ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/220523979

A New Atom-Additive Method for Calculating Partition Coefficients

ARTICLE *in* JOURNAL OF CHEMICAL INFORMATION AND MODELING · MAY 1997

Impact Factor: 3.74 · DOI: 10.1021/ci960169p · Source: DBLP

CITATIONS
253

READS
46

3 AUTHORS, INCLUDING:

Renxiao Wang
Chinese Academy of Sciences
98 PUBLICATIONS **5,115** CITATIONS
SEE PROFILE

Luhua Lai
Peking University
185 PUBLICATIONS **4,565** CITATIONS
SEE PROFILE

Available from: Renxiao Wang
Retrieved on: 10 January 2016

A New Atom-Additive Method for Calculating Partition Coefficients

Renxiao Wang, Ying Fu, and Luhua Lai*

Institute of Physical Chemistry, Peking University, Beijing 100871, P.R. China

Received December 18, 1996[⊗]

A new method is presented for the calculation of partition coefficients of solutes in octanol/water. Our algorithm, XLOGP, is based on the summation of atomic contributions and includes correction factors for some intramolecular interactions. Using this method, we calculate the log *P* of 1831 organic compounds and analyze the derived parameters by multivariate regression to generate the final model. The correlation coefficient for fitting this training database is 0.968, and the standard deviation is 0.37. The result shows that our method for log *P* estimation is applicable to quantitative structure−activity relationship studies and gives better results than other more complicated atom-additive methods.

INTRODUCTION

Many biochemical, pharmacological, and environmental processes are dependent on the hydrophobicity of the molecules involved, and the parametrization of the hydrophobicity of a compound is important in quantitative structure−activity relationship (QSAR) studies. In 1964, Hansch and Fujita[1] pioneered a method for the correlation of biological activity and chemical structure. Since then, the logarithm of the partition coefficient of a solute between octanol and water, log *P*, has been introduced into the regression analyses of QSAR. From the magnitude of the log *P* of a compound, one can infer its ease of transport through the cell membrane and other related events. As such, relationships between biological activities and log *P* can be demonstrated. This has prompted many studies which further our understanding of this parameter.[2−4]

Accurate determination of log *P* of a compound is essential for QSAR studies and, thus, the meaningful prediction of its biological activities. However, measurement of log *P* through synthesis of the compound and its subsequent experimental determination is time consuming and costly. Hence, many methods for computing log *P*, based on a compound's chemical structure, have been proposed.[5] For example, Hansch and Leo,[6,7] divided a compound into basic fragments and calculated its log *P* by the summation of the hydrophobic contributions of these fragments. However, this method failed to work for complex compounds, and correction factors were included to improve the accuracy of the calculations. An automated version of this method, CLOGP,[8,9] has been tested on an elaborate database, Starlist, which includes nearly 8000 compounds and yields good result ($r = 0.970$, $s = 0.398$).[5] Similar approaches based on fragment constants have been put forward by Rekker,[10] Suzuki and Kudo,[11] Moriguchi,[12] and Klopman.[13,14]

An alternative approach for the computation of log *P* is based on additive atomic contributions. Broto[15] suggested that the parameters used in the calculation of log *P* can be obtained by first classifying atoms into different atom types according to their different topological environments which contribute differently to the global log *P* value; they then developed a set of 222 descriptors. Analyzing 1868 compounds, they claimed a precision of about 0.4 log unit. Later, Ghose and Crippen developed a similar procedure but used only 110 descriptors while maintaining the standard deviation of 0.4.[16−18] Although atomic contribution techniques are simple to automate, they cannot deal with long-range interactions.

Since the whole is more than the sum of its parts, any method of calculating log *P* of a molecule from its parts has limitations. Thus, other methods have been proposed based on calculated molecular properties, such as the solvatochromic approach first proposed by Kamlet and Taylor.[19−21] Later, results from quantum mechanical computations were introduced into log *P* calculations. Klopman and Iroff used atomic charges as parameters to calculate the log *P* of simple organic compounds.[22,23] In 1989, Bodor et al. derived a nonlinear model in which atomic charges, molecular surfaces, molecular volumes, ovalities, molecular weights, dipoles, and some other properties were included.[24,25] Although good results could be obtained by these methods, the applicability of such approaches needs to be verified outside the relatively small databases used in these studies. Other more theoretical approaches include those of Hopfinger and Battershell,[26] Essex et al.,[27] and Dunn and Nagy.[28] Free energy perturbation methods have begun to yield useful results for small molecules, but their extension to larger systems is limited by available computer resources.

Among all the current approaches for log *P* calculation, the group/atom contribution method is widely used because it is conceptually simple and it has given fast and accurate estimations for many organic compounds. However, this method is still far from perfect. As a matter of fact, little progress was made in the atom-additive approach since the original reports of Broto and Ghose in the 1980s. Much room is left for improvement.

In this paper, we present a new atom-additive method, XLOGP, for log *P* calculation. We classify atoms by their hybridization states and their neighboring atoms. We also include correction factors to account for some intramolecular interactions; 1831 organic compounds were analyzed by multivariate regression to derive the parameters. The result of log *P* calculation shows that our method is relatively simple and applicable to QSAR studies.

