

ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach

T. J. Hou, K. Xia, W. Zhang, and X. J. Xu*

College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Received August 24, 2003

A novel method for the estimation of aqueous solubility was solely based on simple atom contribution. Each atom in a molecule has its own contribution to aqueous solubility and was developed. Altogether 76 atom types were used to classify atoms with different chemical environments. Moreover, two correction factors, including hydrophobic carbon and square of molecular weight, were used to account for the inter-/intramolecular hydrophobic interactions and bulkiness effect. The contribution coefficients of different atom types and correction factors were generated based on a multiple linear regression using a learning set consisting of 1290 organic compounds. The obtained linear regression model possesses good statistical significance with an overall correlation coefficient (r) of 0.96, a standard deviation (s) of 0.61, and an unsigned mean error (UME) of 0.48. The actual prediction potential of the model was validated through an external test set with 21 pharmaceutically and environmentally interesting compounds. For the test set, a predictive $r = 0.94$, $s = 0.84$, and $UME = 0.52$ were achieved. Comparisons among eight procedures of solubility calculation for those 21 molecules demonstrate that our model bears very good accuracy and is comparable to or even better than most reported techniques based on molecular descriptors. Moreover, we compared the performance of our model to a test set of 120 molecules with a popular group contribution method developed by Klopman et al. For this test set, our model gives a very effective prediction ($r = 0.96$, $s = 0.79$, $UME = 0.57$), which is obviously superior to the predicted results ($r = 0.96$, $s = 0.84$, $UME = 0.70$) given by the Klopman's group contribution approach. Because of the adoption of atoms as the basic units, our addition model does not contain a "missing fragment" problem and thus may be more simple and universal than the group contribution models and can give predictions for any organic molecules. A program, drug-LOGS, had been developed to identify the occurrence of atom types and estimate the aqueous solubility of a molecule.

INTRODUCTION

The solubility ($\log S$) of organic molecules in water should be considered in the design of drugs, because this parameter usually has a significant impact on many ADME-concerned properties of drugs, such as uptake, distribution, transport, and eventually bioavailability. In recent years, the developments of combinatorial chemistry and high-throughput screening (HTS) give us more opportunities to synthesize and give rapid and effective assay to thousands upon thousands of compounds in a very short period.¹ It has been noticed that in order to obtain more drug-like molecules the synthesis of combinatorial libraries tends to result in compounds with suitable lipophilicity and aqueous solubility, than with conventional synthetic strategies. Thus, computational screenings have been suggested and used to select sublibraries with relevant physicochemical properties to the range of known values, such as lipophilicity and solubility, of the orally active drugs. Hence there is a strong interest in fast, reliable, and generally applicable structure-based methods for prediction of aqueous solubility of new drugs before a promising drug candidate has even been synthesized.

Until now, many methods have been proposed for the prediction of solubility.^{2–20} These methods generally consist of multiple linear regression (MLR) or artificial neural

networks (ANN) using various molecular descriptors. These methods can be roughly divided into three categories: (1) experiment-related methods, (2) descriptor-based methods, and (3) group contribution methods. The first class of techniques calculate aqueous solubilities using one or several experimental physicochemical properties^{2–4} such as partition coefficient, melting points, boiling points, or molar volumes. These methods require some experimental values, so they are not applicable to compounds not yet synthesized or isolated. Therefore, these methods only have limited applications. The second class of methods generally uses a diverse set of descriptors such as physicochemical descriptors, molecular properties, and/or collection of relevant structural features, which are correlated to activity by means of various statistical techniques.^{5–14} The methods developed by Huuskonen et al.,^{5,6} EcElroy et al.,⁷ McFarland et al.,⁸ Liu et al.,⁹ Tetko et al.,¹⁰ Engkvist et al.,¹¹ Yan et al.,¹² Butina et al.,¹³ and Wegner et al.¹⁴ belong to this class. The practical superiority of this class of methods is that it does not require the knowledge of any experimental data of the compound because all descriptors needed are calculated directly from a 2D or 3D molecular structure. However, this class of methods bear their inherent deficiencies. First, they usually need many molecular descriptors, which may be difficult to be calculated or can even be obtained using commercial software. For example, in the work of Engkvist et al., the authors used a total of 63 physicochemical and topological

* Corresponding author e-mail: xiaojxu@chem.pku.edu.cn.

descriptors. The dependence of the descriptors calculated from other theoretical models prevents us from estimating the solubility of a molecule using the model from a reference or develop a program to estimate solubility as an automatic fashion. Second, the prediction precision should be affected by the prediction precision of descriptors used in the model. For example, in a log*S* prediction model, partition coefficient (log*P*) is a very important descriptor. If to some molecules the predictions of log*P* are very poor, then the resulting solubility may be poor also. Third, the relationship between the descriptors and the aqueous solubility is usually not straightforward. The third class of methods for prediction of aqueous solubility is based on group contribution.^{15–20} They allow the approximation calculation of solubility by calculating the contribution of relevant substructural units of compounds. The methods proposed by Nirmalakhandan et al.,¹⁵ Suzuki et al.,¹⁶ Kuhne et al.,¹⁷ Lee et al.,¹⁸ and Klopman et al.^{19,20} belong to this category. Among all these methods, only the Klopman's mode is a pure and general group contribution model without using additional experimental parameters. In the newest version of the Klopman's model, the authors used a set of 118 functional groups.²⁰

Group contribution methods may be the most practical means of estimating aqueous solubility. The superiority of this class of methods is that they do not need any descriptors based on other theoretical models. Moreover, this class of methods only needs to count the occurrence of functional groups in a molecule, so they are extremely time-saving. The shortcoming of this approach is also obvious. First, they require a large data set to obtain a contribution of each functional group. Second, they may contain a "missing fragment" problem, which means that if a compound contains a missing fragment which can be defined by the group contribution model, its aqueous solubility cannot be precisely predicted. In our previous work, we proposed a new atom-additive method for calculating octanol/water partition coefficient (log*P*) of organic compounds.²¹ The method, SLOGP v1.0, gives log*P* values by summing the contribution of atom-weighted solvent accessible surface areas and correction factors. Because of the good correlation between log*P* and water solubility and the successful application of the atom additive method to the calculation of log*P* values of organic molecules, we believe that the water solubility of organic compounds can be effectively predicted using simple atom addition methods. So the first aim of this study is to construct a robust predictive model of aqueous solubility only based on simple atom contribution. Certainly, the basic idea of the atom contribution in this paper is indeed not novel, and this procedure has been successfully applied in the estimation of log*P*.^{21,22} Furthermore, we attempt to develop an automatic procedure of log*S* prediction, which can be easily used and extended by other researchers and easily integrated within the suite of ADME prediction developed in our group.^{21,23,24}

