Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods

Arup K. Ghose,* Vellarkad N. Viswanadhan, and John J. Wendoloski

Department of Molecular Structure and Design, Amgen Inc., 1840 DeHavilland Drive, Thousand Oaks, California 91320

Received: November 17, 1997; In Final Form: March 4, 1998

Molecular hydrophobicity (lipophilicity), usually quantified as log *P* (the logarithm of 1-octanol/water partition coefficient), is an important molecular characteristic in drug discovery. ALOGP and CLOGP are two of the most widely used methods for the estimation of log *P*. This work describes an extensive reparametrization of the atomic log *P* values and a detailed comparison of the performance of ALOGP and CLOGP methods on the Pomona Medchem database. Only the *"star list"* compounds having precisely measured log *P* values were used in this analysis. While the overall results with both methods are similar, analysis shows that the CLOGP method is better for very small molecules in the range of 1−20 atoms. The two methods are almost comparable in the range of 21−45 atoms, while the ALOGP method has better accuracy for molecules with more than 45 atoms. Although the rms deviation and the correlation coefficient for CLOGP predictions were marginally better than those for corresponding ALOGP predictions, the latter showed a very stable performance for all classes of organic compounds analyzed. The ALOGP method can be used to compute estimates of most neutral organic compounds having C, H, O, N, S, Se, P, B, Si, and halogens. It also covers most zwitterionic compounds having amine and carboxylic acids and ammonium halide salts. The CLOGP method has improved considerably over the years to cover most neutral organic compounds, but it still has some undefined fragments. Finally, unlike CLOGP and other methods of predicting lipophilicity, the ALOGP method has multiple uses, such as the estimation of local hydrophobicity, the visualization of molecular hydrophobicity maps, and the evaluation of hydrophobic interactions in protein−ligand complexes.

Introduction

The logarithm of the 1-octanol/water partition coefficient (log *P*) is a well-known measure of molecular hydrophobicity (also known as lipophilicity).[1−4] It is used to assess biological properties relevant to drug action, such as lipid solubility, tissue distribution, receptor binding, cellular uptake, metabolism, and bioavailability. Hansch[5] pioneered the extensive use of this parameter in developing variables for quantitative structure−activity relationship (QSAR) equations. 1-Octanol is a natural choice as the hydrophobic solvent in this respect because of its physicochemical similarity to lipids, its ready availability, and its ease of use.

The advent of high throughput synthesis and combinatorial chemistry has created an increasing demand for fast methods for the accurate assessment of log *P* and other relevant molecular properties prior to organic synthesis. This is necessary to achieve the maximal diversity of combinatorial libraries with a minimum number of compounds. Tremendous efforts in the past by theoretical and computational chemists led to several useful computational methods for estimating log *P* values of organic compounds.[2,6−12] Among these, the CLOGP[7] and the ALOGP[10,11,13] methods are the most widely used, owing to two fundamentally different reasons. The CLOGP[7] method was simply the first to be implemented in commercially available software. The ALOGP method[10,11,13] is a later development but is easier to computerize. It not only yields reasonably accurate estimates of log *P* values[10,11,14] but also is applicable in several related areas of computational and medicinal chemistry, including 3D-QSAR applications,[15−19] generation of lipophilic surfaces of molecules,[20−22] and evaluation of hydrophobic interactions in protein−ligand complexes.[22] Though very popular, the ALOGP method[10,11,13] has not been reparametrized since its last revision in 1989.[11]

There have been some attempts in the literature to compare various log *P* computational methods. However, these comparisons were made on small datasets by one of the original authors, and the "best" method often turned out to be the author's own![14,23,24] A recent book[4] provides an excellent overview of the fundamentals and physical chemistry of octanol−water partition coefficients containing experimental, theoretical, and conceptual aspects. This book[4] also collected all the important works that dealt with the predictive performance of various methods. However, even these reviewed comparisons were based on limited datasets. Clearly, tests should include all organic compounds with accurately determined log *P* values for a thorough and objective evaluation. Furthermore, it is also important to classify the database into organic subclasses such as aldehydes, ketones, and so forth, so that the user can be aware of the quality of log *P* predictions for each subclass. We, therefore, employed the largest available database with accurately determined log *P* values (the star list of the MedChem database[25]) and developed procedures for the comparison of the two widely used methods: ALOGP[10,11,13] and CLOGP.[7] The objectives of the present work are to extend and to reevaluate the ALOGP parameters to cover neutral organic molecules of medicinal interest and some charged compounds including unblocked peptides including the amino acids, using a large database of compounds, and to present an analysis of the ALOGP and CLOGP methods that can be a benchmark for evaluating such computational methods.

S1089-5639(98)00230-8 CCC: $15.00
Published on Web 05/05/1998

TABLE 1: Classification of Atoms and Their Contributions to Octanol−Water Partition Coefficient as a Measure of Hydrophobicity

type	description[a]	hydrophobicity[b]	no. of compds	no. of freq of use	type	description[a]	hydrophobicity[b]	no. of compds	no. of freq of use
	C in					N in			
1	:CH_3R, CH4	−1.5603	4088	7642	66	:Al−NH_2	−0.5427	138	141
2	:CH_2R_2	−1.0120	2941	7639	67	:Al_2NH	−0.3168	166	169
3	:CHR_3	−0.6681	848	1406	68	:Al_3N	0.0132	457	481
4	:CR_4	−0.3698	403	527	69	:Ar−NH_2, X−NH_2	−0.3883	687	859
5	:CH_3X	−1.7880	2279	3463	70	:Ar−NH−Al	−0.0389	196	239
6	:CH_2RX	−1.2486	3529	6957	71	:Ar−NAl_2	0.1087	198	218
7	:CH_2X_2	−1.0305	156	162	72	:RCO−N<, >N−X═X	−0.5113	3717	5526
8	:CHR_2X	−0.6805	1640	2846	73	:Ar_2NH, Ar_3N	0.1259	1251	1264
9	:$CHRX_2$	−0.3858	301	315		:Ar_2N−Al, R⋯N⋯R[f]			
10	:CHX_3	0.7555	30	30	74	:R≡N, R═N−	0.1349	1070	1293
11	:CR_3X	−0.2849	436	508	75	:R- -N- -R,[g]R- -N- -X	−0.1624	1927	3083
12	:CR_2X_2	0.0200	82	96	76	:Ar−NO_2. R- -N(- -R)- -O[h]	−2.0585	867	969
13	:CRX_3	0.7894	275	305		RO−NO,			
14	:CX_4	1.6422	44	44	77	:Al−NO_2	−1.9150	21	21
15	:═CH_2	−0.7866	134	153	78	:Ar−N═X, X−N═X	0.4208	270	392
16	:═CHR	−0.3962	794	1307	79	:N^+ (positively charged)	−1.4439	189	189
17	:═CR_2	0.0383	444	530	80	unused	−	−	−
18	:═CHX	−0.8051	260	286		F attached to			
19	:═CRX	−0.2129	314	417	81	:$C^1_{sp^3}$	0.4797	107	115
20	:═CX_2	0.2432	57	59	82	:$C^2_{sp^3}$	0.2358	30	84
21	:≡CH	0.4697	72	88	83	:$C^3_{sp^3}$	0.1029	289	893
22	:≡CR, R═C═R	0.2952	77	79	84	:$C^1_{sp^2}$	0.3566	278	341
23	:≡CX	−	−	−	85	:$C^{2-4}_{sp^2}$, C^1_{sp}, $C^4_{sp^3}$, X	0.1988	18	22
24	:R- -CH- -R	−0.3251	6046	27607		Cl attached to			
25	:R- -CR- -R	0.1492	4086	6356	86	:$C^1_{sp^3}$	0.7443	154	253
26	:R- -CX- -R	0.1539	4328	8624	87	:$C^2_{sp^3}$	0.5337	41	76
27	:R- -CH- -X	0.0005	1010	1384	88	:$C^3_{sp^3}$	0.2996	58	195
28	:R- -CR- -X	0.2361	860	1089	89	:$C^1_{sp^2}$	0.8155	1044	1939
29	:R- -CX- -X	0.3514	545	653	90	:$C^{2-4}_{sp^2}$, C^1_{sp}, $C^4_{sp^3}$, X	0.4856	121	145
30	:X- -CH- -X	0.1814	152	154		Br attached to			
31	:X- -CR- -X	0.0901	133	146	91	:$C^1_{sp^3}$	0.8888	36	43
32	:X- -CX- -X	0.5142	296	429	92	:$C^2_{sp^3}$	0.7452	4	5
33	:R- -CH⋯X	−0.3723	428	478	93	:$C^3_{sp^3}$	0.5034	6	15
34	:R- -CR⋯X	0.2813	673	789	94	:$C^1_{sp^2}$	0.8995	213	257
35	:R- -CX⋯X	0.1191	302	314	95	:$C^{2-4}_{sp^2}$, C^1_{sp}, $C^4_{sp^3}$, X	0.5946	29	35
36	:Al−CH═X	−0.1320	39	39		I attached to			
37	:Ar−CH═X	−0.0244	115	115	96	:$C^1_{sp^3}$	1.4201	10	12
38	:Al−C(═X)−Al	−0.2405	55	65	97	:$C^2_{sp^3}$	1.1472	1	2
39	:Ar−C(═X)−R	−0.0909	434	492	98	:$C^3_{sp^3}$	−	−	−
40	:R−C(═X)−X, R−C≡X, X═C═X	−0.1002	4126	5998	99	:$C^1_{sp^2}$	0.7293	83	116
41	:X−C(═X)−X	0.4182	1399	1527	100	:$C^{2-4}_{sp^2}$, C^1_{sp}, $C^4_{sp^3}$, X	0.7173	15	17
42	:X- -CH⋯X	−0.2147	381	383		halide ions			
43	:X- -CR⋯X	−0.0009	247	261	101	:fluoride ion	−	−	−
44	:X- -CX⋯X	0.1388	205	261	102	:chloride ion	−2.6737	3	3
45	unused	−	−	−	103	:bromide ion	−2.4178	1	1
	H attached to[c]				104	:iodide ion	−3.1121	2	2
46	:$C^0_{sp^3}$ having no X attached to next C	0.7341	2870	19673	105	unused	−	−	−
47	:$C^1_{sp^3}$, $C^0_{sp^2}$	0.6301	7785	53484		S in			
48	:$C^2_{sp^3}$, $C^1_{sp^2}$, C^0_{sp}	0.5180	1056	1370	106	:R−SH	0.6146	23	27
49	:$C^3_{sp^3}$, $C^{2-3}_{sp^2}$, C^{1-3}_{sp}	−0.0371	1599	2247	107	:R_2S, RS−SR	0.5906	675	774
50	:heteroatom	−0.1036	5456	10957	108	:R═S	0.8758	204	211
51	:α-C[d]	0.5234	3072	8801	109	:R−SO−R	−0.4979	58	59
52	:$C^0_{sp^3}$, having 1 X attached to next carbon	0.6666	2891	13030	110	:R−SO_2−R	−0.3786	560	631
53	:$C^0_{sp^3}$, having 2 X attached to next carbon	0.5372	370	952		Si in			
54	:$C^0_{sp^3}$, having 3 X attached to next carbon	0.6338	79	164	111	:>Si< as in silicones	1.5188	11	11
55	:$C^0_{sp^3}$, having 4 or more X attached to next carbon	0.3620	2	4		B in			
	O in				112	:>B^- as in boranes	1.0255	2	2
56	:alcohol	−0.3567	953	1477	113−114	unused	−		
57	:phenol, enol, carboxyl OH	−0.0127	1048	1239		P in			
58	:═O	−0.0233	5138	8248	115	:ylids	−	−	−
59	:Al−O−Al	−0.1541	575	821	116	:R_3−P═X	−0.9359	2	2
60	:Al−O−Ar, Ar_2O	0.0324	2924	4188	117	:X_3−P═X (phosphate)	−0.1726	131	132
	:R⋯O⋯R, R−O−C═X				118	:PX_3 (phosphite)	−0.7966	3	3
61[e]	:- -O	1.0520	902	1928	119	:PR_3 (phosphine)	0.6705	2	2
62	:O^- (negatively charged)	−0.7941	182	363	120	:C−$P(X)_2$═X (phosphonate)	−0.4801	20	20
63	:R−O−O−R	0.4165	8	16					
	Se in								
64	:Any−Se−Any	0.6601	6	6					
65	:═Se	−	−	−					