* To whom all correspondence should be addressed. Tel: 86-10-62751490. Fax: 86-10-62751725. E-mail: lai@ipc.pku.edu.cn.

[⊗] Abstract published in *Advance ACS Abstracts,* March 15, 1997.

S0095-2338(96)00169-2 CCC: $14.00 © 1997 American Chemical Society

METHODS

A relevant training database is imperative for the development of a meaningful model to calculate log *P*. First, the number and complexity of the molecules used need to be statistically sound. We have used 1831 compounds of diverse structure, from the combined collections of Suzuki and Kudo[11] and Klopman.[14] Second, the experimental log *P* values need to come from a reliable source since the literature log *P*'s were determined by various methods under different conditions and they are often different even for the same compound. Here, we have chosen the experimental log *P* values from Hansch and Leo's compilation.[29] All of the compounds and their log *P* values are summarized in the Supporting Information.

All the molecular structures were generated with SYBYL[30] on a SGI INDIGO2 workstation. The structures were then optimized by using the Tripos force field. The structures of complicated compounds were obtained by a systematic conformational search. The final structures were all stored in MOL2 format for further analyses.

We classified carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorus, and halogens in neutral organic compounds into 76 atom types according to their hybridization states and their nearest neighboring atoms. This classification differentiated the electron distribution on the atoms and the approachability of the solvent to the atoms. The atoms belonging to the same atom type were assumed to have similar values in solvent accessible surface and atomic charge. In fact, we had calculated the solvent accessible surfaces and atomic charges for each atom of each molecule in the training database which helped us in atom classification. Besides the basic 76 atom types, four functional groups, i.e., cyano, isothiocyano, nitroso, and nitro groups, were defined as pseudoatom types since these groups are all "terminal" groups and can be treated as a whole. Thus, we used a total of 80 descriptors in atom classification, as shown in Table 1. A program was developed to automatically identify the occurrence of each atom type in a molecule.

The log *P* of a molecule is assumed to be the summation of the contributions of each atom type as described by

$$\log P = \sum_i a_i A_i \qquad (1)$$

where a_i is the contribution coefficient of the *i*th atom type and A_i is the number of occurrences of the *i*th atom type. This equation ignores the possible interactions within the molecule. This model was then submitted to standard multivariate linear regression to get atomic contribution coefficients. The significance of each parameter was evaluated by analyzing their statistical *t* and *F* values.

Although the model described by eq 1 gave reasonably good results for many compounds, many others showed unacceptably large deviations. As reported previously,[9,14,17] for example, many of these compounds fell into three categories: (i) compounds having hydrophobic carbon chains, (ii) compounds existing as zwitterions, such as amino acid, and (iii) compounds having intramolecular hydrogen bonds. Therefore, to account for the interactions which affect the hydrophobicity of these compounds, we have introduced the following correction factors in our model:

(1) Hydrophobic Carbon. The hydrophobicity of compounds having hydrocarbon chains is generally underestimated from the summation of atomic contributions alone because of chain flexibility or possible aggregation of these compounds in the aqueous phase. Thus, certain compensations are required to account for these factors. We defined sp^3- and sp^2-hybridized carbons without any attached heteroatoms, i.e., atom types 1, 2, 4, 5, 8, 9, 12, 13, 18, 19, and 22, to be "hydrophobic carbons". The number of hydrophobic carbons is used as a correction factor for only aliphatic and aromatic hydrocarbons.

(2) Indicator of Amino Acid. log *P* values of amino acids have been largely overestimated (about 2 log units at an average) by the summation of simple atomic contributions because amino acids do not contain free amino and carboxylic acid groups but rather exist as zwitterions. So, when a certain compound is identified as an amino acid, this indicator will be set to 1; otherwise it will be set to 0.

(3) Intramolecular Hydrogen Bond. Intramolecular hydrogen bonds can increase the hydrophobicity of a molecule, but identifying the existence of hydrogen bonds within a compound from its chemical structure is not easy because we need to know its conformations. In our program, we have adopted the following conservative standards to define an intramolecular hydrogen bond: (i) the hydrogen donor could be a heteroatom attached to one or two hydrogens; (ii) the hydrogen acceptor could be an sp^2-hybridized oxygen, fluorine, or hydrogen donor atom defined above; (iii) the distance between the donor and the acceptor should be within a proper range; (iv) the intramolecular hydrogen bond could form a six-membered ring in the molecule. Only when all four conditions are met do we assign an intramolecular hydrogen bond to a given molecule. In this way, we take into account only "reliable" intramolecular hydrogen bonds.

(4) Halogen−Halogen 1−3 (Geminal) Interaction. When two or more halogen atoms are attached to the same atom, some relatively large changes occur because of dipole shielding. In our model, we have found that whether fluorine is involved or not determines the extent of the change. This is because fluorine is much smaller than the other halogens and is extremely electronegative. Thus, we have used two different parameters to account for halogen−halogen 1−3 (geminal) interaction (Table 2). In general, the correction factor is positive when fluorine is involved and negative when fluorine is not.

After including these correction factors, the final log *P* is described as

$$\log P = \sum_i a_i A_i + \sum_j b_j B_j \qquad (2)$$

where a_i and b_j are regression coefficients, A_i is the number of occurrences of the *i*th atom type, and B_j is the number of occurrences of the *j*th correction factor identified by our program.