METHODS

Data Sets. In the development of the atom contribution model, we worked with the Tetko data set.¹⁰ The aqueous solubility was expressed as log*S*, where *S* is the solubility at a temperature of 20–25 °C in mol/L. The original source of the Tetko data set was based on the date set afforded by Huuskonen et al.⁶ The Huuskonen data set includes 1297

diverse compounds taken from the AQUASOL database of the University of Arizona²⁵ and the PHYSPROP database.²⁶ After the revision of Tetko et al., two repetitions including 284 and 522, one NO-oxide compound (minoxidil), one organo-metal (Sn) compound (cyhexatin), and two inner salts (betaine and cephaloridine) were eliminated. Moreover, in the Tetko data set lindane was indicated twice and was also eliminated from the data set. Finally, the whole date set used for the model development in this paper includes 1290 organic compounds. The data set was converted from the SMILES flat file representation²⁷ to the MACCS/sdf structured data file using the Weblabview program developed by Accelrys.²⁸ All molecules were converted successfully. The molecular geometries of all compounds were fully minimized using a molecular mechanism with a MMFF force field.²⁹ The molecules and the experimental log*S* values are listed in Table A. Table A and the MACCS/sdf database file are available in the Supporting Information. It should be noted that in the current work, only the 2D topology information of molecule is required, but if other researchers need the precise 3D structures of these compounds, they can use the structures afforded by us directly.

Based on the data set, three models were developed for solubility prediction. In the first model, the whole data set was divided into a training set of 878 molecules and a validation set of 412 molecules. The 878 molecules were used to obtain the prediction model, and the 412 validated molecules were used to test the prediction potential of the obtained model. The classifications of the training set and the test set used here are the same as those used in Tetko et al. paper.¹⁰ In the second model, all molecules in the data set were used in a linear fitting to obtain the final model. In the third model, 1207 molecules were used to obtain the prediction model, and the other 83 molecules were eliminated from the learning set because these 83 molecules have been included in the second test set used by Klopman et al.²⁰ The molecules in the second test set were not included in the learning set, and thus they can be used as an independent test set.

We used two test sets to validate the actual prediction potential of our model. The first test set is a 21-member test set consisting of drugs and other environmentally interesting compounds such as pesticides. They were selected by Yalkowsky as a challenging test set of complex chemical structures for the validation of solubility models.³⁰ The second test set includes 120 compounds, which has been used by Klopman et al. in the validation of the group contribution model.²⁰ The reason that we selected these 120 molecules as the second test set is that we want to make a direct comparison between the prediction potential of the Klopman's group contribution model and that of the atom contribution model developed in this paper. The molecules in these two test sets were saved in two MACCS/sdf files. The molecules and the experimental log*S* values are listed in Tables 2 and 3. The MACCS/sdf files are available in the Supporting Information.

Atom Typing Rules and Correction Factors. In a group or atom addition model, the aqueous solubility values are computed from an equation as follows

$$\log S = C_D + \sum_i n_i a_i \quad (1)$$

where $\log S$ is the logarithm of the solubility; C_0 is a constant characteristic of solvent; a_i is the contribution coefficient of the i th group or the i th atom type in a molecule; and n_i is the number of occurrences of the i th group or the i th atom type in a molecule. The contribution for each group or atom type is determined by using a multiple linear regression (MLR) or other statistical techniques.

According to eq 1, different atom types should have different contributions to solubility. Here, we defined 76 basic atom types for the elements commonly found in organic molecules (C, O, N, P, S and halogens). The classification scheme differentiates atoms according to (i) element, (ii) hybridization state, and (iii) nature of the neighboring atoms. This establishes the rough theoretical support for the assumption that a certain type of atom has a specific contribution to the aqueous solubility. To allow for portability and simple implementation of the classification system, all atom types are presented in SMARTS.³¹ The SMARTS definitions for all atom types are listed in Table 1. The atom types represented by SMARTS were determined by using the SMARTS system included in OELIB.³² More detailed descriptions about SMARTS and OELIB can be found in refs 31 and 32. Here, we did not define any atom types for hydrogen atoms, which means that all heavy atoms connected with hydrogen are united atom types. In fact, we gave different definitions for heavy atoms connected with a different number of hydrogen atoms, and thus the contribution for hydrogen is implicitly included in that of the central heavy atom.

When we developed the $\log P$ prediction model using the atom contribution approach,²³ we found that the prediction $\log P$ values for many compounds with hydrophobic carbon atoms or intramolecular hydrogen bonds exist with large differences in the experimental values. The large deviations are sometimes explained by inter-/intramolecular hydrophobic interactions and intramolecular hydrogen bond interactions. Here, we found that the inter-/intramolecular hydrophobic interactions can effect the prediction of $\log S$, so we introduced a correction factor named "hydrophobic carbon".

We defined sp^3 - or sp^2 -hybridized carbon without any attached heteroatom (any atom other than carbon) with the 1–4 relationship as hydrophobic carbon. It should be noted that sp^2 -hybridized aromatic carbons were not considered as hydrophobic carbons. Moreover, the sp^2 -hybridized carbons in the ring were also not considered as hydrophobic carbons, because the sp^2 -hybridized carbon in the ring was relatively rigid and it was not easy to adjust conformation to form aggregation.

Here, we used another correction factor, which is the square of molecular weight (MW^2). In the work of Tetko et al.,¹⁰ the calculated results indicate that molecular weight is an important descriptor in prediction model. Here, we found that the square of molecular weight was more effective than molecular weight, so in the current work, MW^2 is used as a correction factor in the multiple linear correlations.

After including correction factors, the $\log S$ is described as

$$\log S = C_0 + \sum_i a_i n_i + \sum_j b_j B_j \quad (2)$$

where C_0 is a constant; a_i and b_j are regression coefficients; n_i is the number of occurrences of the i th atom type in a molecule, and B_j is the number of occurrences of the j th correction factor.

A program named LOGS-FIT was developed to identify the occurrence of each atom type in a compound from sdf files. A standard multiple linear regression analysis was used to get the contribution for each atom type. The contribution coefficient for each atom type was then used to give a prediction for a new molecule.

RESULTS AND DISCUSSION

The program, drug-LOGS v1.0, was developed in C programming language. The program can read a single molecule or multiple molecules (represented in single SYBYL/mol2 file, single MACCS/mol file, SYBYL/mol2 database file, or MACCS/sdf database file), perform atom typing, calculate occurrences of each atom type, detect correction factors, and then calculate $\log S$ using the parameters from MLR. The drug-LOGS program has been embedded into our drug-ADME program as a subroutine. Until now, the drug-ADME program can give several important ADME-concerned properties including $\log P$, $\log S$, and $\log BB$ (blood-brain partitioning).

