA	–1.16	–1.58	5-tBu-5-(3MBE)-BA	–3.55	–3.12
5-Me-5-phenyl-BA	–2.38	–2.28	5-allyl-5-phenyl-BA	–2.37	–2.64
5-Me-5-(3MBE)-BA[d]	–2.60	–2.02	5,5-diethyl-2-S-BA	–2.17	–2.29
5,5-$(CH_2)_2$-BA	–1.89	–0.95	5-Et-5-(1MB)-2-S-BA[c]	–3.68	–3.48
5,5-$(CH_2)_3$-BA	–1.66	–1.38	5,5-$(CH_2)_7$-BA	–2.98	–2.93
5,5-$(CH_2)_4$-BA	–2.35	–1.80	5,5-$(CH_2)_{10}$-BA	–4.59	–3.95
5,5-$(CH_2)_5$-BA	–3.06	–2.19	5,5-$(CH_2)_5$-2-S-BA	–3.46	–2.91
5,5-$(CH_2)_{11}$-BA	–5.80	–5.00	5,5-$(CH_2)_6$-BA	–3.17	–2.57
5IDU[f]	–2.25	–1.65	5′-COC_4H_9-5IDU	–3.40	–3.02
5′-COC_2H_5-5IDU	–2.46	–2.47	5′-COtBu-5IDU	–3.34	–3.07
5′-COC_3H_7-2′-5IDU	–2.84	–2.75	5′-COC_6H_5-5IDU	–3.48	–3.09
5′-$COiC_3H_7$-2′-5IDU	–2.76	–2.81	5′-(COC_6H_4-p-NO2)-5IDU	–3.30	–3.61
4-Hydroxypyridine	1.02	0.19	5′-(COC_6H_4-p-OMe)-5IDU	–3.55	–3.11
ethyl cinnamate	–2.31	–2.13	2-pinene	–3.66	–3.27
acetazolamide	–2.49	–1.84	methazolamide	–1.92	–2.54
phenylacetic acid	–1.38	–1.17	2-heptanone	–1.22	–1.19
malathion	–3.36	–3.26	p-aminophenol	–0.83	–0.97
piperonal	–1.63	–1.51	diazion	–3.76	–3.90
acrylamide	0.92	0.79	sorbitol	1.09	1.25
propoxur	–2.04	–2.49	heptanoic acid	–1.59	–1.39
cinnamaldehyde	–1.61	–1.37	undecane	–4.07	–4.02
quinhydrone	0.22	–0.09	salol	–2.43	–3.11
2,4-dimethylquinoline	–3.47	–3.34	ioxynil	–4.40	–4.56
maleic hydrazide	0.38	0.29	metolazone	–4.47	–4.89
2,3-Dichlorophenoxyacetic acid	–2.52	–2.83	3-bromophenyl isothiocyanate	–3.80	–3.69
daminozide	0.22	0.17	propyzamide	–3.34	–3.97
pentamethylmelamine	0.20	–0.14	ethofumesate	–3.46	–3.54
cis-1,2-dimethylcyclohexane	–2.69	–2.95	2,3,6-trichlorophenoxyacetic acid	–3.36	–3.71
thiabendazole	–3.73	–3.74			
allopurinol	–1.02	–0.93	m-fluorobenzoic acid	–1.52	–1.59
p-iodo-benzylisothiocyanates	–4.11	–4.10	2,2-dimethylhex-3-yne	–2.29	–2.30
thymol	–2.22	–2.29	norflurazon	–4.04	–4.13
p-chloroaniline	–1.67	–1.58	triadimefon	–3.05	–2.95
crotonic acid	0.23	0.49	dinicotinic acid	–0.46	–0.81
dimethirimol	–1.77	–2.46	suberic acid	–1.73	–1.42
cortisone	–3.81	–3.35	quinidine	–3.70	–4.35
fonofos	–4.20	–4.02	phoxim	–4.75	–4.44
coumarin	–2.53	–2.27	aspirin	–1.62	–1.94
glucose	0.68	0.94	DMPA	–4.80	–5.16
2-nitrobenzaldehyde	–1.97	–1.83	aspartic acid	1.11	1.20
protoporphyrin	–7.02	–7.15	cystine	–0.52	–0.25
tetrachloromethane	–3.40	–3.33	3-methyl-1-butene	–0.69	–0.52
Shikimic acid	–0.06	–0.06	caffeine	–0.98	–0.96
fructose	0.64	0.65	thiram	–3.90	–3.95
sulfanilic	–1.21	–1.61	3-Methyl-2-pentanone	–0.74	–0.86
sulfamerazine	–3.12	–2.64	thiourea	0.25	0.81