[a] R represents any group linked through carbon; X represents any heteroatom (O, N, S, P, Se, and halogens); Al and Ar represent aliphatic and aromatic groups, respectively; ═ represents a double bond; ≡ represents a triple bond; - - represents a aromatic bonds as in benzene or delocalized bonds such as the N−O bond in a nitro group; ⋯ represents aromatic single bonds as the C−N bond in pyrrole. [b] Atomic hydrophobicity in the unit of log *P*(octanol−water). [c] The subscript represents hybridization and the superscript its formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the formal bond orders with electronegative atoms; the C- -N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds; the C⋯X bond order in pyrrole or furan may be considered as 1. [d] An α-C may be defined as a C attached through a single bond with −C═X, −C≡X, −C- -X. [e] As in nitro, *N*-oxides. [f] Pyrrole-type structure. [g] Pyridine-type structure. [h] Pyridine *N*-oxide type structure.

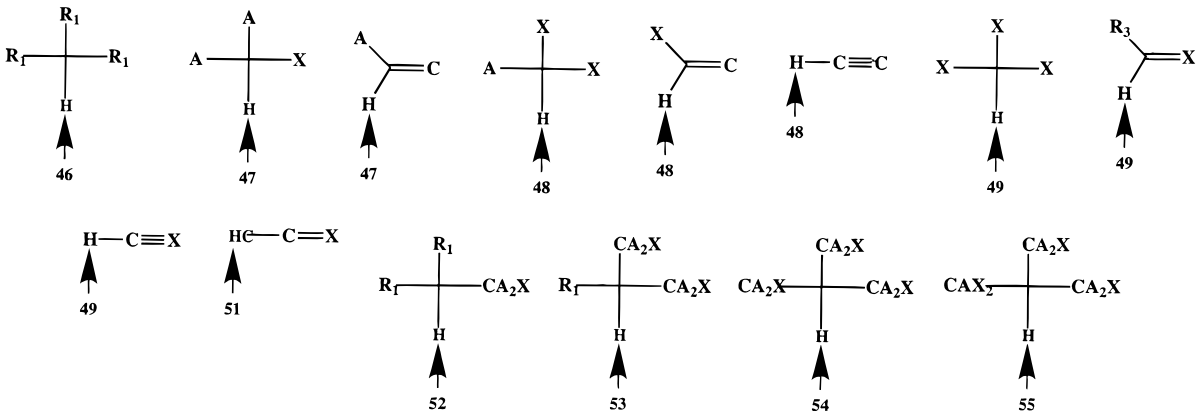

Figure 1. Description of various types of hydrogen. R_1 = H or CA_3 where A = C or H, X = any electronegative atom, R_3 = C or H or X. For atom types 53−55, the electronegative atom may be attached to the same or different carbon atom (as shown above).

TABLE 2: Comparative Evaluation of ALOGP (Current and Old Parameters) and CLOGP Methods for the Training[a] and Test Set

data set	method of calc	correl coeff	rms dev	max dev	min dev	max log *P*	min log *P*	predictive r^2	no. of data points
train	CLOGP	0.96	0.50	4.58	0.00	9.96	−4.41	0.91	8141
	ALOGP	0.95	0.55	3.66	0.00	9.96	−4.41	0.90	8364
	ALOGP (old)[b]	0.91	0.73	4.68	0.00	9.96	−4.41	0.81	8364
test	CLOGP	0.96	0.51	3.46	0.00	9.07	−3.15	0.91	912
	ALOGP	0.95	0.55	2.75	0.00	9.07	−3.52	0.90	931
	ALOGP (old)	0.91	0.75	4.24	0.00	9.07	−3.52	0.81	931

[a] The explained variance of the regression analysis = 0.895. The definition and necessity of the use of *explained variance* have been described in Purcell, W. P.; Bass, G. E.; Clayton, G. E. *Strategy of Drug: A Guide to Biological Activity*; Wiley: New York, 1973; p 23. [b] The parameters for the four new atom types were taken from the new parameter set.

Computational Methodology

Computational methods developed and applied in the present work are directed toward two tasks: first, development of new parameters for the ALOGP method; second, comparison of the ALOGP and CLOGP methods. The atom types used in the ALOGP method have been discussed extensively in the literature.[10,11,14,20] The present work uses the same atom types, with additions for boron, silicon, charged oxygen as in a carboxylate, and charged nitrogen as in Lys and halide ions. These atom types are shown in Table 1. Developing new parameters for these types entails the following: (i) collection of reliable experimental values of log *P* for common organic compounds along with their chemical structure; (ii) interpretation and conversion of the structures suitable for graphical display and atom type classification according to the scheme presented in Table 1; (iii) division of the compounds into two sets, *training* and *test*; (iv) evaluation of the atomic parameters using linear regression analysis from the training set; and (v) evaluation of the fitted (training) or predicted (test) log *P* values from the atomic parameters. Comparison of the ALOGP and CLOGP methods involves classification of the log *P* database by the criteria of organic structural class (aldehydes, heterocyclic aromatics, etc.) and molecular size, followed by the calculation of statistical parameters for each of the subclasses and comparison of statistical results. These steps are discussed below.

1. New Parameters for the ALOGP Method. *(i) Molecular Database.* The largest compilation of experimentally determined log *P* values of organic compounds known today is a result of the extensive efforts of Hansch and Leo.[2] The Pomona Medchem Database[25] consists of the experimental log *P* values of nearly 30 000 compounds along with the structural information in SMILES notation.[26] A subset of approximately 9000 compounds, with very accurately determined log *P* values, forms the star list.[25] A text (ASCII) version of the star list database was used for extracting the measured log *P* values and SMILES[26] strings for the structures.