Prediction Models. Any additive method, either by fragment or atom, needs a relevant scheme for fragment/atom classification. The quality of such a classification scheme can be evaluated by how well the calculated $\log S$ values agree with their experimental counterparts. Here, the training set of 878 molecules used by Tetko et al.¹⁰ was used to determine the most suitable atom typing rules. Originally, we used the atom typing system of the SLOGP model. In SLOGP, we used an atom typing system of 100 atom types. Using the SLOGP atom typing system, a fairly good correlation between experimental and predicted solubility was obtained ($n = 878$, $r = 0.95$, $s = 0.61$), but we also think that the SLOGP system needs more elaborate improvement. Using the SLOGP atom typing system, the occurrence of many atom types is very low, for example, the occurrence of atom types 8, 69, 78, or 79 is only two times and that of atom type 84 is only one time. Moreover, three atoms in the training set were not defined using the SLOGP atom typing system. Due to random correlation the obtained contribution coefficients for these atom types with low occurrences may exist with large deviations from the right answers. To get a more reliable prediction model, we make some modifications to the old atom typing system according to three principles. First, enough occurrences of the atom types are necessary to yield reliable results. Second, each atom type should be independent, and the redundant atom types should not be allowed. Third, the number of atom types should be controlled to as few as possible in order to avoid overfitting problem. According to the above principle, the atom typing system was carefully adjusted. The final atom typing system includes 76 atom types. The SMARTS definitions for these 76 atom types are listed in Table 1. When using the new atom typing rules and two correction factors, the obtained linear model for the training set with 878 molecules are statistically significant ($r = 0.96$, $s = 0.59$).

In previous work of Tetko et al.,¹⁰ the authors used a validation set of 412 compounds to optimize the training

Table 1. Atom Typing Rules and Their Contributions to logS

no.	SMARTS representation	occurrence ^a	contribution1 ^a	occurrence ^a	contribution2 ^b
1	[CX4;H4]	199	0.277	301	0.273
1	[CX4;H3]	199	0.277	301	0.273
2	[CX4;H3][#6]	22	-0.045	36	-0.099
3	[CX4;H3][CX4,c,F,Cl,Br,I]	549	-0.392	838	-0.409
4	[CX4;H3][CX3,c,F,Cl,Br,I] = [#8,#7]	63	0.254	86	0.205
5	[CX4;H3][CX4,c,F,Cl,Br,I]~[#8,#7]	203	-0.067	307	-0.059
6	[CX4;H2]	8	0.101	12	0.184
7	[CX4;H2][#6]	213	-0.089	303	-0.052
8	[CX4;H2]([#6])[#6]	65	-0.314	97	-0.179
9	[CX4;H2]([#6 × 4,c,F,Cl,Br,I])[#6 × 4,c,F,Cl,Br,I]	766	-0.307	1136	-0.334
10	[CX4;H2][#6] = [#8,#7]	131	0.180	191	0.183
11	[CX4;H2][CX4,c,F,Cl,Br,I]~[#8,#7]	79	0.061	107	0.051
12	[CX4;H2]([CX4,c,F,Cl,Br,I])[CX4,c,F,Cl,Br,I]~[#8,#7]	244	-0.075	362	-0.094
13	[CX4;H2]-[OH,NH2,NH]	99	-0.066	157	-0.003
14	[CX4;H]	11	0.017	18	0.105
15	[CX4;H]([#6])[#6]	87	-0.149	135	-0.216
16	[CX4;H]([#6])([#6])[#6]	54	-0.025	91	0.113
17	[CX4;H]([#6 × 4,c,F,Cl,Br,I])([#6 × 4,c,F,Cl,Br,I])[#6 × 4,c,F,Cl,Br,I]	238	-0.231	377	-0.201
18	[CX4;H1]-[OH,NH2,NH]	147	-0.368	211	-0.339
19	[CX4;H0]	30	-0.553	52	-0.526
19	[CX4;H0][#6]	30	-0.553	52	-0.526
19	[CX4;H0]([#6])[#6]	30	-0.553	52	-0.526
20	[CX4;H0]([#6])([#6])[#6]	65	-0.477	113	-0.491
21	[CX4;H0]([#6])([#6])([#6])[#6]	111	0.216	179	0.206
21	[CX4;H0]([#6 × 4,c])([#6 × 4,c])([#6 × 4,c])[#6]	111	0.216	179	0.206
21	[CX4;H0]([#6 × 4,c])([#6 × 4,c])([#6 × 4,c])[#6 × 4,c]	111	0.216	179	0.206
22	[C;H2] = *	38	-0.126	53	-0.227
23	[C;H1] = *	152	-0.323	244	-0.332
24	[C;H0] = *	112	-0.302	188	-0.275
25	[C;H1] = O	23	-0.420	25	-0.419
26	[C;H0] = O	331	-0.964	482	-0.950
27	[C;r] = O	133	-0.915	217	-0.868
28	C(=C)=C	26	-0.379	39	-0.407
28	C#*	26	-0.379	39	-0.407
29	[c;H1](~c)~c	23	-0.227	34	-0.324
30	[c;H1;r6](~c)~c	2494	-0.310	3667	-0.307
31	[c;H1](~c)~[a;!c]	124	-0.021	167	-0.107
32	[c;H1](~[a;!c])~[a;!c]	25	-0.431	37	-0.602
33	[c;H0](~[a;!c])~[a;!c]	8	0.301	13	-0.206
34	[c;H0](~[#6])(~[a;!c])~[a;!c]	56	-1.022	78	-1.180
35	[c;H0](~*)(~c)~c	139	-0.481	207	-0.416
35	[c;H0](~C)(~c)~[a;!c]	139	-0.481	207	-0.416
36	[c;H0](~[CX4,F,Cl,Br,I])(~c)~c	622	-0.226	940	-0.226
36	[c;H0](~[CX4,F,Cl,Br,I])(~c)~[a;!c]	622	-0.226	940	-0.226
37	[c;H0](~c)(~[c])~c	292	-0.700	429	-0.719
38	[c;H0](~[#6;!F;!Cl])(~c)~c	551	-0.359	816	-0.383
39	[c;H0](~[#6])(~c)~[a;!c]	69	-0.791	109	-0.856
40	[#8;H1]	99	0.564	154	0.460
41	[#8;H1]C	300	0.332	454	0.303
42	[#8;H0]	155	-0.306	239	-0.332
43	[#8;H0]C=O	111	-0.246	163	-0.239
44	[#8;H0]([#6 × 4])[#6 × 4]	47	-0.299	73	-0.295
45	[#8] = C	419	0.533	616	0.535
45	[#8] = c	419	0.533	616	0.535
46	[#8] = C([#6])[#6]	104	0.398	166	0.370
46	[#8] = c([#6])[#6]	104	0.398	166	0.370
47	[o]	16	-0.561	29	-0.082
48	[OX1]~S	98	-1.083	137	-0.092
48	[OX1]~P	98	-1.083	137	-0.092
49	[OX2](*)[P]	21	-1.245	65	-0.690
50	[N;H2]	127	0.200	185	0.205
50	[N;H2][C]	127	0.200	185	0.205
51	[N;H1]	82	-0.025	108	-0.047
52	[N;H1]([#6 × 4])	42	-0.129	66	-0.122
53	[N;H1]([#6 × 4])[#6 × 4]	5	0.078	6	0.006
54	[N;H1;r]	57	0.436	93	0.166
55	[N;H0]	8	0.447	11	0.210
56	[N;H0]([#6 × 4])	85	-0.387	125	-0.481
56	[N;H0]([#6 × 4])[#6 × 4]	85	-0.387	125	-0.481
56	[N;H0]([#6 × 4])([#6 × 4])[#6 × 4]	85	-0.387	125	-0.481
57	[N;H0;r]([#6 × 4])([#6 × 4])[#6 × 4]	35	-0.026	44	-0.042
58	[n]	69	0.327	94	0.451
59	[n]~[n]	147	0.064	204	0.196