RESULTS

We generated our model for log *P* calculation by analyzing 1831 compounds of diverse structures. The initial model was based on the summation of basic atomic contributions alone in which a total of 80 atom types were used. This model yielded farily satisfactory results, $n = 1831$, $r = 0.953$, SD = 0.43, $F(80,1750) = 219.6$, which were comparable to or better than those obtained by other methods using similar strategies.[14,15,17] Further analysis showed that log *P* values derived from the summation of atom contributions alone did not deviate much from their experimental counterparts for

Table 1. Atom Types Used in XLOGP

no.	description[a]	contribution	no. of compd	frequency of use	no.	description[a]	contribution	no. of compd	frequency of use
				sp^3 Carbon in					
1	CH_3R ($\pi = 0$)[b]	0.484	419	692	**10**	CHR_2X	−0.417	190	348
2	CH_3R ($\pi \neq 0$)[b]	0.168	325	414	**11**	$CHRX_2$, CHX_3	−0.454	75	75
3	CH_3X	−0.181	337	456	**12**	CR_4($\pi = 0$)	−0.378	38	44
4	CH_2R_2 ($\pi = 0$)	0.358	326	722	**13**	CR_4($\pi \neq 0$)	0.223	19	19
5	CH_2R_2 ($\pi \neq 0$)	0.009	275	340	**14**	CR_3X	−0.598	30	32
6	CH_2RX	−0.344	548	766	**15**	CR_2X_2	−0.396	4	4
7	CH_2X_2	−0.439	9	9	**16**	CRX_3	−0.699	79	86
8	CHR_3 ($\pi = 0$)	0.051	49	67	**17**	CX_4	−0.362	12	12
9	CHR_3 ($\pi \neq 0$)	−0.138	67	74					
				sp^2 Carbon in					
18	R═CH_2	0.395	41	52	**22**	R═CR_2	0.098	45	51
19	R═CHR	0.236	161	248	**23**	R═CRX, R═CX_2	−0.108	65	85
20	R═CHX	−0.166	93	104	**24**	X═CR_2, X═CXR	1.637	749	917
21	X═CHR, X═CHX	1.726	99	102	**25**	X═CX_2	1.774	244	247
				Aromatic Carbon in					
26	R···C(H)···R	0.281	1349	6032	**31**	R···C(R)···X	0.079	41	46
27	R···C(H)···X	0.142	187	278	**32**	R···C(X)···X	0.200	64	94
28	X···C(H)···X	0.715	18	18	**33**	X···C(A)···X	0.869	23	25
29	R···C(R)···R	0.302	776	1016	**34**	A···C(···A)···A[c]	0.316	120	318
30	R···C(X)···R	−0.064	1033	1832					
				sp Carbon in					
35	R≡CH	0.054	4	4	**36**	R≡CR, R≡CX, R═C═R	0.347	6	8
				Hydrogen					
37	H	0.046	1824	17899					
				sp^3 Oxygen in					
38	R−OH ($\pi = 0$)	−0.399	197	283	**41**	R−O−R	0.397	517	619
39	R−OH ($\pi \neq 0$)	−0.029	405	452	**42**	R−O−X, X−O−X	0.068	6	15
40	X−OH	−0.330	11	11	**43**	π−O−π (in ring)[d]	0.327	12	12
				sp^2 Oxygen in					
44	O═R	−2.057	863	1041	**45**	O═X	0.218	90	185
				sp^3 Nitrogen in					
46	R−NH_2 ($\pi = 0$)	−0.582	58	60	**50**	R−NH−X, X−NH−X	−0.381	38	39
47	R−NH_2 ($\pi \neq 0$)	−0.449	178	192	**51**	NR_3	0.443	65	75
48	X−NH_2	−0.774	49	49	**52**	NR_2X, NRX_2, NX_3	−0.117	28	31
49	R−NH−R	0.040	49	50					
				sp^2 Nitrogen in					
53	R═NH, R═NR	−2.052	122	129	**55**	X═NR	0.321	15	19
54	R═NX	−1.716	70	82	**56**	X═NX	−0.921	14	17
				Aromatic Nitrogen					
57	A···N···A[e]	−0.704	239	301					
				Trigonal Planar Nitrogen in[f]					
58	R−NH−R	0.119	9	9	**61**	NA_3	0.587	3	3
59	R−NH−X, X−NH−X	1.192	37	37	**62**	NA_3 (in ring)[g]	0.668	61	61
60	A−NH−A (in ring)[g]	0.434	64	65					
				Amide Nitrogen in					
63	−NH_2	−0.791	93	99	**65**	−NR_2, −NRX	0.016	94	103
64	−NHR, −NHX	−0.212	289	337					
				sp^3 Sulfur in					
66	A−SH	0.752	5	5	**68**	π−S−π (in ring)[h]	0.964	23	23
67	R−S−R, R−S−X	1.071	44	45					
				sp^2 Sulfur					
69	S═R	−1.817	14	14					
				Sulfoxide Sulfur					
70	A−SO−A	−1.214	5	5					
				Sulfone Sulfur					
71	A−SO_2−A	−0.778	81	88					
				Fluorine					
72	F	0.493	140	333					
				Chlorine					
73	Cl	1.010	233	409					
				Bromine					
74	Br	1.187	86	109					
				Iodine					
75	I	1.489	39	40					
				Phosphorus					
76	A−PO(A)−A	−0.802	4	4					
				Terminal Groups					
77	−CN	−0.256	79	84	**79**	−NO	0.077	37	40
78	−NCS	1.626	23	24	**80**	−NO_2	0.264	135	154