(ii) Structural Interpretation. The star list structures encoded in the SMILES notation were first converted to 3D molecular structures using the 2D-to-3D converter module of Galaxy.[27] During this process, Galaxy also classified the atom types according to Table 1. An accurate atom classification procedure is the key to the success of the ALOGP method, and hence, the atom type classification of Galaxy was thoroughly checked for its correctness. Errors may come either from the misinterpretation of the atom notation in Table 1 or from the bond type representation in ambiguous structures. Galaxy,[27] for example, did not interpret the *N*-oxides properly, since it expected the N−O bond type to be delocalized; however, it was represented as a double bond in the SMILES-encoded star list database. Such structures were identified using the substructure search options in Galaxy, and the corresponding bond types were corrected. Azulenes were classified as aliphatic; although the database had only a few azulenes to make a definite conclusion, we feel that they are best represented as aromatic. All of the other organic compounds were interpreted properly, as ascertained by manual cross-checking of a large number of compounds using MaclogP.[28] Most other atom types shown in Table 1 are assigned in a straightforward manner, since they are determined both by their orbital hybridization state and by the atoms directly attached to them. Since, in an organic molecule, hydrogen is a recurring constituent of its solvent accessible surface, its classification was done by considering the nature of the directly attached atom and that of the neighbor of its direct attachments to account for the inductive polarization of the C−H bond. Some of the hydrogen atom types (46−55) are illustrated in greater detail in Figure 1. The subscript on the carbon to which the hydrogen is attached represents its formal oxidation number, which in turn may be considered as the number of electronegative atoms attached to it. The junction atoms in polynuclear heterocyclic rings are also classified uniquely by prioritizing the pyridine-type ring over the pyrrole-type ring over the benzene ring. For example, the junction

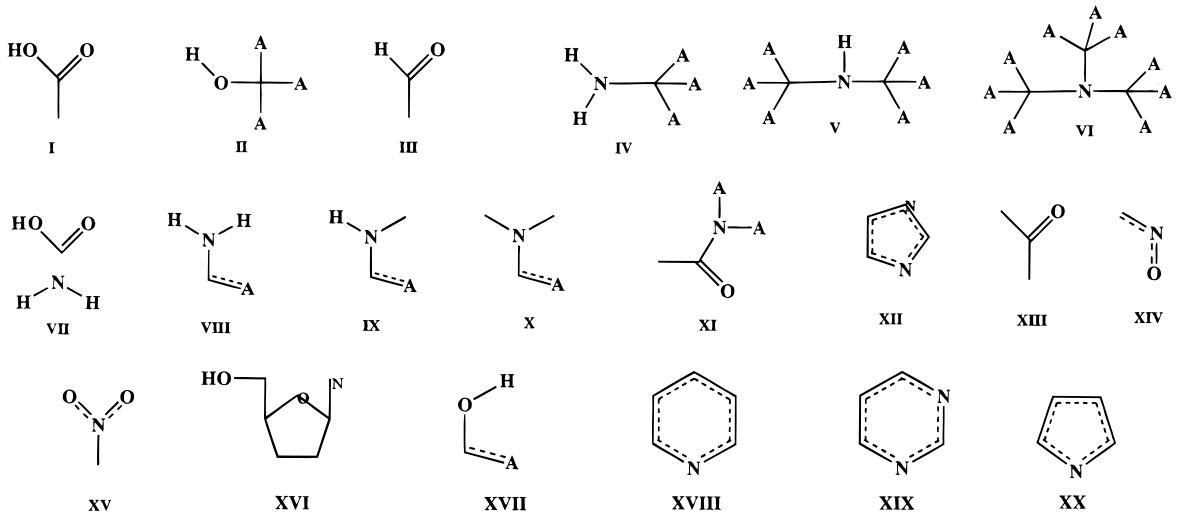

Figure 2. The representation of the substructures searched for the comparative analysis of the ALOGP and CLOGP methods. In these structures, the atoms are C unless otherwise indicated. "A" represents any element; in the Galaxy software, this is equivalent to picking only the bond and not the atom.

atoms in quinolines belong to type 28, and not to type 26. For larger cyclic structures, the ring atoms were classified as aliphatic systems. However, the absence of measured log *P* values of such systems in the current dataset did not justify such classification.

(iii) Training and Test Sets. The whole molecular dataset was divided into two subsets, training and test. The test set constituted every tenth molecule of the full dataset. The training set was the remaining 90% of the compounds. The objectives here are to use the training set to evaluate the atomic log *P* parameters and to use the test set to evaluate its predictive power. Both the training and the test sets were converted to a Galaxy 3D molecular databases along with the measured log *P* and CLOGP estimated values obtained from the log *P* master database.

(iv) Parameter Evaluation. The atomic log *P* parameters were determined from the general eq 1

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

where n_i is the number of atoms of type i and a_i is the atomic log *P* contribution. The tabulation of data of this massive regression was done via the 2D-QSAR module of Galaxy.[27] A generalized inversion method[29] was used to evaluate the least squares solutions of the atomic parameters.

2. Comparison of the ALOGP and CLOGP Methods and Results. *(i) Classification of the Star List Database.* A detailed examination of the star list database showed that it is quite comprehensive in representing all of the organic functional groups of interest in medicinal chemistry. Twenty types of organic and bio-organic functional groups, ubiquitous in the medicinal chemistry literature, have been identified from this database for the purpose of the present comparisons. These are carboxylic acids, alcohols, aldehydes, aliphatic primary amines, aliphatic secondary amines, aliphatic tertiary amines, unblocked peptides including the amino acids, aromatic primary amines, aromatic secondary amines, aromatic tertiary amines, carboxamides, imidazoles, ketones, *N*-oxides, nitro compounds, nucleosides, phenols, pyridines, pyrimidines, and pyrroles. Clearly, multifunctional compounds belong to more than one class, and hence, the structural classes do overlap. Nevertheless, a classification of this kind is useful in qualitatively delineating the shortcomings of a method in terms of specific structural classes. Since atomic/fragment contribution methods assume additivity of log *P*, it is important to explore the limits of the additivity approximation. It stands to reason that high molecular weight compounds that are conformationally flexible (e.g., peptides) will be harder to model using the additivity assumption. Therefore, we classified the database into 11 subclasses (bins) on the basis of the molecular size (total number of atoms in a molecule) for performance evaluation.

(ii) Statistical Parameters for the Assessment of log *P* Models. To assess the calculated log *P* values from the two models, we used the Pearson correlation coefficient (*R*), rms deviation, maximum and minimum deviations for each set, and "predictive r^2".[30] These parameters were calculated for the overall training and test sets as well as for each subclass of molecules described earlier.

Predictive r^2 [30] measures the quality of predictions relative to a simple "no model" guess, the average of all experimental log *P* values for a given set of molecules. This is given by

$$\text{predictive } r^2 = (\text{SD} - \text{"press"})/\text{SD} \qquad (2)$$

where SD is the sum of squared deviations of each measured log *P* value from their mean, and "press" is the predictive sum of squared differences (the sum of squared differences between the actual and predicted log *P* values). Negative values for predictive r^2 indicate that log *P* is better estimated by the "mean of values" rather than by the model under consideration. The ALOGP values as obtained from the regression analysis for the training set and the calculated CLOGP values of molecules in the star list were appended to the molecular databases (training and test*)* devised earlier for the analysis of the properties. The ALOGP values of the test set were evaluated from the database module after replacing the original atomic hydrophobicity parameter file of Galaxy software by the current values.

Results and Discussion

New Parameters for the ALOGP Method. The atom classification scheme shown in Table 1 is an attempt to discretize the electronic effects, solvent accessibility, and so forth of an atom from a topological consideration. These properties are the most critical in determining the relative contribution of different atom types to log *P* for small organic molecules with relatively few degrees of freedom. Thus, the atomic parameters