Table 1 (Continued)

no.	SMARTS representation	occurrence ^a	contribution1 ^a	occurrence ^a	contribution2 ^b
59	[n]~[*]~[n]	147	0.064	204	0.196
60	[nH]	31	-0.182	49	-0.068
61	[N] = *	20	-0.293	29	-0.245
62	N#*	12	0.255	17	0.166
62	N(= *) = *	12	0.255	17	0.166
63	[NX3](O)=O	28	-0.223	53	-0.139
63	[NX3](O)-O	28	-0.223	53	-0.139
64	[S;H1]	4	-0.436	6	-0.645
65	[S;H0]	49	-1.080	85	-1.118
66	s	22	-0.905	35	-0.696
66	[S;H0]1* = ** = *1	22	-0.905	35	-0.696
66	[S;H0]1*:*:*1	22	-0.905	35	-0.696
66	[S;H0]1* = **:*1	22	-0.905	35	-0.696
66	[S;H0]1* = *~*~* = *1	22	-0.905	35	-0.696
66	[S;H0]1*:*~*~*1	22	-0.905	35	-0.696
66	[S;H0]1* = *~*~*:*1	22	-0.905	35	-0.696
67	[SX3] = [OX1]	50	1.802	68	-0.109
67	[SX4](= [OX1]) = [OX1]	50	1.802	68	-0.109
67	[SX3]-[OX1]	50	1.802	68	-0.109
67	[SX4](-[OX1])-[OX1]	50	1.802	68	-0.109
68	[PX3]	10	1.578	26	0.510
68	[PX4]	10	1.578	26	0.510
69	[F]	49	-0.332	73	-0.293
70	[Cl]	127	-0.657	245	-0.705
71	[Br]	24	-1.006	39	-0.993
72	[I]	7	-1.886	9	-1.771
73	[F]c	14	-0.436	23	-0.438
74	[Cl]c	322	-0.954	472	-0.964
75	[Br]c	18	-1.276	28	-1.312
76	[I]c	15	-1.631	20	-1.514
78	hydrophobic carbon	997	-0.253	1497	-0.230
79	MW ²		0.00008		0.00008
	constant		0.500		0.518

^a Based on a training set of 878 molecules. ^b Based on the whole data set of 1290 molecules.

process of ANN. Here, we predicted the solubility of the molecules in the validation set based on the contribution coefficient from the training set. The predictions for the validation set ($r = 0.95$, $s = 0.63$) are a little worse than those for the training set. The good prediction for the test set means that the obtained model is reliable. The contribution coefficient and the occurrence for each atom type and correction factor are listed in Table 1. It seems that the predictions to the validation set of 412 molecules using the Tetko's ANN model ($r = 0.96$, $s = 0.60$) are better than those using the drug-LOGS model. It should be noted that in Tetko's work, the validation set employed is to control the termination of the learning of ANN. The validation set was really included in the model development. The prediction used for the validation set was only the optimized results of ANN, not the actual predicted results of the model. Because of the good nonlinear regression ability of ANN, the models produced by ANN may be very statistically significant, but the actual prediction potentials of the models may be questionable. For example, in the work of Engkvist et al.,¹¹ the authors developed a model based on a training set of 1160 and a validation set of 130 using ANN. The average squared correlation coefficient (r^2) was 0.96 (standard deviation 0.56) for the training set and 0.95 (standard deviation 0.57) for the test set. Using the model, Engkvist predicted the solubilities of a test set of 2767 molecules. However, the result for the independent validation set was rather poor, yielding a squared correlation coefficient of 0.79 and a standard deviation of 1.18. Therefore, the authors developed a second model. This time the training set includes

3042 molecules and the test set includes 309 molecules. The statistical significance of the second model is worse than that of the first model. The average squared correlation coefficient (r^2) was 0.91 (standard deviation 0.84) for the training set and 0.89 (standard deviation 0.87) for the test set, respectively. But for the independent validation set of 307 molecules, a squared correlation coefficient of 0.86 (standard deviation 0.80) was obtained. The reason that the first model performs worse for the test set is that the data set used for learning is small, and thus overfitting problem was encountered in the fitting of ANN.

Additionally, all molecules in the training set and the validation set were used to develop the log S prediction model. Based on the data set of 1290 molecules, we achieved a regression coefficient of 0.96 and a standard deviation of 0.62, which is a little worse than the fitting calculated only using the training set of 878 molecules ($r = 0.96$, $s = 0.59$). To further test the robust of the model, we have performed a leave-one-out cross-validation on the whole data set, which give nearly the same results of MLR ($q = 0.95$, $s = 0.63$). The contribution efficient and the occurrence for each atom type and correction factor calculated using the whole data set are listed in Table 1. The correlation between the experimental and the predicted log S values are illustrated in Figure 2. According to Table 1, it can be found that for most atom types the contribution coefficients do not exist with large differences whether using 878 molecules or 1290 molecules. But for some atom types with low occurrences the contribution coefficients exist with obvious differences. For example, using 878 molecules the coefficient for type

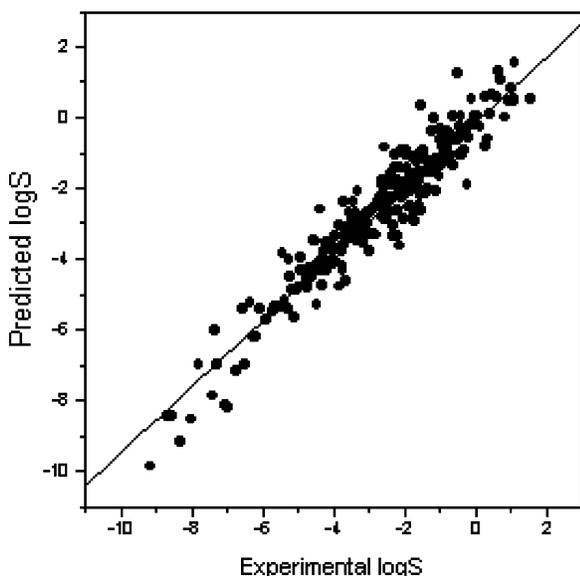


Figure 1. The predicted versus observed aqueous solubilities for the whole data set.