APPENDIX I. Solubility Data of Compounds from the Literature in the Test Set

compound	log S_w* exp.	log S_w* cal.	compound	log S_w exp.	log S_w cal.
phenytoin	–3.90	–3.38	urea	1.02	1.68
3-hexanone	–0.74	–0.72	androstenedione	–4.62	–4.59
5-ethyl-5-heptyl-barbituric acid	–3.77	–3.78	o,p′-DDE	–6.36	–6.80
pyrazon	–2.83	–3.50	N2,O-diacetylacyclovir	–2.70	–1.62
m-tolyl isothiocyanate	–3.30	–2.99	D-mannitol	0.60	1.29
2,4,6-trichlorophenoxyacetic acid	–3.40	–3.73	p-toluenesulfonamide	–1.57	–1.91
norflurazon	–4.75	–4.44	cumene hydroperoxide	–2.36	–2.11
methyl-dibutyl phosphate	–2.86	–2.59	p-(tert-amyl)phenol	–2.52	–3.10
chloropicrin	–3.39	–3.79	3,4,5-trichlorophenoxy acetic acid	–3.40	–3.71
1,1,2-trichloroethane	–1.45	–1.37	3-thenoic acid	–1.12	–0.98
2,4-DB	–3.13	–3.36	2-methylnaphthalene	–3.62	–3.29
methidathion	–3.95	–3.44	m-toluic acid	–1.58	–1.59
3-methylpentane	–1.64	–1.69	iodofenphos	–6.62	–6.11
α-Furole	0.04	–0.10	p-sec-butylphenol	–1.77	–2.32
Acamol	–1.03	–1.35	nicotinic acid	–0.84	–0.42



Substituents W	X	Y	Z	R1	R2	log S_w exp.	log S_w cal.
C	O	C	O	H	Et	–3.68	–3.79
C	N(Et)	N	O	H	Me	–3.32	–3.29
N	N(Et)	N	O	H	Me	–2.62	–2.75
N	N(Et)	N	O	2,4-Me_2	H	–4.55	–3.50
N	N(ePr)	N	O	H	Me	–2.88	–3.16
N	N(Et)	N	S	H	Me	–4.63	–3.80
N	N(c-Bu)	N	O	H	Me	–3.54	–3.61
N	N(Et)	N	O	2-OMe,4-Me	H	–5.15	–3.27
N	N(Et)	N	O	2-Cl	Me	–4.11	–3.76
N	N(Et)	N	O	2-N(Me)EtOH	H	–3.36	–2.57

* exp.: experimental value; cal.: calculated value.
[a] POM: pentanolyoxymethyl.
[b] OOM: octanoloxymethyl.
[c] BA: barbituric acid.
[d] 3MBE: 3-methylbut-2-enyl.
[e] 1MB: 1-methylbutyl
[f] 5IDU: 5-Iodo-2′-deoxyuridine.

pounds. The derived model was validated with the test set of compounds not included in the training set. Because this model is based on a set of calculated molecular descriptors instead of molecular fragments, it may be used in solubility estimation of compounds with new functional groups and new ring systems, and also in virtual screening and library design.

ACKNOWLEDGMENTS

The authors thank Pieter Stouten, Philip Burton, Mark Johnson, Thomas Vidmar, and Gerald Maggiora for their helpful discussions.

REFERENCES

1. A. K. Ghose, V. N. Viswanadhan, and J. J. Wendoloski. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1**:55–68 (1999).
2. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug Deliv. Rev.* **23**:3–25 (1997).
3. D. E. Clark. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **88**: 815–821 (1999).
4. C. Hansch, J. E. Quinlan, and G. L. Lawrence. The linear free energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **33**:347–350 (1968).
5. E. J. Lien, L. L. Lien, and H. Gao. Structure-system-activity relationship analysis of drug disposition. In F. Sanz, J. Giraldo, and F. Manaut (eds.), *QSAR and Molecular Modeling: Concept, Computational tools and Biologic Applications*, Prous Science Publishers, Barcelona, 1995 pp. 94–100.
6. S. H. Yalkowsky and S. C. Valvani. Solubility and partitioning of nonelectrolytes in water. *J. Pharm. Sci.* **69**:912–922 (1980).