[a] Definitions: −, single bond; ═, double bond; ≡, triple bond; ···, aromatic bond; R, any group linked through carbon; X, any heteroatom (O, N, S, P, and halogens); A, any atom except hydrogen, i.e., R or X; π, any atom involved in a conjugated system, such as sp^3- and sp^2-hybridized atoms and aromatic atoms. [b] $\pi = 0$ represents that no π atom is a neighboring atom, while $\pi \neq 0$ represents that this atom is connected to a conjugated system. [c] The joint aromatic carbon in polycyclic aromatic systems. [d] As in a furan ring. [e] This only represents the nitrogen atom in a six-membered aromatic ring, such as in a pyridine ring. [f] This is a special atom type adopted by the Tripos force field. When a nitrogen is connected with two or three π atoms, i.e., π−NH−π or π−N(A)−π, it adopts a trigonal planar geometric structure instead of a tetrahedron. [g] As in a pyrrole ring or some other five-membered ring. [h] As in a thiophene ring.

Table 2. Correction Factors Used in XLOGP

description[a]	contributions	no. of affected compd	standard deviation before using correction	standard deviation after using correction
hydrophobic carbon	0.19	88	0.53	0.19
amino acid indicator	−2.27	14	1.53	0.36
intramolecular H-bond	0.60	147	0.59	0.44
halogen 1−3 pair (F−F, F−X)[b]	0.08	84	0.54	0.46
halogen 1−3 pair (X−X)[b]	−0.26	20	0.56	0.37

[a] Detailed description can be found in the Methods section in the article. [b] X represents Cl, Br, and I.

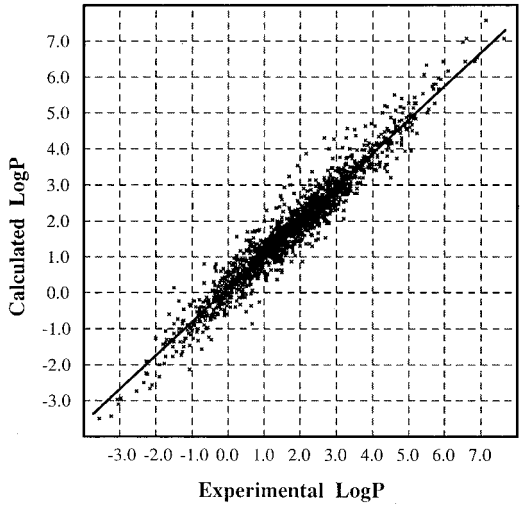


Figure 1. Correlation between the experimental and calculated log *P* values of 1831 organic compounds.

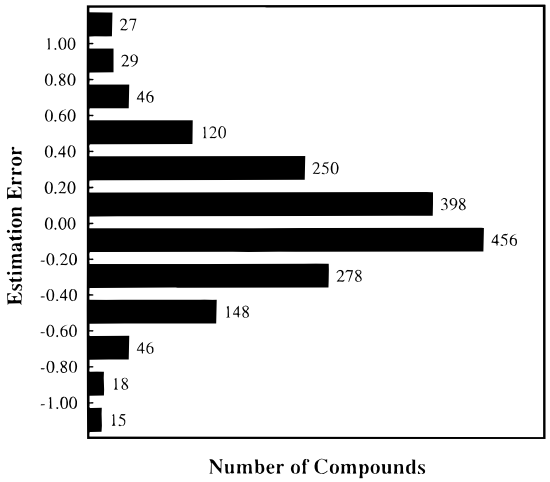


Figure 2. Distribution histogram of the estimation errors.

monofunctional compounds. However, large errors were found with certain classes of compounds such as hydrocarbons, amino acids, and chemicals with multiple functional groups. In such cases, correction factors were needed to account for the possible intramolecular interactions.

We included five correction factors in our final model, as shown in Table 2. All of these correction factors were found to be statistically significant in multivariate regression analysis. The final model for log *P* calculation was obtained by correlating these 85 descriptors with the experimental log *P* values. The regressional coefficients for each atom type and the correction factors are listed in Tables 1 and 2, respectively. This final model yielded better results than those obtained with eq 1: $n = 1831$, $r = 0.968$, SD = 0.37, $F(85,1745) = 301.0$. The standard deviation of 0.37 is within the experimental error of 0.4.

Figure 1 shows the correlation between the experimental and calculated log *P* values. The slope and intercept of the regressional line are 0.94 and 0.11, respectively. The histogram of the estimation errors is shown in Figure 2, where a near-Gaussian error distribution curve centered at zero can be seen. The experimental and calculated log *P* values of the 1831 compounds are summarized in the Supporting Information. Furthermore, we have performed leave-one-out cross-validation on the whole training database, which gives approximately the same result as multivariate regressional analysis ($n = 1831$, $r = 0.963$, SD = 0.40). The result of this cross-validation test demonstrates the applicability of our model for log *P* calculation.