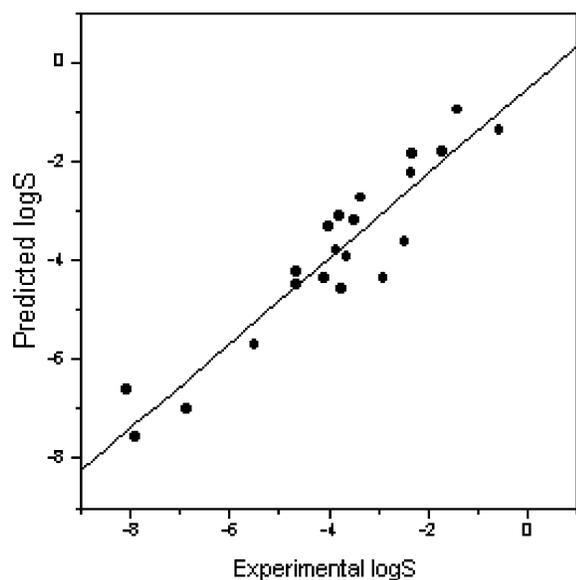


Figure 2. The predicted versus observed aqueous solubilities for test set 1.

32 is -0.432 , while using 1290 molecule, this value decreased to -0.602 . It is not strange because insufficient samples in the data set may produce unstable results caused by a random correlation of MLR. For MLR, sufficient samples in the data set should be necessary in order to obtain effective and reliable results. Here, the contribution coefficients based on 1290 molecules should be more reliable than those only based on 878 molecules, so in our $\log S$ prediction model, the coefficients calculated using 1290 molecules were used.

It is very informative to study the contribution coefficient of each atom type. For example, the solubility of halogen atoms clearly decreases in the following order: $-F$ (-0.293), $-Cl$ (-0.705), $-Br$ (-0.993), $-I$ (-1.771). With the attachment to an aromatic ring the solubility generally lowers the solubility of halogens besides I. The detailed information obtained here can only be afforded by the atom/group contribution approaches, while it cannot be afforded by the experiment-related or the descriptor-based techniques.

Correction Factors. Here, to consider the intra-/inter-molecular hydrophobic interaction and the bulk effect, we considered two correction factors including hydrophobic carbon and “square of molecular weight” in model development.

The correction factor of hydrophobic carbon is very important in our model. When we do not consider this correction factor, the solubility of many compounds with hydrophobic carbon or long aliphatic chains was greatly overestimated. We think these molecules may introduce intramolecular hydrophobic folding or intermolecular aggregation and influence the experimental solubility. When we do not consider any correction factor, based on the data set of 1290 molecules a linear model with $r = 0.951$ and $s = 0.651$ was obtained. After introducing the correction factor of hydrophobic carbon, the correlation of the model was improved obviously, and MLR generates a model with $r = 0.955$ and $s = 0.624$.

In the work of Tetko et al., the calculated results indicate that molecular weight (MW) is an important descriptor in prediction models. After introducing MW in fitting, we found that the linear correlation of the model nearly does not have any improvement ($r = 0.955$, $s = 0.624$). But if we introduced the square of molecular weight (MW^2) in fitting, the linear correlation of the model increased a little ($r = 0.956$, $s = 0.621$). So, in our work, MW^2 was used as a correction factor in the prediction of $\log S$.

Actual Prediction of $\log S$ to Test Set 1. Only from the correlation between the experimental $\log S$ and the calculated values to the molecules in training set, our prediction model is very significant. But it is well-known that the actual prediction power may only be determined based on a list of compounds as the test set. Moreover, we want to know if our model can give comparative predictions with other $\log S$ calculation procedures.

The contribution coefficients based on the data set of 1290 molecules were used to predict the solubilities of test set 1. This test set was designed by Yalkowsky, which is compiled of 21 commonly used compounds of pharmaceutical and environmental interest.¹⁹ The reason that we selected this set of compounds is that this test set was also used by several other researchers.^{6,9,10,12,14,17,20} The predicted and experimental $\log S$ values of test set 1 are presented in Table 2 and Figure 2. They show that there is a significant correlation between the predicted and experimental values ($r = 0.94$, $s = 0.84$). However, the correlation is not as good as that of the data set used for model development. Inspection of the results indicates that three of the compounds were predicted poorly with errors over one log unit; the worst is for phenolphthalein with an error of 1.5 log unit.

Table 2 contains the corresponding linear coefficient (r) and standard deviation (s) values using the other seven $\log S$ prediction approaches. From Table 2, it can be found that the correlation coefficient ($r = 0.94$) of our model is higher than those of the other approaches except for the Huuskonen's ANN model ($r = 0.95$) and the Tetko's ANN model ($r = 0.95$). From the standard deviation, our model ($s = 0.64$) performs similarly with the Huuskonen's ANN model ($s = 0.63$) and the Tetko's ANN model ($s = 0.64$), while better than the other models. It should be noted that the prediction to these 21 compounds cannot give a decisive rank of all these $\log S$ calculation procedures because the number of

Table 2. Observed and Predicted Aqueous Solubilities for Test Set 1 of 21 Compounds

no.	name	log S_{exp}	Hou	Klopman ²⁰	Kühne ¹⁷	Yan ¹²		Huuskonen ⁶		Tetko ¹⁰	Liu ⁹	Wegner ¹⁴
			MLR	MLR	MLR	MLR	ANN	MLR	ANN	ANN	ANN	GA
1	2,2',4,5,5'-PCB	-7.89	-7.57	-7.90	-7.47	-7.12	-7.85	-7.40	-7.55	-7.21	-7.57	-7.55
2	benzocaine	-2.32	-1.86	-1.71	n/a ^a	-1.81	-2.19	-1.85	-1.45	-1.79	-1.63	-2.05
3	aspirin	-1.72	-1.81	-1.52	-1.93	-1.44	-1.87	-1.74	-2.10	-1.69	-1.81	-1.81
4	theophylline	-1.39	-0.97	-1.07	-0.54	-1.07	-1.27	-0.78	-0.73	-1.71	-0.69	-1.21
5	antipyrine	-0.56	-1.39		-1.90	-2.99	-1.31	-1.20	-1.41	-1.29	-0.89	-1.74
6	atrazine	-3.85	-3.79	-3.05	-3.95	-2.44	-3.83	-2.18	-1.51	-3.51	-3.70	-2.82
7	phenobarbital	-2.34	-2.25	-2.08	-2.41	-2.81	-2.80	-2.88	-2.50	-2.97	-2.89	-2.36
8	diuron	-3.80	-3.13	-2.85	-3.38	-3.26	-3.70	-3.20	-2.85	-2.86	-3.01	-3.31
9	nitrofurantoin	-3.47	-3.21	-2.19	-2.62	-0.45	-2.52	-3.03	-2.89	-3.42	-3.09	-2.82
10	phenytoin	-3.99	-3.33	-3.47	-5.25	-2.99	-3.18	-3.48	-3.09	-3.40	-3.52	-2.90
11	diazepam	-3.76	-4.58		-4.51	-5.19	-4.81	-4.26	-4.08	-4.05	-4.37	-4.14
12	testosterone	-4.09	-4.35	-5.17	-4.62	-4.11	-4.52	-4.17	-4.49	-3.98	-4.13	-4.27
13	lindane	-4.64	-4.25	-4.88	-5.80	-3.93	-5.04	-5.34	-4.91	-4.71	-4.91	-3.98
14	parathion	-4.66	-4.51	-3.94	-4.59	-2.64	-3.66	-3.98	-3.64	-4.13	-4.31	-4.06
15	diazinon	-3.64	-3.93	-5.29	-4.98	-3.06	-2.66	-4.10	-3.56	-4.01	-3.43	-4.18
16	phenolphthalein	-2.90	-4.37	-4.48	-4.61	-4.28	-4.62	-4.05	-4.16	-3.99	-4.31	-4.64
17	malathion	-3.37	-2.74	-2.94	-3.48	-3.45	-2.79	-3.63	-2.52	-3.24	-3.73	-2.96
18	chlorpyrifos	-5.49	-5.70	-5.77	-3.75	-4.33	-4.79	-5.46	-4.50	-5.61	-5.31	-6.41
19	prostaglandin	-2.47	-3.63	-4.21	n/a ^a	-4.06	-3.07	-4.35	-3.80	-3.29	-3.52	-3.98
20	p,p'-DDT	-8.08	-6.62	-8.00	-7.75	-6.60	-7.86	-7.82	-7.93	-7.67	-7.59	-6.85
21	chlordane	-6.86	-7.00	-7.55	-6.51	-6.41	-7.66	-8.35	-7.32	-7.29	-7.23	-6.47
<i>n</i>			21	21(19)	19	21	21	21	21	21	21	21
<i>r</i>			0.94	0.84(0.92) ^b	0.87	0.75	0.92	0.91	0.89	0.95	0.95	0.91
<i>s</i>			0.64	1.24(0.86) ^b	1.08	1.20	0.77	0.88	0.91	0.63	0.64	0.79