TABLE 3: Comparative Evaluation of ALOGP and CLOGP Methods among Various Classes of Organic Compounds for the Training Set[a]

compd type	method of calc	correl coeff	rms dev	max dev	min dev	max log *P*	min log *P*	predictive r^2	no. of data points
carboxylic acid (I)	CLOGP	0.97	0.36	2.28	0.00	6.30	−2.60	0.93	448
	ALOGP	0.93	0.52	1.90	0.00	6.30	−2.60	0.86	456
alcohol (II)	CLOGP	0.94	0.81	4.58	0.00	8.42	−3.70	0.79	880
	ALOGP	0.96	0.52	3.17	0.00	8.42	−3.70	0.91	889
aldehyde (III)	CLOGP	0.95	0.45	1.60	0.00	8.02	−0.10	0.89	36
	ALOGP	0.97	0.41	1.37	0.00	8.02	−0.10	0.91	36
aliphatic primary amine (IV)	CLOGP	0.91	0.79	2.57	0.00	8.38	−4.41	0.82	274
	ALOGP	0.95	0.57	2.85	0.00	8.38	−4.41	0.90	274
aliphatic secondary amine (V)	CLOGP	0.95	0.52	2.25	0.00	4.90	−4.00	0.90	140
	ALOGP	0.97	0.41	1.25	0.00	4.90	−4.00	0.94	143
aliphatic tertiary amine (VI)	CLOGP	0.95	0.57	2.39	0.00	7.57	−3.80	0.89	297
	ALOGP	0.94	0.60	2.30	0.00	7.57	−3.80	0.88	308
peptides (VII)[b]	CLOGP	0.81	1.00	2.57	0.00	1.63	−4.41	−0.08	148
	ALOGP	0.85	0.62	2.85	0.01	1.63	−4.41	0.58	148
aromatic primary amine (VIII)	CLOGP	0.93	0.51	2.07	0.00	5.08	−2.20	0.82	480
	ALOGP	0.90	0.54	3.01	0.00	5.08	−2.20	0.80	482
aromatic secondary amine (IX)	CLOGP	0.94	0.55	1.55	0.01	5.18	−0.80	0.83	59
	ALOGP	0.94	0.52	1.14	0.06	5.18	−0.80	0.85	59
aromatic tertiary amine (X)	CLOGP	0.87	0.67	4.42	0.00	5.34	−0.90	0.73	104
	ALOGP	0.91	0.54	1.54	0.00	5.34	−0.90	0.82	104
carboxamide (XI)	CLOGP	0.94	0.57	2.57	0.00	5.66	−3.51	0.88	1077
	ALOGP	0.94	0.55	2.85	0.00	5.66	−3.51	0.89	1129
imidazole (XII)	CLOGP	0.95	0.63	3.34	0.00	6.06	−3.56	0.85	508
	ALOGP	0.96	0.49	3.66	0.00	6.06	−3.56	0.91	510
ketone (XIII)	CLOGP	0.91	0.62	3.71	0.00	7.57	−2.10	0.80	434
	ALOGP	0.89	0.65	2.34	0.00	7.57	−2.10	0.78	441
N-oxides (XIV)	CLOGP	0.80	0.98	2.91	0.00	3.65	−1.40	−0.01	63
	ALOGP	0.87	0.69	1.40	0.01	3.65	−1.40	0.49	64
nitro (XV)	CLOGP	0.94	0.53	3.34	0.00	5.29	−1.59	0.84	798
	ALOGP	0.91	0.55	3.66	0.00	5.29	−1.59	0.83	836
nucleoside (XVI)	CLOGP	0.78	1.12	2.24	0.15	1.35	−1.89	−1.82	43
	ALOGP	0.83	0.40	1.20	0.00	1.35	−1.89	0.65	43
phenol (XVII)	CLOGP	0.94	0.57	3.71	0.00	9.96	−3.51	0.89	491
	ALOGP	0.93	0.64	2.48	0.00	9.96	−3.51	0.86	494
pyridine (XVIII)	CLOGP	0.95	0.42	2.71	0.00	7.00	−2.44	0.89	379
	ALOGP	0.92	0.52	2.27	0.00	7.00	−2.44	0.83	380
pyrimidine (XIX)	CLOGP	0.91	0.69	2.24	0.00	5.70	−1.42	0.69	247
	ALOGP	0.92	0.51	1.75	0.00	5.70	−1.42	0.83	247
pyrrole (XX)	CLOGP	0.93	0.55	2.08	0.00	6.40	−2.21	0.84	129
	ALOGP	0.93	0.54	1.54	0.00	6.40	−2.21	0.85	129

[a] See Figure 2 for the substructures that were searched in different chemical classes. [b] Unblocked peptides including the amino acids.

shown in Table 1 represent the atomic contributions to log *P* and may be considered as atomic measures of hydrophobicity. Table 1 also shows the number of compounds that contain the corresponding atom type and the total number of occurrences of each atom type in the database. Data shown in Table 1 were obtained from a regression model based on 8364 molecules, covering a large variety of organic structures. The statistics of this analysis are shown in Table 2. The reliability of the atomic parameters is evident from the high correlation coefficient, predictive r^2, and other statistics obtained for both the training and the test sets.

It is noteworthy that the old set of atomic parameters gave a good correlation ($r = 0.91$) and a reasonable standard deviation (rms = 0.73 and 0.75 for the training and and test sets, respectively) with the star list database. However, all statistical parameters calculated using the new atomic parameters improved significantly relative to those calculated using the old parameters. The sign of atomic values for the carbon types is consistent with previous studies, with a few exceptions (types 17, 20, 21, 35, and 37). In these cases, the number of observations was small (10 or less) in the previous database, explaining the change of hydrophobic character of these types. Atomic values of types 40 and 43 are close to zero in both new and old parameter sets. Hence, the change of sign in these cases is not important. Notably, the sign of the atomic values for all hydrogen types is unchanged from the previous set. Notably, the hydrogen attached to a heteroatom is less hydrophilic in the present set. In the old set, heteroatoms bonded to hydrogen (types 56, 57, 66, 67, 69, and 70) were given positive atomic values. In the present set, more realistically, they are all represented with negative values, distributing the hydrophilic character of polar groups such as hydroxyls and amines, to both hydrogens and heteroatoms. A similar redistribution of hydrophilic character is also seen in the cases of phosphates and phosphanates, changing the sign of atomic values (types 117 and 120) in these cases. Hydrophobicity values for halides in the present set are quite similar to those of the previous set. We made an effort to assign atomic values to ionic halides, despite their low level of occurrence in the star list database (types 101−103). In the case of bromide ion (type 103), with a solitary occurrence in the database, the derived atomic value is not a statistically valid parameter. A physically more realistic value may be the average of chloride and iodide atomic values, which is −2.8929. More experimental values for such ions would allow a better assignment of the ALOGP parameters.

Since hydrogen types (46−49 and 51−55) are always bonded to carbons, the "negative" contribution of carbons may not be very important, as the hydrocarbon surface will always be represented as hydrophobic. Compared to the previous set of parameters, some of the carbon atom types appear to have a more negative hydrophobicity, while the hydrogens have more positive values. However, the log *P* contribution of a methyl group in a hydrocarbon chain remains virtually unaffected: 0.642 (current), 0.648 (previous). Substituting a non-hydrogen atom for hydrogen on a saturated carbon can have two effects. If the non-hydrogen atom is a carbon, the substitution increases the overall hydrophobicity. If the non-hydrogen atom is an electronegative atom, it decreases the hydrophobicity for a single

TABLE 4: Comparative Evaluation of ALOGP and CLOGP Methods among Various Classes of Organic Compounds for the Test Set[a]

compd type	method of calc	correl coeff	rms dev	max dev	min dev	max log *P*	min log *P*	predictive r^2	no. of data points
carboxylic acid (I)	CLOGP	0.95	0.43	1.52	0.00	6.06	−1.26	0.88	61
	ALOGP	0.92	0.54	2.19	0.01	6.06	−1.26	0.80	61
alcohol (II)	CLOGP	0.95	0.84	3.46	0.01	5.76	−3.02	0.80	100
	ALOGP	0.97	0.42	1.24	0.00	5.76	−3.02	0.95	100
aldehyde (III)	CLOGP	0.99	0.13	0.21	0.00	1.76	−0.01	0.96	5
	ALOGP	0.89	0.39	0.54	0.08	1.76	−0.01	0.64	5
aliphatic primary amine (IV)	CLOGP	0.84	0.92	2.37	0.00	3.54	−3.15	0.67	30
	ALOGP	0.94	0.57	1.17	0.00	3.54	−3.15	0.87	30
aliphatic secondary amine (V)	CLOGP	0.97	0.45	1.35	0.08	4.17	−2.56	0.93	20
	ALOGP	0.98	0.36	0.70	0.03	4.17	−2.56	0.95	20
aliphatic tertiary amine (VI)	CLOGP	0.95	0.43	1.09	0.01	5.76	0.64	0.90	24
	ALOGP	0.92	0.53	1.27	0.02	5.76	0.64	0.85	24
peptides (VII)[b]	CLOGP	0.73	1.12	2.37	0.14	−0.78	−3.15	−2.17	19
	ALOGP	0.85	0.68	1.17	0.15	−0.78	−3.15	−0.19	19
aromatic primary amine (VIII)	CLOGP	0.94	0.50	1.33	0.02	4.31	−0.96	0.85	47
	ALOGP	0.92	0.58	1.62	0.01	4.31	−3.52	0.85	48
aromatic secondary amine (IX)	CLOGP	0.97	0.81	1.41	0.02	2.99	−0.89	0.65	9
	ALOGP	0.97	0.37	0.64	0.05	2.99	−0.89	0.93	9
aromatic tertiary amine (X)	CLOGP	0.94	0.45	0.92	0.03	4.41	0.03	0.87	11
	ALOGP	0.94	0.48	1.11	0.15	4.41	0.03	0.85	11
carboxamide (XI)	CLOGP	0.94	0.60	2.37	0.00	3.92	−3.15	0.881	26
	ALOGP	0.95	0.53	1.47	0.01	4.24	−3.15	0.91	128
imidazole (XII)	CLOGP	0.94	0.89	2.73	0.00	4.96	−2.47	0.75	48
	ALOGP	0.98	0.40	1.31	0.00	4.96	−2.47	0.95	49
ketone (XIII)	CLOGP	0.96	0.58	1.52	0.03	9.07	0.30	0.89	21
	ALOGP	0.95	0.53	1.06	0.01	9.07	0.30	0.91	21
N-oxide (XIV)	CLOGP	0.83	0.90	2.50	0.01	3.64	−0.88	0.42	8
	ALOGP	0.88	0.87	1.23	0.26	3.64	−0.88	0.47	8
nitro (XV)	CLOGP	0.94	0.68	2.73	0.00	4.53	−1.59	0.72	90
	ALOGP	0.88	0.62	1.66	0.01	4.53	−1.59	0.75	96
nucleoside (XVI)	CLOGP	0.90	1.17	2.02	0.38	0.95	−1.30	−0.39	4
	ALOGP	0.97	0.25	0.31	0.09	0.95	−1.30	0.9	44
phenol (XVII)	CLOGP	0.90	0.76	3.46	0.00	4.82	−1.77	0.77	58
	ALOGP	0.93	0.64	2.19	0.03	4.82	−1.77	0.84	58
pyridine (XVIII)	CLOGP	0.97	0.29	0.75	0.00	3.27	−0.65	0.93	37
	ALOGP	0.88	0.51	1.59	0.01	3.27	−0.65	0.76	37
pyrimidine (XIX)	CLOGP	0.91	0.58	1.15	0.03	2.61	−0.05	0.48	12
	ALOGP	0.90	0.49	1.03	0.06	2.61	−0.05	0.63	12
pyrrole (XX)	CLOGP	0.86	0.73	1.52	0.00	5.43	0.71	0.73	10
	ALOGP	0.93	0.56	1.12	0.05	5.43	0.71	0.84	10