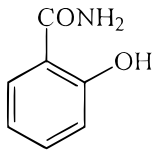
A computer program in C language, XLOGP, has been developed based on our final model. To use this program to calculate the log *P* of a compound, one only needs to input the molecular structure prepared in the MOL2 file format. The program can then perform atom classification and assign correction factors, and using the coefficients listed in Tables 1 and 2, it will calculate log *P* and provide detailed information on the calculation of the submitted molecule. When we run the program on an SGI INDIGO2 workstation, the whole computing process is typically within 1 s/molecule. A typical XLOGP output is shown in Figure 3.

DISCUSSION

Atom Classification. Any method which quantitates the hydrophobicity of a compound by the summation of the characteristics of its parts requires a relevant method for group/atom classification. Good group/atom descriptors should discriminate between the significant contributions from those that are not. The quality of these descriptors can be evaluated by how well the computed log *P* values agree with their experimental counterparts. Thus, we need to define a set of relevant descriptors and compare the resultant log *P*'s with the experimental log *P*'s to establish the usefulness of our model.

We have developed a set of 80 descriptors in our model which is smaller than Ghose's set of 110[17] and much smaller than Broto's set of 222.[15] However, using less parameters does not weaken the power of our model which yields satisfactory results even when we use the addition of basic atomic contributions alone. Our classification of the atom is based on the type of element, its hybridization state, and its nearest neighboring atoms. In fact, we have used the atom types defined in the Tripos force field as our starting point and then developed each of them into a more elaborate class for log *P* calculation. The result shows that the descriptors are well defined and the smaller number in the set has made programming easier.

The coefficients obtained by different methods of log *P* calculation usually cannot be compared because these parameters depend greatly on how the group/atom is clas-



salicylamide measured logP = 0.89

Menu for LogP calculation:

No.	Atom type in SYBYL	Atom type in XLOGP	Contribution
1	C.ar	29	0.30
2	C.ar	26	0.28
3	C.ar	26	0.28
4	C.ar	26	0.28
5	C.ar	26	0.28
6	C.ar	30	-0.06
7	C.2	24	1.64
8	O.2	44	-2.06
9	N.am	63	-0.79
10	O.3	39	-0.03
11	H	37	0.05
12	H	37	0.05
13	H	37	0.05
14	H	37	0.05
15	H	37	0.05
16	H	37	0.05
17	H	37	0.05
Hydrophobic carbon:			0.00
Amino acid indicator:			0.00
Intramolecular hydrogen bond:			0.60
F_F, F_X 1-3 pair (X=Cl,Br,I):			0.00
X_X 1-3 pair (X=Cl,Br,I):			0.00

Final LogP= 1.04

Figure 3. Example of XLOGP's output.

sified. For instance, in our model the hydrophobic contribution of hydrogen is about 0.05, which is much lower than that reported by Rekker (0.18) or Leo (0.23), but some conclusions agreeing with other approaches can still be drawn. The carbon atoms in $-CH_3$, $-CH_2-$, and $-CH\langle$ are generally more hydrophobic than the $\rangle C\langle$ atoms. In addition, the presence of fluorine, chlorine, bromine, and iodine increases the hydrophobicity, which has a positive contribution on log *P*, while the presence of nitrogen and oxygen atoms generally decreases the hydrophobicity, which has a negative contribution on log *P*. Moreover, sp^3-hybridized oxygen and nitrogen, i.e., atom types *o* and ψ, become more hydrophobic when they are attached to conjugated systems.

Are methods using fragmental contributions more accurate than those using atomic contributions? Our result suggests that the accuracy of log *P* estimation of the two methods is comparable, but the method based on atomic addition are simpler to automate. They use fewer parameters than fragmental methods, and in principle, "atom typing" can never be exhaustive to describe the infinite variety of structures present in organic molecules. Furthermore, "atom typing" does not produce ambiguous results, inherent in fragmental methods, due to the different ways of dissecting a molecule or the different interpretation of the correction rules. In some applications of hydrophobic parameters such as the molecular lipophilicity potential (MLP) approaches,[31−33] the use of atom-centered parameters is preferred.

Correction Factors. The calculation by eq 1 illustrates the limitation of the initial model which does not consider the interactions within and among the molecules. When we introduce five correction factors for some intramolecular interactions, the correlation between the calculated log *P* and the experimental values for certain classes of compounds,

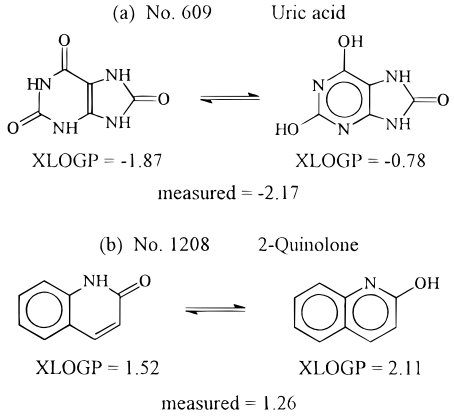


Figure 4. Tautomerization effect.