^a Predicted values are not indicated. ^b Values with and without (in parentheses) outliers.

compounds in this test set is rather limited. But the comparison at least demonstrates that our model yielded acceptable estimations for the tested compounds.

Actual Prediction of log S to Test Set 2. The second test set of 120 molecules was used in the development of the Klopman's group contribution model.²⁰ After careful observation, we found that 83 molecules in test set 2 are included in our training set. So in order to make an unbiased comparison between our model and the Klopman's model, we eliminated these 83 duplicates from the learning set, and thus the new learning set only includes 1207 molecules. Using the new learning set, we give a new fitting and obtained the corresponding contribution coefficients. Now, test set 2 can be used as an unbiased test of the accuracy of our model. The results were then compared with those obtained using the Klopman's model.

The calculations results using these two models are listed in Table 3. The correlation between the experimental and the predicted log S values for test set 2 is illustrated in Figure 3. The standard deviation was found to be 0.76 with our model. But with the Klopman's model, the standard deviation was found to be 0.84. It should be noted that the standard deviation reported by Klopman et al. is 0.79, but when we used the calculated results in ref 20 and performed a linear fitting, we found that the standard deviation is 0.84, not 0.79. Using our model, we achieved an unsigned mean error of 0.57 log unit, which is obviously better than that (0.70) using the Klopman's model. It is clear that our model may be more effective than the Klopman's model. The high correlation coefficient ($r = 0.96$) and low unsigned mean error (0.57) show that the predicted results and the experimental values are consistent. Among all those molecules, the predictions for compounds 91, 92, and 102 are the worst. For these three compounds, the unsigned mean errors between the predicted and experimental values are larger than 2.0 kcal/mol. Among

these three compounds, compound 91 was also highly underestimated by the Klopman's model (error: -2.71 log unit), and thus this compound may be treated as an outlier. Not considering this compound, the linear correlation between the experimental solubilities and the predicted values increased from 0.94 to 0.96, and standard deviation from 0.79 to 0.74. Now we cannot give an exact explanation of these deviations. They may be brought by experimental errors, inadequate atom typing rules, or insufficient correction factors.

The basic ideas of the Klopman's model and our model are similar, but the only difference is that the basic unit in the Klopman's model is a functional group while in our model the basic unit is an atom. From principle, the group-based approach may give a better consideration of the electrostatic distribution and interactions between constituted parts in a group than the atom-based ones. But for a fragment-based method, the classifications of the basic fragments are very difficult, and an additive method will not be able to do the calculations for any compound containing a missing fragment. So, sometimes the fragment-based approach may not calculate the compounds with an undefined fragment. For example, the old version of the Klopman's model cannot give predictions for many compounds with missing fragments. But for an atom-based method, we can give a full description of atoms with different chemical environments very easily. Moreover, based on the limited experimental data, we cannot define excessive types of functional groups; otherwise, the obtained contribution coefficient from multiple linear regressions may be not reliable caused by overfitting. Furthermore, from the technical respect, the programming and extension of the atom-based approaches are much simpler and easier than those the fragment-based approaches. For example, in our program, based on OELIB, the definition and the determination of the

Table 3. Observed and Predicted Aqueous Solubilities for Test Set 2 of 120 Compounds