[a] See Figure 2 for the substructures that were searched in different chemical classes. [b] Unblocked peptides including the amino acids.

substitution because of two opposing effects, polarization and shielding. In the case of aromatic carbons, any substitution involving carbon or heteroatoms shows increased hydrophobicity. As expected, most oxygen and nitrogen atom types show negative hydrophobicity. Zwitterionic unblocked peptides including the amino acids, were the main source of log *P* values for the estimation of atomic parameters for charged oxygen and nitrogen types. The two types (O^- and N^+) occur in pairs in most cases. In other words, these two variables are linearly dependent, and there is no clear mathematical basis by which to assign their individual hydrophobicity contribution. One way to separate these variables may be to study the log *P* values of a sufficient number of sodium carboxylates and ammonium halides. In the absence of such a dataset, we used the relative solvation free energies[31] of acetate and ammonium ions as a guide in estimating their *relative* hydrophobicity contributions.

Comparison of the Performance of CLOGP and ALOGP Methods. A major objective of this work is to make an unbiased and comprehensive comparison of the performance of the ALOGP and CLOGP methods. In addition to the usual statistical parameters, the rms deviation and the Pearson correlation coefficient, we used several others in different sets of compounds to compare their performance. These statistics are shown in Tables 2−6. One should be aware that the performance of the two methods is best judged from the test set results and not from those of the training set. In the training set, one can easily increase the number of independent variables and improve the statistics. However, that would decrease the predictive power of the method because several of the independent variables would then be under-represented. Over the years the number of fragment and correction factors used in the CLOGP method has grown considerably. In the early publications, this number was slightly larger than 200, but the current number is not available in the literature. However, this number may not be directly comparable with the number of independent variables used in the ALOGP method because, in the CLOGP method, the fragmental constants and correction factors are not derived by the usual regression procedures for the entire database but are obtained incrementally for different series with periodic alterations for the old values as indicated by new log *P* data. A detailed comparison of the training and test set results is possible only for the ALOGP method here, as the training and test subsets for the CLOGP method were not available to us. Also, the number of fragments and correction factors used in the current version of CLOGP is not available in the literature. Therefore, the statistical parameters shown in Tables 2−6 (training versus test) should be interpreted with caution. The overall statistics of the study for the training and test sets are shown in Table 2. Notably, the number of data points (molecules) used for the two methods differed, since the CLOGP method failed to evaluate the log *P* values for a number of cases both in the training and test sets. The correlation coefficient, rms deviation, and predictive r^2 for the CLOGP method are marginally better than those for the ALOGP method, whereas the maximum deviation (max dev) is greater for the CLOGP method (training or test). The statistics remained almost unchanged for the training and test sets for either method.

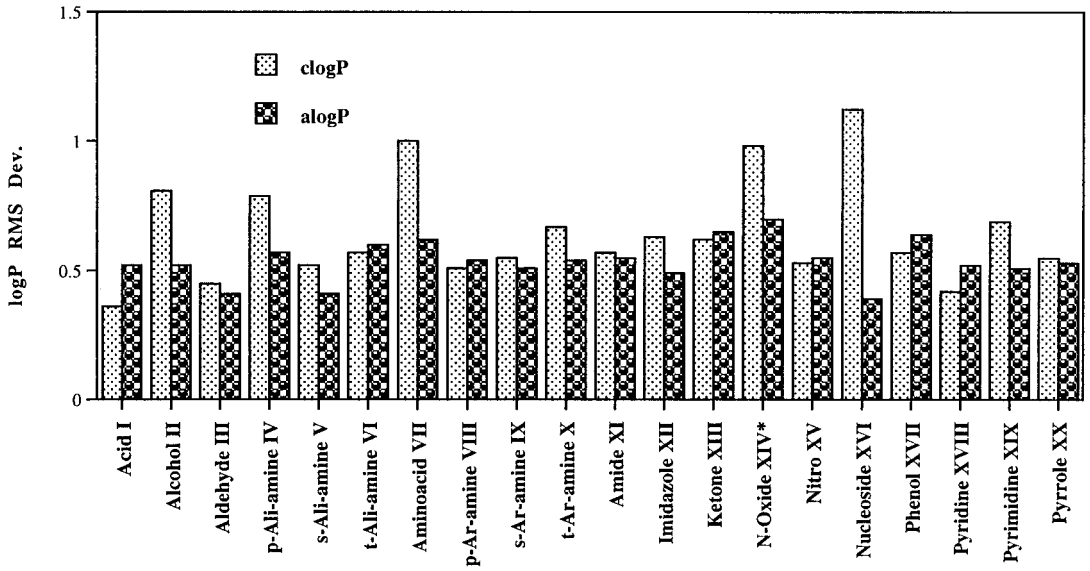

Figure 3. A comparative evaluation of the ALOGP and CLOGP methods among various classes of organic compounds in the training set. Dr. A Leo of Pomona College recently showed us that his MaclogP program, version 2.03, does a better job for the *N*-oxides. For 82 data points, it showed 0.96 correlation coefficient, 0.2 rms deviation, and 0.92 predictive r^2. The master database that we analyze here is compatible with the MaclogP version 2.0.

TABLE 5: Comparative Evaluation of ALOGP and CLOGP Methods for Molecules, Classified by the Total Number of Atoms in the Training Set

atom range	method of calc	correl coeff	rms dev	max dev	min dev	max log *P*	min log *P*	predictive r^2	no. of data points
0−15	CLOGP	0.98	0.31	1.73	0.00	6.07	−3.21	0.96	822
	ALOGP	0.95	0.50	1.86	0.00	6.07	−3.21	0.90	831
16−20	CLOGP	0.97	0.32	2.37	0.00	6.42	−4.00	0.94	1406
	ALOGP	0.93	0.48	1.71	0.00	6.42	−4.00	0.86	1419
21−25	CLOGP	0.98	0.41	2.97	0.00	9.14	−4.41	0.95	1657
	ALOGP	0.96	0.52	2.85	0.00	9.14	−4.41	0.92	1698
26−30	CLOGP	0.96	0.47	2.39	0.00	7.54	−4.20	0.91	1388
	ALOGP	0.94	0.54	3.17	0.00	7.54	−4.20	0.88	1427
31−35	CLOGP	0.95	0.53	2.91	0.00	6.63	−3.09	0.87	977
	ALOGP	0.93	0.54	2.73	0.00	6.63	−3.09	0.86	1007
36−40	CLOGP	0.95	0.53	2.59	0.00	7.10	−3.51	0.89	599
	ALOGP	0.94	0.57	3.01	0.00	7.10	−3.51	0.88	619
41−45	CLOGP	0.95	0.62	2.43	0.00	5.85	−3.70	0.86	478
	ALOGP	0.94	0.60	2.64	0.00	5.85	−3.70	0.87	495
46−50	CLOGP	0.94	0.68	2.28	0.00	6.30	−2.33	0.83	264
	ALOGP	0.94	0.61	2.11	0.01	6.30	−2.33	0.86	276
51−55	CLOGP	0.96	0.71	3.00	0.00	7.41	−2.64	0.90	208
	ALOGP	0.96	0.64	2.10	0.00	7.48	−2.64	0.92	222
56−60	CLOGP	0.96	0.82	4.20	0.00	8.06	−3.05	0.91	121
	ALOGP	0.96	0.72	2.91	0.00	8.06	−3.05	0.93	134
60+	CLOGP	0.88	1.11	4.58	0.00	9.96	−2.82	0.75	221
	ALOGP	0.93	0.79	3.66	0.00	9.96	−2.82	0.86	236