7. Y. Ran and S. H. Yalkowsky. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**:354–357 (2001).
8. N. Jain and S. H. Yalkowsky. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **90**: 234–252 (2001).
9. G. Klopman, S. Wang, and D. M. Balthasar. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **32**:474–482 (1992).
10. J. Huuskonen. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**:773–777 (2000).
11. N. Bodor and M. J. Huang. A new method for the estimation of the aqueous solubility of organic compounds. *J. Pharm. Sci.* **81**: 954–959 (1992).
12. S. H. Yalkowsky and R. M. Dannelfelser. *The Arizona Database Of Aqueous Solubility*, College of Pharmacy, University of Arizona, Tucson, Arizona, 1990.
13. M. M. Morelock, L. L. Choi, G. L. Bell, and J. Wright. Estimation and correlation of drug water solubility with pharmacological parameters required for biologic activity. *J. Pharm. Sci.* **83**:948–952 (1994).
14. R. J. Prankerd and R. H. McKeown. Physico-chemical properties of barbituric acid derivatives: iv. Solubilities of 5,5-disubstituted barbituric acids in water. *Int. J. Pharm.* **112**:1–15 (1994).
15. G. C. Williams and P. J. Sinko. Oral absorption of the HIV protease inhibitors: A current update. *Adv. Drug Delvi. Rev.* **39**:211–238 (1999).
16. I. Fichan, C. Larroche, and J. B. Gros. Water solubility, vapor pressure, and activity coefficients of terpenes and terpenoids. *J. Chem. Eng. Data* **44**:56–62 (1999).
17. B. J. Aungst. P-glycoprotein, secretory transport, and other barriers to the oral delivery of anti-HIV drugs. *Adv. Drug Deliv. Rev.* **39**:105–116 (1999).
18. A. Kristl. Estimation of aqueous solubility for some guanine derivatives using partition coefficient and melting temperature. *J. Pharm. Sci.* **88**:109–110 (1999).
19. P. B. Myrdal, A. M. Manka, and S. H. Yalkowsky. AQUAFAC 3: aqueous functional group activity coefficients: application to the estimation of aqueous solubility. *Chemosphere* **30**:1619–1637 (1995).
20. A. Kristl and G. Vesnaver. Thermodynamic investigation of the effect of octanol-water mutual miscibility on the partitioning and solubility of some guanine derivatives. *J. Chem. Soc., Faraday Trans.* **91**:995–998 (1995).
21. A. Kristl. Thermodynamic investigation of the effect of the mutual miscibility of some higher alkanols and water on the partitioning and solubility of some guanine derivatives. *J. Chem. Soc., Faraday Trans.* **92**:1721–1224 (1996).
22. V. J. Stella, S. Martodihardjo, and V. M. Rao. Aqueous solubility and dissolution rate does not adequately predict in vivo performance: a probe utilizing some N- acyloxymethylphenytoin prodrugs. *J. Pharm. Sci.* **88**:775–779 (1999).
23. Chemical Computing Groups, Inc. MOE 2000.02, 1255 University Street, Montreal, Quebec, Canada, H3B 3x3.
24. H. Gao. Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J. Chem. Inf. Comput. Sci.* **41**:402–407 (2001).
25. ACD (Available Chemicals Directory), available from MDL Information Systems Inc., 14600 Catalina St., San Leandro, California 94577.
26. MDL Drug Data Report 99.2; MDL Information System, Inc., 1999.
27. R. S. Pearlman and K. M. Smith. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **9**:339–353 (1998).
28. H. Gao, J. A. Katzenellenbogen, R. Garg, and C. Hansch. Comparative QSAR analysis of estrogen receptor ligands. *Chem. Rev.* **99**:723–744 (1999).
29. NeuralWare, Inc. Technical Publishing Group, Pittsburgh, Pennsylvania 15276.