no.	name	exptl logS	calcd logS ^a	calcd logS ^b
1	carbamic acid, ethyl ester	0.85	-0.12	-0.00
2	benzamide	-0.96	-0.93	-1.54
3	glycine	0.52	1.52	0.76
4	L-serine	-0.02	1.29	0.89
5	L-glutamine	-0.55	0.26	0.41
6	benz a anthracene, 7,12-dimethyl-	-7.02	-7.76	-8.17
7	lindane	-4.59	-6.04	-4.61
8	L-leucine	-0.8	-0.91	-1.02
9	L-methionine	-0.42	-0.62	-0.47
10	L-phenylalanine	-0.92	-1.00	-1.26
11	L-valine	-0.30	-0.46	-0.51
12	endrin	-6.29	-6.04	-6.24
13	L-tryptophan	-1.28	-1.84	-2.18
14	L-isoleucine	-0.59	-0.91	-1.04
15	4-chlorobenzoic acid	-3.31	-2.84	-2.08
16	L-arginine	0.00	0.42	0.11
17	codeine	-1.52	-2.53	-1.87
18	1,2,3-propanetricarboxylic acid,2-hydroxy-	0.51	1.77	0.91
19	2-propenamide	0.95	-0.24	-0.29
20	2-propenoic acid, 2-methyl-	0.00	-0.17	-0.00
21	2-propenoic acid, 2-methyl-, methyl ester	-0.80	-1.17	-0.40
22	1,2-benzisothiazol-3(2H)-one, 1,1-dioxide	-1.64	-2.16	-1.60
23	1-naphthalenesulfonic acid, 2-amino-	-1.70	-2.02	-2.92
24	9,10-anthracenedione	-5.19	-5.87	-4.17
25	1,2-benzenedicarboxylic acid, butyl phenylmethyl ester	-5.64	-4.77	-4.49
26	9H-carbazole	-5.27	-4.62	-4.04
27	benzenamine, 2-nitro-	-1.96	-1.43	-1.94
28	1,2-benzenedicarboxylic acid	-2.11	-1.02	-1.22
29	phenol, 2-methoxy-	-1.96	-1.43	-0.95
30	1-naphthalenol	-2.22	-1.59	-2.85
31	1,2-dicyanobenzene	-2.38	-0.97	-1.12
32	benzenamine, N,N-diethyl-	-3.03	-1.98	-1.89
33	biphenyl	-4.30	-3.24	-3.81
34	1,1'-biphenyl-4-ol	-3.48	-2.53	-3.40
35	10H-phenothiazine	-5.10	-6.17	-3.45
36	1,1'-biphenyl-4,4'-diamine	-2.70	-3.84	-3.56
37	1,2-benzenediamine	-0.42	-0.50	-1.04
38	2-propanol, 1,2-dichloro-	-0.11	-0.41	-1.02
39	2-propenoic acid, methyl ester	-0.22	-0.76	-0.40
40	2-imidazolidinethione	-0.71	-0.64	-0.44
41	2-furancarboxaldehyde	-0.10	0.00	-0.68
42	benzene, 1,3,5-trinitro-	-2.89	-1.77	-3.59
43	1,2,3-propanetriol, triacetate	-0.60	0.73	-0.52
44	diazene, diphenyl-	-2.75	-2.90	-3.53
45	acetamide, N-phenyl-	-1.33	-1.89	-1.49
46	diethylthiourea, N,N'-	-1.46	-2.56	-1.05
47	2-propenoic acid, 2-methylpropyl ester	-1.21	-1.87	-1.14
48	ethanesulfonic acid, 2-amino-	-0.09	1.07	0.90
49	2-pentanone, 4-methyl-	-0.74	-1.05	-1.00
50	2-pentene	-2.54	-1.86	-1.96
51	butanedioic acid	-0.20	0.65	1.12
52	2,4-hexadienoic acid, (E,E)-	-1.77	-0.50	-1.50
53	2-propanol, 1,1'-iminobis-	0.81	0.91	0.73
54	endosulfan	-6.15	-5.25	-5.62
55	anthranilic acid, o-	-1.52	-0.75	-1.47
56	2-naphthalenesulfonic acid, 5-amino-	-2.35	-2.53	-2.92
57	dinitrotoluene, 2,4-	-2.82	-2.85	-3.14
58	hydrazine, 1,2-diphenyl-	-2.92	-1.57	-3.14
59	benzaldehyde, 4-hydroxy-	-0.96	0.48	-0.90
60	benzaldehyde, 4-methoxy-	-1.49	-0.27	-1.39
61	4-heptanone	-1.30	-2.52	-1.49
62	2-butenal	0.32	0.09	-0.46
63	1-butanol, 3-methyl-, acetate	-1.92	-1.27	-0.05
64	1-naphthalenamine	-1.92	-3.04	-3.14
65	2-naphthalenol	-2.28	-2.09	-2.85
66	D-glucopyranoside, 2-(hydroxymethyl)phenyl	-0.85	0.26	-0.85
67	2-propenoic acid, 3-phenyl-	-2.48	-3.40	-1.87
68	2-propenoic acid, ethyl ester	-0.74	-1.24	-0.74
69	4-pyrimidinone, 2,3-dihydro-2-thioxo-	-2.26	-1.33	-1.77
70	acetic acid, hexyl ester	-2.46	-1.79	-1.98
71	mercaptobenzothiazole, 2-	-3.15	-2.03	-3.30
72	benzoic acid, 4-amino-	-0.40	-1.52	-0.99
73	acenaphthylene	-3.96	-4.27	-4.09

Table 3 (Continued)

no.	name	exptl logS	calcd logS ^a	calcd logS ^b
74	dibenzo-p-dioxin	-5.31	-4.63	-3.75
75	1,1-ethanediol, 2,2,2-trichloro-	0.72	0.63	-1.04
76	DL-alanine	0.26	0.31	0.37
77	decanoic acid	-3.44	-3.70	-3.36
78	2-propanol, 1,1,1-trifluoro-	0.30	0.03	-0.83
79	guanidine, cyano-	-0.31	1.02	-0.03
80	5-nonanone	-2.59	-3.80	-2.56
81	1,2-dinitrobenzene	-3.10	-2.55	-2.84
82	2,3-dichloro-2-methyl-butane	-2.69	-3.78	-2.23
83	1,2-diiodoethylene	-3.22	-2.28	-3.04
84	3-methyl-3-hexanol	-1.00	-0.88	-1.09
85	ethane, 1,2-diethoxy-	-0.77	0.16	-0.06
86	4-methylpentanol	-1.14	-1.16	-1.05
87	1-phenylethanol	-0.92	-0.56	-1.08
88	1-hexen-3-one	-0.83	-1.73	-1.35
89	1,2,3,6,7,8-hexahydropyrene	-5.96	-7.25	-6.14
90	dicamba	-1.70	-3.17	-2.94
91	dodine acetate	-2.63	-5.34	-5.52
92	biphenyl, 3,4-dichloro-	-7.44	-6.30	-5.43
93	asulam	-1.66	-1.21	-1.57
94	O-tert-butyl carbamate	0.10	-1.34	-0.57
95	3-methyl-3-heptanol	-1.60	-1.38	-1.63
96	2,4',5-PCB	-6.25	-6.30	-6.21
97	2,3-dimethyl-1-butanol	-0.39	-0.83	-0.97
98	ditolyl ether	-4.85	-3.84	-4.39
99	3-methyl-2-heptanol	-1.72	-2.45	-1.99
100	2',3,4,4',5'-PCB	-7.39	-8.25	-7.73
101	2,3',4',5-PCB	-7.25	-7.63	-6.98
102	dichlorodibenzo-p-dioxin, 2,7-	-7.82	-6.59	-5.34
103	2,3,4,2',4',5'-PCB	-8.32	-8.62	-8.46
104	2,2',3,3',4,4',5,5'-PCB	-9.16	-9.33	-9.88
105	2,3,4,2',5'-PCB	-7.91	-8.07	-7.73
106	2,3,3',4,4',5-PCB	-7.82	-8.76	-8.46
107	2,3,4'-PCB	-6.26	-6.35	-6.21
108	2-chlorodibenzo-p-dioxin	-5.82	-5.66	-4.55
109	2,2',3,3',4,4',5,5',6-nonachlorobiphenyl	-10.26	-9.65	-10.56
110	2,2',3,5'-PCB	-6.47	-6.92	-6.98
111	2,2',3,5,5',6-hexachlorobiphenyl	-7.42	-8.45	-8.46
112	2,2',3,4,4',5',6-heptachlorobiphenyl	-7.92	-8.93	-9.18
113	2,2',3,3',4,5,5',6,6'-PCB	-10.41	-9.65	-10.56
114	2,2',3,4,5,5',-hexachlorobiphenyl	-7.68	-7.45	-8.46
115	2,2',3,4,5,5',6-heptachlorobiphenyl	-8.94	-8.93	-9.18
116	2,2',3,4,6-PCB	-7.43	-7.87	-7.73
117	2,3,4,5,2',3'-PCB	-8.78	-8.45	-8.46
118	2,3,6-PCB	-6.29	-6.86	-6.21
119	2,2',4,6,6'-PCB	-7.32	-8.55	-7.73
120	2,3,3',4,4',6-hexachlorobiphenyl	-7.66	-8.61	-8.46
<i>r</i>			0.96	0.96
<i>s</i>			0.84	0.79
<i>MUE</i> ^c			0.70	0.57

^a Predicted values using the Klopman's group contribution. ^b Predicted values using the drug-LOGS model. ^c *MUE* represents mean unsigned error.

atom types are very simple, and the whole program is very short and easy to be interpreted.