The whole dataset was then searched for various types of common organic functional groups shown in Figure 2, and the results for each subclass were then separately analyzed. Clearly, this type of analysis is helpful to the user to decide what is the most appropriate method for a given subclass or to point to the limitations of a given method. The results of this analysis are shown in Table 3 (training set), Table 4 (test set), and Figure 3 (training set). The number of compounds in some classes of the test set is indeed small. If the statistics of the fit for such classes are different from the training set, the reader may rely on the corresponding results of Table 3. It is seen that the ALOGP method gives a more stable performance relative to CLOGP when all the different structural classes are considered. Thus, the CLOGP method results in rms deviations greater than 0.65 in seven subclasses (alcohols, aliphatic primary amines, unblocked peptides, including the amino acids, aromatic tertiary amines, *N*-oxides, nucleosides and pyrimidines). Except for *N*-oxides, none of the structural classes shows rms deviations greater than 0.65 in the case of the ALOGP method. With CLOGP, predictive r^2 is negative for unblocked peptides, including the amino acids and nucleosides, indicating that this method in its current version may not be generally applicable to bio-organics. By the criterion of predictive r^2, most organic structural classes are reasonably well-predicted by both methods. The latest version of CLOGP as available in MacLogP (v 2.03), gives vastly better values for *N*-oxides than the previous CLOGP version (results shown in Tables 3−6). These results are summarized in the footnote of Figure 3. The maximum deviation (max dev) in Tables 3−6 is an indication of how far a really poor prediction can deviate from the measured value. For most structural classes, this parameter is higher with the CLOGP method. This parameter is most often 4 to 5 times the rms deviation. If the same set of compounds was poorly predicted by both methods, it would be indicative of either a problem in measuring the log *P* values experimentally or a problem in structural representation (for example, neglect of tautomeric structure or of acid/base behavior). Picking poorly predicted compounds is often relative to the magnitude of the measured value. A deviation of 1.5, for example, for a compound with a log *P* value of 8.0 may not seem very bad.

TABLE 6: Comparative Evaluation of ALOGP and CLOGP Methods for Molecules, Classified by the Total Number of Atoms in the Test Set

atom range	method of calc	correl coeff	rms dev	max dev	min dev	max log *P*	min log *P*	predictive r^2	no. of data points
1−15	CLOGP	0.98	0.26	1.44	0.00	4.04	−1.80	0.95	104
	ALOGP	0.95	0.39	1.20	0.00	4.04	−1.80	0.89	105
16−20	CLOGP	0.98	0.26	1.30	0.00	4.56	−3.15	0.96	193
	ALOGP	0.91	0.51	2.19	0.01	4.56	−3.15	0.83	194
21−25	CLOGP	0.98	0.38	1.82	0.00	7.43	−1.87	0.96	167
	ALOGP	0.97	0.52	1.66	0.00	7.43	−1.87	0.93	170
26−30	CLOGP	0.96	0.56	2.50	0.00	5.43	−2.45	0.88	132
	ALOGP	0.93	0.58	1.82	0.01	5.43	−2.45	0.87	136
31−35	CLOGP	0.93	0.66	2.73	0.01	4.44	−2.60	0.81	99
	ALOGP	0.94	0.51	1.47	0.01	4.44	−2.60	0.88	100
36−40	CLOGP	0.96	0.49	1.35	0.00	6.42	−2.56	0.92	70
	ALOGP	0.92	0.71	2.75	0.01	6.42	−3.52	0.84	77
41−50	CLOGP	0.95	0.67	2.11	0.00	6.26	−2.72	0.87	80
	ALOGP	0.95	0.60	1.27	0.00	6.26	−2.72	0.90	81
51−60	CLOGP	0.95	0.71	2.04	0.04	5.50	−2.80	0.90	40
	ALOGP	0.97	0.53	1.16	0.01	5.50	−2.80	0.94	41
60+	CLOGP	0.89	1.17	3.46	0.01	9.07	−2.80	0.75	32
	ALOGP	0.95	0.75	1.83	0.04	9.07	−2.80	0.90	32

TABLE 7: List of the Compounds Having a Predicted and Observed log P Difference Greater than 2.0 by the ALOGP Method

compound	log *P*	CLOGP	ALOGP	Δ*c*	Δ*a*
1 1-(4-aminobutoxy)-*d*8-THC	8.38	7.51	6.15	−0.87	−2.23
2 1-nitrosourea, 1-(*N*-oxo-2,2,6,6,-tetramethylpiperidin-4-yl)-3-glucos-3-yl)	1.87	−2.33	−1.04	−4.20	−2.91
3 phosphonic amide, *N*,*N*-diMe-*P*,*P*-di-1-pyrrolidinyl	1.88	1.93	−0.19	0.05	−2.07
4 2,4-diNO_2-$C_6H_3NHN{=}C(CN)COOET$	4.14	2.89	1.92	−1.25	−2.22
5 Gly-Gly-Gly	−2.68	−4.29	−5.53	−1.61	−2.85
6 basagran	2.80	2.80	0.68	0.00	−2.12
7 D_8-THC-dimethylheptyl	9.96	9.09	7.48	−0.87	−2.48
8 dibenzodioxin, 1,2,3,7-tetrachloro	8.22	7.31	5.78	−0.91	−2.44
9 dibenzodioxin, 1,2,4-trichloro	7.47	6.71	5.11	−0.76	−2.36
10 methanetricarboxamide, hexamethyl	−3.09	−0.99	−0.36	2.10	2.73
11 ketobemidone, isopropyl carbonate	1.72	1.99	3.78	0.27	2.06
12 5′-chlorocyclocytosine	−3.10		0.07		3.17
13 ketobemidone, methyl carbonate	0.75	1.16	3.05	0.41	2.30
14 DEF	3.23	3.24	5.87	0.01	2.64
15 dimethyldiethoxysilane	0.61	0.61	2.69	0.00	2.08
16 1-naphthol-3-sulfonic acid	−0.18	−0.14	1.83	0.04	2.01
17 quninoline-2-carboxaldehyde, *N*-phenylguanylhydrazone	0.99	0.99	3.26	0.00	2.27
18 β-keto-ω-hydroxyacylanilide, 2,6-diisopropyl analogue	4.20	4.43	6.20	0.23	2.00
19 xylazine	1.00	1.00	3.27	0.00	2.27
20 2-*N*,*N*-dipentylamino-1-phenylethanol hydrochloride	2.86	5.14	4.96	2.29	2.10
21 hexahydropyrimidine, 2-nitromethylene-3-(6-chloropyrid-3-yl) methyl	−0.62	0.44	1.46	1.06	2.08
22 triflumizole	1.40	1.59	3.76	0.19	2.36
23 ketobemidone, ethyl carbonate	1.29	1.68	3.40	0.40	2.11
24 1-(2,3,5-tribenzoyl-D-ribofuranosyl)-2-NO_2-imidazole	1.19	4.53	4.85	3.34	3.66
25 6-purinethione, 4-carboxyethyl	−1.65	−1.10	0.80	0.55	2.45
26 2,6-diisopropylacetanilide, α-diphenylacetyl	3.96		6.30		2.34
27 benzoylacetanilide, 2,6-diisopropyl	2.87	3.20	4.92	0.34	2.05
28 diphenylguanidine	−0.05	−0.05	2.58	0.00	2.63
29 pararosaniline	−0.21		2.80		3.01

However, such a deviation for a molecule with log *P* close to 0.0 may seem very bad. We tried to analyze the compounds with deviations greater than 2.0 in either method. In the process, we detected a few structural problems. For example, the anions of a few pyridinium salts were not represented in the SMILES strings. These poorly predicted compounds in these two methods are shown in Tables 7 and 8. The number of such bad compounds is only 29 for the ALOGP method (Table 7), while that number for the CLOGP method is 62. Methanetricarboxamide hexamethyl (compound **10** in Table 7 or compound **41** in Table 8) is possibly acidic, and the measured log *P* value is probably not corrected for its ionization state. Overall, the number of common compounds in the two lists is not high. This shows that, though both methods are additive−constitutive, they do not share the same limitations.

