

# Prediction of Physicochemical Parameters by Atomic Contributions

Scott A. Wildman and Gordon M. Crippen\*

College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received April 19, 1999

We present a new atom type classification system for use in atom-based calculation of partition coefficient ( $\log P$ ) and molar refractivity (MR) designed in part to address published concerns of previous atomic methods. The 68 atomic contributions to  $\log P$  have been determined by fitting an extensive training set of 9920 molecules, with  $r^2 = 0.918$  and  $\sigma = 0.677$ . A separate set of 3412 molecules was used for the determination of contributions to MR with  $r^2 = 0.997$  and  $\sigma = 1.43$ . Both calculations are shown to have high predictive ability.

## INTRODUCTION

For many years, the octanol–water partition coefficient ( $\log P$ ) has been used as a measure of lipophilicity. Pioneering work by Hansch and Leo has led to the use of  $\log P$  in quantitative structure–activity relation methods,<sup>1</sup> as a general description of cell permeability.  $\log P$  has since become a standard property determined for potential drug molecules, noting Lipinski's "rule of 5".<sup>2</sup> The invention of combinatorial synthesis and high-throughput screening has led to vast numbers of new potential drugs being produced and tested each day. With this technology comes an increasing need for quick and accurate determination or calculation of  $\log P$ .

A recent review<sup>3</sup> summarizes the many different approaches to calculating  $\log P$  found in the literature and includes a list of available  $\log P$  calculation software. The two main calculation schemes remain the fragment methods and the atom-based approach. The fragment method, developed by Rekker<sup>4</sup> and by Hansch<sup>5</sup> and refined by Klopman,<sup>6</sup> Kudo,<sup>7</sup> and others,<sup>8</sup> has become a standard calculation and is available in many common software packages. This method involves the estimation of  $\log P$  based on the contributions of functional groups and fragments attached to a base molecule. The atomic contribution method was developed by Broto<sup>9</sup> and by Ghose and Crippen<sup>10</sup> and later refined by Ghose and co-workers.<sup>11,12</sup> This method assigns to the individual atoms in the molecule additive contributions to molecular  $\log P$ . This is accomplished by classifying atoms into chemically distinct types and fitting the contributions on a data set of experimentally determined  $\log P$  values. Localization of lipophilicity to specific atoms in the molecule allows for extended use of the  $\log P$  calculation, including generation of a molecular lipophilicity map.

In the past few years, there have been many comments in the literature about various problems with atomic  $\log P$  calculation methods. Several references have been made to ambiguity in the classification system, the large number of atom types (and therefore adjustable parameters), unrealistic values of some atom contributions, a perceived failure at prediction, and bias toward underestimation of  $\log P$ . We present a novel definition of atomic  $\log P$  atom types and contributions designed in part to address these complaints.

Additionally, calculation of molar refractivity (MR), a common descriptor accounting for molecular size and polarizability,<sup>10</sup> is accomplished in the same manner.

## METHODS

The partition coefficient ( $\log P$ ) and molar refractivity (MR) of small molecules can be calculated as the sum of the contributions of each of the atoms in the molecules. While these properties are not strictly additive, intramolecular interactions can be accounted for by classifying atoms into different types based on attached and neighboring atoms. Following classification, the property can be calculated as

$$P_{\text{calc}} = \sum_i n_i a_i \quad (1)$$

where  $P_{\text{calc}}$  is the property to be calculated ( $\log P$  or MR),  $n_i$  is the number of atoms of type  $i$  present in the molecule, and  $a_i$  is the contribution for atoms of type  $i$ .

The classification system was developed by declaring different types for atoms with different nearest neighbors. For instance, the first carbon of  $\text{CH}_3\text{C}$  is treated as different from that of  $\text{CH}_3\text{N}$ . Atoms were further divided on the basis of second neighbors and aromaticity; however, the resulting list of atom types was thought to be too long for practical use. Those types with similar neighbors and those with similar atomic contributions to  $\log P$ , as shown with a test fit of experimental data, were grouped together into a single common type. The reduced set of atom types was designed such that only atoms with similar chemical nature and the same approximate atomic contribution to  $\log P$  were combined. Atom types were also combined so that no type was too poorly represented. In this way, redundant atom types and intercorrelations were reduced in the classification system.