Table 3. List of Compounds with a Large Calculation Error

no.	log P_{exptl}[a]	log P_{calcd}[b]	error	name
241	2.69	4.31	1.62	α,α-diphenylpropionic acid
367	−4.41	−3.38	1.03	ornithine
443	−1.05	−2.12	−1.07	Ado
525	1.96	3.55	1.59	2,4,5-tribromoimidazole
558	−1.62	−0.57	1.05	2-pyrimidone
569	−1.34	−0.23	1.11	2,3-butanedione
577	0.52	−0.53	−1.05	2-methyl-2-imidazoline
596	3.53	4.56	1.03	2,3,4,5,6-pentachloropyridine
621	−0.35	0.72	1.07	8-valerolactone
638	3.23	2.10	−1.13	pentafluorophenol
669	2.36	1.33	−1.03	3,5-dinitrophenol
689	−1.50	0.14	1.63	picolinic acid
713	0.65	−0.36	−1.01	3-hydroxypicolinamide
737	0.88	−0.24	−1.12	2-aminonicotinamide
738	0.70	−0.84	−1.54	6-aminonicotinamide
792	1.48	2.71	1.23	1-phenyl-4-chlorotetrazole
877	0.26	1.59	1.33	benzohydroximic acid
967	3.89	1.99	−1.90	2-(trifluoromethyl)-5,6-dinitro-benzimidazole
968	2.83	4.03	1.20	4,5,6,7-tetrachloro-2-methyl-benzimidazole
984	4.50	3.21	−1.29	*m*-(trifluoromethyl)trifluoro-methane sulfonanilide
1344	2.73	1.56	−1.17	hexamethylmelamine
1389	0.55	1.73	1.18	2-(*N*,*N*-dimethylamino)-6-(5-nitro-2-furyl)-1,3-thiazin-4-one
1394	2.52	1.50	−1.02	benzoylacetone
1437	0.68	1.73	1.05	fusaric acid
1489	5.08	3.96	−1.12	1-(2-Cl-5-CF_3-phenylhydra-zono)-1-cyanoacetone
1555	0.75	1.76	1.01	2,3,6-trimethyl-4-hydroxy-acetanilide
1595	1.34	2.37	1.03	benzidine
1604	−0.33	0.82	1.15	sulfsomidine
1643	−0.65	0.80	1.45	sotalol
1649	4.40	3.37	−1.03	4-isothiocyanodiphenyl sulfoxide
1650	1.59	2.95	1.36	niflumic acid
1725	1.70	3.24	1.54	chlorambucil
1740	1.74	3.03	1.29	apigenim
1743	2.36	3.37	1.01	1-methyl-4-phenyl-7-chloro-quinazolin-2-one
1755	1.52	2.61	1.09	sulfaphenazole
1758	3.32	4.59	1.27	bisphenol
1777	2.09	0.88	−1.21	(phenoxymethyl)penicillin
1778	1.40	0.38	−1.02	(α-hydroxybenzyl)penicillin
1810	3.27	4.33	1.06	phenolphthalol
1816	2.18	3.88	1.70	buquinolate
1824	1.07	2.32	1.25	pipamperone
1826	2.96	4.28	1.32	flupentixol

[a] log P_{exptl} is the experimental value. [b] log P_{calcd} is the calculated value.

Table 4. Experimental and Calculated log *P* Values of 19 Drugs

no.	drug	log P_{expt}[a]	XLOGP	Moriguchi[b]	Rekker[b]	Hansch−Leo[b]	Suzuki−Kudo[b]
1	atropine	1.83	2.29	2.21	1.88	1.32	0.03
2	chloramphenicol	1.14	1.46	1.23	0.32	0.69	−0.75
3	chlorothiazide	−0.24	−0.58	−0.36	−0.68	−1.24	−0.44
4	chlorpromazine	5.19	4.91	3.77	5.10	5.20	3.89
5	cimetidine	0.40	0.20	0.82	0.63	0.21	3.33
6	diazepam	2.99	2.98	3.36	3.18	3.32	1.23
7	diltiazem	2.70	3.14	2.67	4.53	3.55	1.96
8	diphenhydramine	3.27	3.74	3.26	3.41	2.93	3.35
9	flufenamic acid	5.25	4.45	3.86	5.81	5.58	5.16
10	haloperidol	4.30	4.35	4.01	3.57	3.52	3.43
11	imipramine	4.80	4.26	3.88	4.43	4.41	3.38
12	lidocaine	2.26	2.47	2.52	2.30	1.36	0.91
13	phenobarbital	1.47	1.77	0.78	1.23	1.37	1.29
14	phenytoin	2.47	2.23	1.80	2.76	2.09	2.01
15	procainamide	0.88	1.27	1.72	1.11	1.11	0.65
16	propranolol	2.98	2.98	2.53	3.46	2.75	2.15
17	tetracaine	3.73	2.73	2.64	3.55	3.65	2.90
18	trimethoprim	0.91	0.72	1.26	−0.07	0.66	0.57
19	verapamil	3.79	5.29	3.23	6.15	3.53	6.49

[a] The experimental log *P* values are cited from ref 29. [b] The calculated log *P* values by different methods are cited from ref 34.

such as hydrocarbons and amino acids, is much improved (Table 2). Overall, the standard deviation of the entire training set is reduced to under 0.4, and the *r* value is increased from 0.953 to 0.968.