To some extent, an additive method is the art of fragment/atom classification. The definition of atom types should be suitable, and too few atom types may not represent the different chemical environments effectively. Certainly, we do not mean that more atom types can produce better results. Sometimes, two or several atom types may not be fully independent, and the addition of redundant atom types cannot effectively enhance the prediction of the model. The number of atom types used here is much smaller than Klopman's set of 118. However, using fewer atom types does not weaken the power of our model. We think that after our iterative adjustment of atom types we have developed each of them into a more elaborate class for logS calculation.

CONCLUSION

A new atom contribution approach was used to correlate the aqueous solubility of 1290 organic compounds. The model was only based on 2D-molecular topology. The aqueous solubility was calculated from contributions of 76 atom types and 2 corrections factors. The model was able to predict the aqueous solubility of a diverse set of 1290 organic compounds with an overall correlation coefficient of 0.96 and a standard deviation of 0.62 log unit between the calculated and experimental data. The actual prediction of our model was validated through two external test sets. Comparison of the calculated results for test set 1 among several logS models demonstrates that our model can give similar results with the best model reported. Comparison of

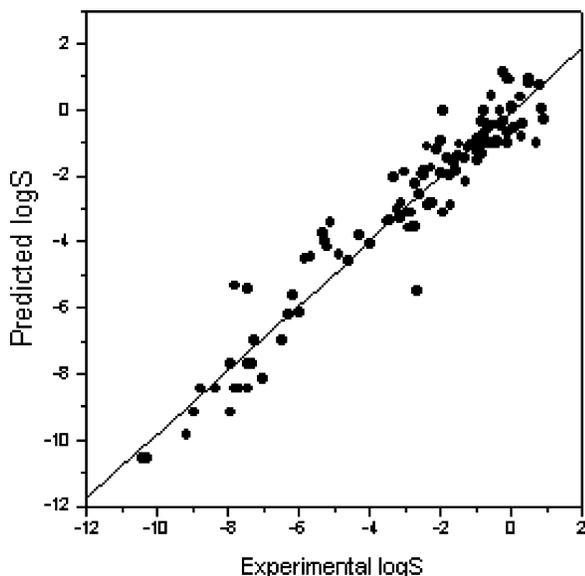


Figure 3. The predicted versus observed aqueous solubilities for test set 2.

the calculated results of test set 2 between the Klopman's model and drug-LOGS demonstrates that the drug-LOGS model is more predictive than the Klopman's model. Moreover, because our model uses an atom as the basic unit, it does not contain a missing fragment problem and can be used as a universal and effective approach for the prediction of solubility for any organic molecule.

ACKNOWLEDGMENT

This project is supported by National Natural Science Foundation of China (NSFC 29992590-2 and 29873003). We thank Dr. Igor V. Tetko for providing us data of the analyzed molecules.

Supporting Information Available: The methods proposed here and all the parameters for calculations on $\log S$ have been incorporated into a computer program called drug-LOGS. The drug-LOGS computer code can be obtained by contacting the authors. The drug-LOGS program has been tested on Linux operation systems. The CAS number, compound name, and experimental and calculated $\log S$ values for molecules of the data set are listed in Table A. The structures of the data set and two test sets are saved in MACCS/sdf files named data_set.sdf, test_set1.sdf, and test_set2.sdf (experimental $\log S$ values of all compounds are included). Table A and all structural files can be downloaded from the website: <http://www.cadd.chem.pku.edu.cn>. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- Hou, T. J.; Xu, X. J. Recent development and application of virtual screening in drug discovery: An overview. *Curr. Pharm. Des.* In press.
- Yalkowsky, S. H. Estimation of the Aqueous Solubility of Complex Organic Compounds. *Chemosphere* **1993**, *26*, 1239–1261.
- Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Non-Electrolytes. *J. Pharm. Sci.* **2000**, *90*, 234–252.
- Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354–357.
- 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.
- Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
- EcElroy, 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.
- McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355–1359.
- Liu, R.; So, S.-S. Development of Quantitative Structure–Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
- 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.
- Engkvist, O.; Wrede, P. High-Throughput, *In Silico* Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247–1249.
- Yan, A. X.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
- Butina, D.; Gola, J. M. R. Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837–841.
- Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.
- Nirmalakhandan, N. N. P.; Speece, R. E. Prediction of aqueous solubility of organic chemicals based on molecular structure. 2. Application to PNAs, PCBs, PCDDs, et. *Environ. Sci. Technol.* **1989**, *23*, 708–713.
- Suzuki, T. Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149–166.
- Kühne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group contribution method to estimate water solubility of organic chemicals. *Chemosphere* **1995**, *30*, 2061–2077.
- Lee, Y.; Myrdal, P. B.; Yalkowsky, S. H. Aqueous functional group activity coefficients (AQUAFAC) 4: Application to complex organic compounds. *Chemosphere* **1996**, *33*, 2129–2144.
- Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.
- 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.
- Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 2. Prediction of Partition Coefficient by Atom-additive Approach Based on Atom-weighted Solvent Accessible Surface Areas. *J. Chem. Comput. Sci.* **2003**, *43*, 1058–1067.
- Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. I. partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.
- Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 1. Applications of Genetic Algorithms on the Prediction of Blood-brain Partitioning of a Large Set Drugs. *J. Mol. Model.* **2002**, *8*, 337–349.
- Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* published ASAP.
- Yalkowsky, S. H.; Dannelfelser, R. M. *The ARIZONA dATABASE of Aqueous Solubility*; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.
- Syracuse Research Corporation. *Physical/Chemical Property Database-(PHYSPROP)*; SRC Environmental Science Center: Syracuse, NY, 1994.
- Weinenger, D. SMILES: a Chemical Language for Information Systems. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- Webblabviewer, Accelrys Inc., San Diego, USA. <http://www.accelrys.com>.
- Halgren, T. A. Merck molecular force field 0.1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Dekker: New York, 1992.
- James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual Daylight 4.62, Daylight Chemical Information Systems, Inc., Los Altos, 2001.
- OELIB, <http://www.eyesopen.com/oelib/>.