Next, the whole database was divided according to the size of the molecules. Molecular size may be specified in a number of ways (e.g., by molecular weight, number of atoms, number of non-hydrogen atoms). All three criteria led to similar results. Results shown in Tables 5 and 6 (for training and test sets) used the total number of atoms as the criterion for classifying the database. Figure 4 shows a comparative histogram plot of rms deviations for molecules of different sizes using the two methods. Clearly, the CLOGP method is superior for molecules with very small size, i.e., about 20 atoms or less. The two methods were very similar in performance for molecules in the range of 21−45 atoms. For molecules with over 45 atoms, the ALOGP method outperformed CLOGP. The statistics in the training set and the test set were very similar. Importantly, both methods lead to gradually weaker predictions as the size of the molecule becomes larger, as evident from Figure 4. This figure shows that the largest rms deviations were obtained when molecules with over 60 atoms were considered. This shows the limits of the additivity approximation inherent in both methods. Further improvements in these methods, therefore, should include corrections for longer range nonbonded interactions such as intramolecular hydrogen bonding and molecular flexibility, which tend to be more important for larger molecules. However, for larger molecules such as polypeptides and proteins, the additivity assumption is expected to break down, and better predictions are likely to be obtained by the considerations of exposed surface area in addition to atom types.

TABLE 8: List of the Compounds Having Predicted and Observed log *P* Difference Greater than 2.0 by the CLOGP Method

compound	log *P*	CLOGP	ALOGP	Δ*c*	Δ*a*
1 podophyllotoxin	2.01	−0.99	2.11	−3.00	0.10
2 morpholinodaunorubicin	2.31	−0.17	1.28	−2.48	−1.03
3 tirapazamine	−0.30	−2.67	−0.56	−2.37	−0.26
4 4,6-diiodoresorcyl-1,3-diglucoside	−0.25	−2.88	−1.36	−2.64	−1.11
5 guanosine	−1.89	−3.90	−2.38	−2.01	−0.49
6 1-acetyl-7-methoxy-*N*-methylmitosene	2.39	0.31	1.12	−2.08	−1.27
7 azacitidine	−2.17	−4.30	−1.83	−2.13	0.34
8 *N*-acetyl-Ara-C	−1.35	−3.43	−2.41	−2.08	−1.06
9 1-nitrosourea, 1-(*N*-oxo-2,2,6,6-tetramethylpiperidin-4-yl)-3-(glucos-3-yl)	1.87	−2.33	−1.04	−4.20	−2.91
10 1-methyl-2-nitro-5-(CH═N(O)CH_3)imidazole	0.04	−2.28	0.30	−2.32	0.26
11 naloxone (5*R*,9*R*,13*R*,14*S*)	2.09	−0.34	1.45	−2.43	−0.64
12 benzotriazine, 1,4-di-*N*-oxide-3-amino-7-(2,3-dihydroxy)propoxy	−1.10	−4.01	−1.63	−2.91	−0.53
13 benzotriazine, 1,4-di-*N*-oxide-3-amino-6,7-dimethyl	0.56	−1.72	0.41	−2.28	−0.15
14 morpholinodaunorubicin, 13-dihydro	2.03	−0.40	1.33	−2.43	−0.70
15 2′-deoxyadenosine, *N*6-acetyl	−0.19	−2.43	−1.39	−2.24	−1.20
16 2′-deoxycytidine, *N*4-benzoyl	0.67	−1.59	0.03	−2.26	−0.64
17 *P*-nitrophenylmaltoside	−1.39	−3.49	−2.19	−2.10	−0.80
18 benzotriazine-1,4-di-*N*-oxide, 3-amino-6-chloro	0.41	−1.92	0.10	−2.33	−0.31
19 α-dihydrograyanotoxin II	1.51	−0.97	0.48	−2.48	−1.03
20 2-cyanomorpholinodoxorubicin, 12-imino	1.97	−1.74	0.53	−3.71	−1.44
21 2-cyanomorpholinodoxorubicin	1.98	−1.25	0.48	−3.23	−1.50
22 chloralose-β	1.12	−1.17	−0.08	−2.29	−1.20
23 microlenin	1.66	−0.71	2.54	−2.37	0.88
24 benzotriazine, 1,4-di-*N*-oxide, 3-amino-7-methyl	0.20	−2.17	−0.08	−2.37	−0.28
25 strychnine, bromthymol blue salt	1.93	−0.30	1.15	−2.23	−0.78
26 sucrose	−3.70	−5.72	−4.31	−2.02	−0.61
27 ouabagenin	−0.02	−4.60	−1.31	−4.58	−1.29
28 dithianon	2.84	−0.13	3.49	−2.97	0.65
29 gibberellin-α-3,2-α	0.24	−2.04	0.44	−2.28	0.20
30 cytidine, N_4-benzoyl	0.30	−1.94	−0.74	−2.24	−1.04
31 benzotriazine-1,4-di-*N*-oxide, 3-amino-7-methoxy	0.00	−2.31	−0.58	−2.31	−0.58
32 morpholinodoxorubicin, 12-imino	1.80	−1.19	0.67	−2.99	−1.13
33 3-nitrotriazole, 1-(*N*-methoxypropyl)thioacetamido	0.63	−1.38	0.10	−2.02	−0.53
34 2-cyanomorpholinodaunorubicin	2.59	−0.73	1.15	−3.32	−1.44
35 morpholinoadriamycin	1.73	−0.69	0.61	−2.42	−1.12
36 2-cyanomorpholinodoxorubicin, 13-dihydro	1.56	−2.10	0.31	−3.66	−1.25
37 2-aminostrychnine	0.54	−1.53	0.40	−2.07	−0.14
38 2-nitrostrychnine	1.61	−0.56	1.04	−2.17	−0.57
39 decitabine	−1.89	−3.95	−1.06	−2.06	0.83
40 benzotriazine, 1,4-di-*N*-oxide, 3-acetylamino	−0.60	−2.73	−0.70	−2.14	−0.10
41 methanetricarboxamide, hexamethyl	−3.09	−0.99	−0.36	2.10	2.73
42 naphthalene, octachloro	6.42	8.54	8.05	2.12	1.63
43 Leu-Leu-Leu-Val	−0.51	1.67	0.01	2.18	0.52
44 Ile-Tyr-Ile-Val	−1.09	0.97	0.23	2.06	1.32
45 acetylcholine bromide	−3.61	−1.22	−3.61	2.39	0.00
46 Ile-Ile-Val-Val	−1.41	1.14	−0.24	2.55	1.17
47 MT-124	4.33	6.92	5.56	2.59	1.23
48 propaquizafop	4.60	7.04	4.24	2.44	−0.36
49 *N*-benzylcinchonine chloride	−0.55	2.16	−0.06	2.71	0.49
50 Pro-Leu-Leu-Leu	−1.06	1.19	−0.51	2.25	0.55
51 5,5-dimethyl-3-(piperidin-1-yl)-2-cyclohexen-1-one	0.83	3.01	2.40	2.18	1.57
52 leucinylvalinylvaline	−2.10	−0.05	−1.15	2.05	0.95
53 2-*N*,*N*-dipentylamino-1-phenylethanol hydrochloride	2.86	5.14	4.96	2.29	2.10
54 Ile-Ala-Ala-Ile	−2.82	−0.71	−1.93	2.11	0.89
55 Val-Pro-Val-Leu	−1.91	0.50	−1.29	2.40	0.62
56 leucinylisoleucinylisoleucine	−1.11	1.00	−0.24	2.11	0.87
57 Leu-Pro-Leu-Leu	−0.92	1.55	−0.51	2.47	0.41
58 Leu-Leu-Pro-Leu	−1.00	1.55	−0.51	2.55	0.49
59 Leu-Leu-Leu-Pro	−1.18	1.39	−0.51	2.57	0.67
60 5,5-Dibenzo-30-Crown-10-Ether, di-(*m*-*tert*-butyl)	3.32	5.87	5.25	2.55	1.93
61 1-(2,3,5-tribenzoyl-D-ribofuranosyl)-2-NO_2-imidazole	1.19	4.53	4.85	3.34	3.66
62 H8-pyrazino[2′,1′:6,1]pyrid[3,4-*b*]indol, benzamidoethyl	1.52	3.55	3.19	2.03	1.67

In most cases studied here using the ALOGP method, the correlation coefficient and other statistics were closely similar for the training and test sets. This indicates that the size of the current database (∼9K compounds) is nearly optimal, and rederiving the atomic parameters for even bigger databases is unlikely to improve predictions. *o*-Hydroxybenzoic acids (salicylic acid derivatives) and *o*-hydroxybenzaldehyde consistently showed negative deviation between −0.4 and −1.2 with the ALOGP method but were better predicted by the CLOGP method. Intramolecular hydrogen bonds such as those between the carbon (C═O) and the hydroxyl groups are not represented or corrected in the ALOGP method, but the CLOGP method has a hydrogen bonding correction factor. With the ALOGP predictions somewhat lower negative deviations (between −0.1 and −0.9) were observed among the *o*-nitrophenols. When the hydroxyl and carbonyl groups were separated by single bonds, deviations were both positive and negative, indicating the hydrogen bonding is less important because of the conformational entropy associated with the bond rotation. Predictions with the ALOGP method could be improved with some atom subclassification, for example, for the carbonyl oxygen and hydroxyl hydrogen when they are separated by double or aromatic bonds. In principle, any structural feature causing systematic deviation can be accounted for in the ALOGP method by introducing new subtypes, avoiding correlated atom types. Introduction of a new atom class by identifying such structural features will be statistically more significant than a totally