We have introduced the concept of "hydrophobic carbon" to correct for the underestimated values of aliphatic and aromatic hydrocarbons. The hydrophobic carbons in a hydrocarbon molecule are given positive compensations. The rationale behind this correction factor is that hydrocarbons have special properties in the aqueous phase as compared to the compounds with heteroatoms. We find this correction factor works very well for hydrocarbons. However, for other heteroatom-containing series, such as alcohol, amine, and carboxylic acid, we have not seen the same trend. Inclusion of this correction factor does not improve the results. Additional considerations are required to cope with the hydrophobic portions of a molecule.

The correction for intramolecular hydrogen bonds is important in our model. We find that a correction factor of 0.60 is very close to that reported by Leo,[5] 0.63. An intramolecular hydrogen bond is strictly defined in our model, i.e., the atoms should be able to form a six-membered ring within the molecule. This ensures that the formation of an intramolecular hydrogen bond is not only possible but reliable. We have tried other definitions in our analysis but they do not work as well as this one.

Tautomerization effect affects the results of our calculation of log *P*. As illustrated in Figure 4, when the structure submitted for calculation is not of the dominant tautomer, the resultant log *P* will deviate from the experimental value. Thus, the ratio of the two tautomers at equilibrium is needed for the accurate prediction of log *P* for such compounds. However, we have not introduced any additional correction factor for tautomerization to our model. Instead, we have identified all the compounds in the training database which may undergo tautomerization and dealt with the predominant tautomers only.

In our current model, we have included five correction factors which compensate for some estimation errors in certain compounds. The method works well for most of the 1831 compounds tested, but as listed in Table 3, 42 compounds in the training database showed deviations greater than or equal to 1.0 log unit ($1.0 \approx 3\sigma$). We do not have an explanation now but are currently working on developing more elaborate sets of correction factors to improve the model.

Comparison to Other Methods. To compare different methods for log *P* calculation, experts in the field may need to come together and work out a list of compounds as the standard "test set". But, accuracy alone cannot tell the whole story. Some other features, such as efficiency and practicality, should be considered to evaluate a method of log *P* estimation as well.

Two atom-additive approaches have been reported before ours: one by Broto et al.[15] and the other by Ghose and Crippen.[16−18] Among the 1868 compounds used by Broto et al., none possesses the potential to form internal hydrogen bonds. This greatly simplifies the calculation but weakens the power of their model. They have developed a set of 222 descriptors, but some descriptors are statistically insignificant for such a limited database. Ghose and Crippen have reduced the descriptors to 110 while maintaining a standard deviation of 0.4, but their training database is only one-half the size of ours (830 vs 1831). Compared with these two approaches, our model uses less descriptors but yields better results for more diverse structures. Since they have not reported any improvements on their methods, we merit ours an alternative, if not a better, approach for log *P* calculation.

In a recent article, Moriguchi has compared the reliability of several fragmental methods using a list of 22 compounds.[34] We tested XLOGP on 19 of these compounds sicne three of them, disopyramide, propafenone, and furosemide, do not have reliable measured values according to Leo.[35] The measured and calculated log *P* values by these different methods are listed in Table 4. The correlations between the measured and calculated values are as follows: XLOGP, $n = 19$, $r = 0.942$, $s = 0.52$; Moriguchi's method, $n = 19$, $r = 0.933$, $s = 0.53$; Rekker's method, $n = 19$, $r = 0.917$, $s = 0.77$; Hansch−Leo's method, $n = 19$, $r = 0.970$, $s = 0.42$; Suzuki−Kudo's method, $n = 19$, $r = 0.737$, $s = 1.25$.

The above data show that XLOGP is the second best among the five methods. But we must bear in mind that 19 compounds are not statistically abundant for such a test. Changing just one or two samples will change the outcome of the comparison. In fact, XLOGP yielded acceptable estimations for all 19 molecules except verapamil. When

this compound is removed from the current list, the calculation shows that XLOGP gives the best result among all these methods.

Compared to CLOGP, the well-established method by Hansch and Leo, our method is simpler and easier to use. XLOGP is a single program written in C language. Therefore, it can be easily distributed and installed on different types of computers. We use a readily available software, SYBYL, to prepare the input files for XLOGP. This has freed us from writing programs for topology generation, and this will also free the user from learning unfamiliar rules to use XLOGP. Besides SYBYL users, anyone who can prepare the structures of their molecules in the MOL2 format can perform XLOGP calculations. After which, the input structures can be directly incorporated into further QSAR analysis such as CoMFA in SYBYL. This strategy has worked very well in our lab.

CONCLUSION

We have developed a new automated method, XLOGP, for calculating the partition coefficients of solutes in octanol/water, log *P*. We calculate log *P* of a molecule by the summation of its atomic contributions, and we account for some intramolecular interactions by including the respective correction factors. In this model, 80 descriptors and five correction factors are used to classify the various atoms in a molecule. The final log *P* model is then derived from multivariate regression analysis on 1831 organic compounds of diverse structures. Compared to other methods, ours gives either comparable or better estimations of log *P* values. Other appealing features of our method are its ease of computation and its compatibility with the popular software SYBYL.