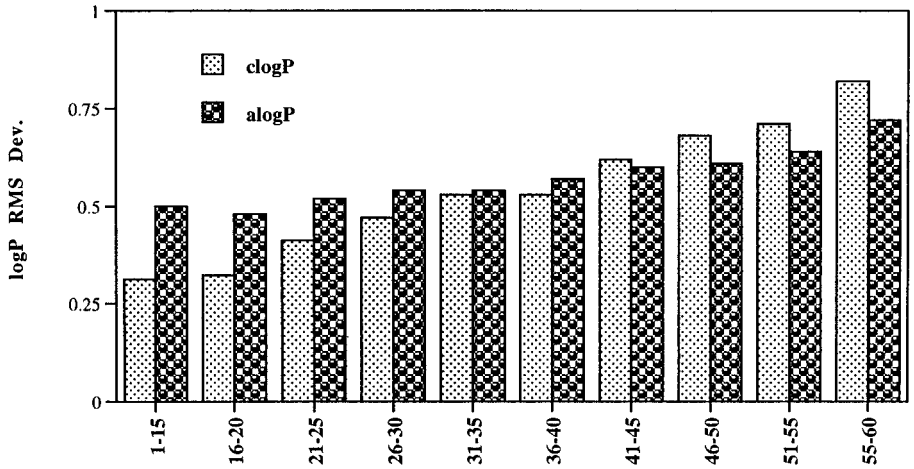

Figure 4. Performance of the ALOGP and CLOGP methods in compounds of different size in the training set.

different atom classification scheme with a considerably larger number of atom types.

The ALOGP method is intrinsically atomistic, and hence, it is useful for drug design in many ways. Some examples include the estimation of local or overall hydrophobicity in a molecule for physicochemical property-based 2D- or 3D-QSAR,[16] estimation of molecular similarity,[11] molecular mimicry,[32,33] automated and semiautomated pharmacophore modeling,[34,35] evaluation of hydrophobic potential surfaces,[20] and scoring protein−ligand interactions.[22] "Atomic" parameters were also derived on the basis of the CLOGP estimates of lipophilicity. These parameters have been used in deriving lipophilicity fields[36,37] and 3D-QSAR models[38] and in empirically scoring ligand−receptor interactions.[36,39] Thus, "atomization" of lipophilicity is a conceptual tool that finds increasing applicability in molecular design.

Concluding Remarks

There have been reports in the literature[12,40] questioning the validity of fragment-atom-based approaches for log *P* prediction. High correlations (∼0.9 or higher) obtained in the present study clearly demonstrate the validity of these approaches in the case of small organic molecules, indicating that the long range interactions are probably not very important for predicting log *P* in a majority of cases of small molecules. The CLOGP method showed marginally better performance when all types of molecules are considered as a whole. Despite relatively fewer variables used in the ALOGP method, its performance is at least as good as the CLOGP method for a majority of the compound subclasses and is very close to the CLOGP method in overall performance.

The main limitation of the CLOGP method is that it cannot be applied to a considerable portion of the database, due to the presence of undefined fragments. In contrast, the ALOGP method is applicable to most neutral organic compounds and selective charged compounds in the dataset. The CLOGP method uses a large number of parameters and correction factors. To develop these parameters, molecules with about 20 atoms or less were often considered. Thus, the CLOGP method gives much better performance for molecules with about 20 atoms or fewer. However, as larger molecules (with 40−60 atoms or more) are considered, predictions of this method become more error-prone with rms deviations of greater than 1 log unit for some molecules with over 60 atoms. The ALOGP method gives a stable performance in all classes of organic compounds tested, with much less variability in the statistical quality of results among different subclasses. The CLOGP method gives much larger deviations in the cases of alcohols, primary amines, unblocked peptides and nucleosides.

Acknowledgment. The authors would like to thank Dr. Timothy Harvey and Dr. Steve Jordan for their comments and suggestions.

References and Notes

(1) Hansch, C. In *Correlation Analysis in Chemistry*; Chapman, N. B., Shorter, J., Eds.; Wiley: New York, 1978.
(2) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry*; Wiley: New York, 1979.
(3) Pliska, V.; Testa, B.; Waterbeemd, H. *Lipophilicity in drug action and toxicology*; Pliska, V., Testa, B., Waterbeemd, H., Eds.; VCH Publishers: New York, 1996.
(4) Sangster, J. *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*; John Wiley & Sons: Chichester, 1997; Vol. 2.
(5) Hansch, C. *Acc. Chem. Res.* **1993**, *26*, 147−153.
(6) Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1975**, *18*, 865−868.
(7) Leo, A. J. *CLOGP*, version 3.63; Daylight Chemical Information Systems: Irvine, CA, 1991.
(8) Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity*; VCH Publishers: New York, 1992.
(9) Klopman, G.; Namboodiri, K.; Schochet, M. *J. Comput. Chem.* **1985**, *6*, 28−38.
(10) (a) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565−577. (b) Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.
(11) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(12) Bodor, N.; Buchwald, P. *J. Phys. Chem.* **1997**, *101*, 3404−3412.
(13) Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80−90.
(14) Viswanadhan, V. N.; Rami Reddy, M.; Bacquet, R. J.; Erion, M. D. *J. Comput. Chem.* **1993**, *14*, 1019−1026.
(15) Ghose, A. K.; Crippen, G. M. *J. Med. Chem.* **1985**, *28*, 333−46.
(16) Ghose, A. K.; Crippen, G. M.; Revankar, G. R.; McKernan, P. A.; Smee, D. F.; Robins, R. K. *J. Med. Chem.* **1989**, *32*, 746−756.
(17) Ghose, A. K.; Crippen, G. M. *Mol. Pharmacol.* **1990**, *37*, 725−34.
(18) Viswanadhan, V. N.; Ghose, A. K.; Weinstein, J. N. *Biochim. Biophys. Acta* **1990**, *1039*, 356−366.
(19) Viswanadhan, V. N.; Ghose, A. K.; Hanna, N. B.; Matsumoto, S. S.; Avery, T. L.; Revankar, G. R.; Robins, R. K. *J. Med. Chem.* **1991**, *34*, 526−532.
(20) Furet, P.; Sele, A.; Cohen, N. C. *J. Mol. Graphics* **1988**, *6*, 182−189.
(21) Heiden, W.; Moekel, G.; Brickmann, J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 503−514.
(22) Viswanadhan, V. N.; Rami Reddy, M.; Wlodawer, A.; Varney, M. D.; Weinstein, J. N. *J. Med. Chem.* **1996**, *39*, 705−712.
(23) Leo, A. J. in *Methods of Calculating Partition Coefficients*; Hansch, C., Sammes, P. G., Taylor, Eds.; Pergamon Press: New York, 1990; Vol. 4, pp 295−319.
(24) Mannhold, R.; Rekker, R. F.; Sonntag, C.; Laak, A. M.; Dross, K.; Polymeropoulos, E. E. *J. Pharm. Sci.* **1995**, *84*, 1410−1419.

(25) *Star list Database*; Biobyte Corporation: Pomona, CA, 1997.
(26) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(27) *Galaxy*, version 2.3; American Molecular Technologies, San Antonio, TX, 1997.
(28) *MacLogP*, version 2.0; Biobyte Corp. Claremont, CA, 1997.
(29) Krishnamurthy, E. V.; Sen, S. K. *Computer Based Numerical Algorithms*; East-West Press: New Delhi, 1976.
(30) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(31) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(32) Ghose, A. K.; Viswanadhan, V. N.; Sanghvi, Y. S.; Nord, L. D.; Willis, R. C.; Revankar, G. R.; Robins, R. K. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8242−8246.
(33) Ghose, A. K.; Sanghvi, Y. S.; Larson, S. B.; Revankar, G. R.; Robins, R. K. *J. Am. Chem. Soc.* **1990**, *112*, 3622−3628.
(34) Ghose, A. K.; Logan, M. E.; Treasurywala, A. M.; Wang, H.; Wahl, R. C.; Tomczuk, B.; Gowravaram, M.; Jaeger, E. P.; Wendoloski, J. J. *J. Am. Chem. Soc.* **1995**, *117*, 4671−4682.
(35) Ghose, A. K.; Wendoloski, J. J. In *3D-QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic: Dordrecht, The Netherlands, 1997.
(36) Wireko, F. C.; Kellogg, G. E.; Abraham, D. J. *J. Med. Chem.* **1991**, *34*, 758−767.
(37) Abraham, D. J.; Kellogg, G. E. *J. Comput.-AIded Mol. Des.* **1994**, *8*, 41−49.
(38) Oprea, T. I.; Waller, C. L.; Marshall, G. R. *Drug Des. Discovery* **1994**, *12*, 29−51.
(39) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−52.
(40) Bodor, N.; Gabanyi, Z.; Wong, C.-K. *J. Am. Chem. Soc.* **1989**, *111*, 3783.