ACKNOWLEDGMENT

The authors thank the Chinese State Commission of Science and Technology for their financial support and Prof. Renli Li at Beijing Medical University for his kind offer of references.

Supporting Information Available: Table of the training database of XLOGP, containing 1831 organic compounds of diverse structures, in MOL2 format and the measured log *P* of these compounds (37 pages). See any current masthead page for ordering and Internet access instructions. The program XLOGP as well as all the compounds stored in MOL2 format used in the training database is available by contacting the authors.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(2) Le, A. The Octanol-Water Partition Coefficient of Atomatic Solutes: the Effect of Electronic Interactions, Alkyl Chains, Hydrogen Bonds, and ortho-Substitution. *J. Chem. Soc., Perkin Trans. II* **1983**, *2*, 825−838.
(3) Hansch, C.; Bjorkroth, J. P.; Leo, A. Hydrophobicity and Central Nervous System Agents: the Principle of Minimal Hydrophobicity in Drug Design. *J. Pharm. Sci.* **1987**, *76*, 663−687.
(4) Hansch, C.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hamment Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178−180.
(5) Leo, A. Calculating $\log P_{oct}$ from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(6) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and their Uses. *Chem. Rev.* **1971**, *71*, 525−616.
(7) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(8) Chou, J. T.; Jurs, P. C. Computer-assisted computation of partition coefficient from molecular structures using fragment constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172−178.
(9) Leo, A. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon: Oxford, 1990; Vol. 4, pp 295−319.
(10) Rekker, R. F. *The Hydrophobic Fragment Constant*; Elsevier: New York, 1977.
(11) Suzuki, T.; Kudo, Y. Automatic logP estimation based on combined additive modeling methods. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 155−198.
(12) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(13) Klopman, G.; Wang, S. A Computer Automated Structure Evaluation (CASE) Approach to Calculation of Partition Coefficient. *J. Comput. Chem.* **1991**, *12*, 1025−1032.
(14) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated logP Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(15) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.* **1984**, *19*, 71−78.
(16) 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.
(17) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. III. Modeling Hydrophobic Interactions. *J. Comput. Chem.* **1988**, *9*, 80−90.
(18) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(19) Kamlet, M.; Doherty, R.; Abboud, J.-L.; Abraham, M.; Taft, R. Linear Solvation Energy Relationships: 36. Molecular Properties Governing Solubilities of Organic Nonelectrolytes In Water. *J. Pharm. Sci.* **1986**, *75*, 338−349.
(20) Kamlet, M.; Doherty, R.; Abraham, M.; Marcus, Y.; Taft, R. Linear Solvation Energy Relationships. 46. An Improved Equation for Correlation and Prediction of Octanol/Water Partition Coefficients of Organic Nonelectrolytes (Including Strong Hydrogen Bond Donor Solutes) *J. Phys. Chem.* **1988**, *92*, 5244−5255.
(21) Leahy, D.; Morris, J.; Taylor, P.; Wait, A. Model Solvent Systems for QSAR. Part 3. An LSER Analysis of the 'Critical Quartet' New Light on Hydrogen Bond Strength and Directionality. *J. Chem. Soc., Perkin Trans. II* **1992**, *2*, 705−722.
(22) Klopman, G.; Iroff, L. D. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem.* **1981**, *2*, 157−160.
(23) Klopman, G.; Namboordiri, K.; Schochet, M. Simple Method of Computing the Partition Coefficient. *J. Comput. Chem.* **1985**, *6*, 28−38.
(24) Bodor, N.; Gabanyi, Z.; Wong, C. K. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(25) Bodor, N.; Huang, M. J. An extended version of a novel method for the estimation of partition coefficients. *J. Pharm. Sci.* **1992**, *81*, 272−281.
(26) Hopfinger, A. J.; Battershell, R. D. Application of SCAR to drug design 1. Prediction of octanol-water partition coefficients using solvent-dependent conformational analyses. *J. Med. Chem.* **1976**, *19*, 569−573.
(27) Essex, J. W.; Reynolds, C. A.; Richards, W. G. Theoretical Determination of Partition Coefficients. *J. Am. Chem. Soc.* **1992**, *114*, 3634−3639.
(28) Dunn, W. J., III; Nagy, P. I. Relative LogP and Solution Structure for Small Organic Solutes in the Chloroform/Water System Using Monte Carlo Methods. *J. Comput. Chem.* **1992**, *13*, 468−477.
(29) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: hydrophobic, electronic, and steric constants*; American Chemistry Society: Washington, DC, 1995; Vol. 2.
(30) SYBYL 6.2, Tripos Assoc. Inc., St. Louis, MO, 1995.
(31) Furet, P.; Sele, A.; Cohen, N. C. 3D molecular lipophilicity potential profiles: a new tool in molecular modeling. *J. Mol. Graph.* **1988**, *6*, 182−189.
(32) Abraham, D. J.; Kellogg, G. E. The effect of physical organic properties on hydrophobic fields. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 41−49.
(33) Gaillard, P.; Carrupt, P.-A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a tool in 3D QSAR: Method and applications. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83−96.
(34) Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of Reliability of logP Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42*, 976−978.
(35) Leo, A. Critique of Recent Comparison of logP Calculation Methods. *Chem. Pharm. Bull.* **1995**, *43*, 512−513.

CI960169P