The final atom classification system has 68 basic atom types. It is comprehensive for the elements commonly found in organic molecules (C, H, N, O, S, P, halogens), and also includes metals and noble gasses. It is designed such that each atom present in the molecule will match one and only one atom type, thereby removing ambiguity from the typing system. In order to ensure that all organic molecules can be

**Table 1.** Atom Type Descriptions and Contributions

type	descriptions	SMARTS <sup>a</sup>	log <i>P</i>	obsd	MR	obsd
C1	1°, 2° aliphatic	'[CH4]', '[CH3]C', '[CH2](C)C'	0.1441	5080	2.503	2560
C2	3°, 4° aliphatic	'[CH](C)(C)C', '[C](C)(C)(C)C'	0.0000	1014	2.433	587
C3 <sup>b</sup>	1°, 2° heteroatom	'[CH3]([N,O,P,S,F,Cl,Br,I])'	-0.2035	5452	2.753	1513
C4 <sup>c</sup>	3°, 4° heteroatom	'[CH2X4]([N,O,P,S,F,Cl,Br,I])'	-0.2051	2431	2.731	847
		'[CH1X4]([N,O,P,S,F,Cl,Br,I])'				
C5	C = heteroatom	'[C] = [A#X]'	-0.2783	5758	5.007	961
C6	C = C aliphatic	'[CH2] = C', '[CH1](=C)A', '[CH0](=C)(A)A', '[C](=C)=C'	0.1551	1062	3.513	570
C7	acetylene, nitrile	'[CX2]#A'	0.00170	465	3.888	215
C8	1° aromatic carbon	'[CH3]c'	0.08452	1085	2.464	231
C9	1° aromatic heteroatom	'[CH3][a#X]'	-0.1444	200	2.412	9
C10	2° aromatic	'[CH2X4]a'	-0.0516	2016	2.488	247
C11	3° aromatic	'[CHX4]a'	0.1193	825	2.582	114
C12	4° aromatic	'[CH0X4]a'	-0.0967	456	2.576	36
C13 <sup>d</sup>	aromatic heteroatom	'[cH0]-[!](C,N,O,S,F,Cl,Br,I)'	-0.5443	30	4.041	33
C14	aromatic halide	'[c][#9]'	0.0000	350	3.257	25
C15	aromatic halide	'[c][#17]'	0.2450	1329	3.564	61
C16	aromatic halide	'[c][#35]'	0.1980	298	3.180	47
C17	aromatic halide	'[c][#53]'	0.0000	124	3.104	19
C18	aromatic	'[cH]'	0.1581	7915	3.350	926
C19	aromatic bridgehead	'[c](:a)(:a):a'	0.2955	1179	4.346	64
C20	4° aromatic	'[c](:a)(:a)-a'	0.2713	514	3.904	19
C21	4° aromatic	'[c](:a)(:a)-C'	0.1360	5050	3.509	703
C22	4° aromatic	'[c](:a)(:a)-N'	0.4619	3428	4.067	95
C23	4° aromatic	'[c](:a)(:a)-O'	0.5437	2427	3.853	167
C24	4° aromatic	'[c](:a)(:a)-S'	0.1893	851	2.673	21
C25	4° aromatic	'[c](:a)(:a) = C', '[c](:a)(:a) = N', '[c](:a)(:a) = O'	-0.8186	661	3.135	4
C26	C = C aromatic	'[C](=C)(a)A', '[C](=C)(c)a', '[CH](=C)a', '[C] = c'	0.2640	344	4.305	57
C27 <sup>e</sup>	aliphatic heteroatom	'[CX4]([N,O,P,S,F,Cl,Br,I])'	0.2148	24	2.693	101
CS	carbon supplemental	'[#6]' not matching any basic C type	0.08129	0	3.243	0
H1	hydrocarbon	'[#1][#6]', '[#1][#1]'	0.1230	9852	1.057	3361
H2 <sup>f</sup>	alcohol	'[#1]O[CX4]', '[#1]Oc', '[#1]O[!(C,N,O,S)]', '[#1]![C,N,O)]'	-0.2677	1744	1.395	493
H3	amine	'[#1][#7]', '[#1]O[#7]'	0.2142	4954	0.9627	216
H4	acid	'[#1]OC = [#6]', '[#1]OC = [#7]', '[#1]OC = O', '[#1]OC = S', '[#1]OO', '[#1]OS'	0.2980	622	1.805	81
HS	hydrogen supplemental	'[#1]' not matching any basic H type	0.1125	0	1.112	0
N1	1° amine	'[NH2+0]A'	-1.0190	1014	2.262	70
N2	2° amine	'[NH+0](A)A'	-0.7096	2040	2.173	81
N3	1° aromatic amine	'[NH2+0]a'	-1.0270	696	2.827	30
N4	2° aromatic amine	'[NH+0](A)a', '[NH+0](a)a'	-0.5188	1346	3.000	21
N5	imine	'[NH+0] = A', '[NH+0] = a'	0.08387	48	1.757	2
N6	substituted imine	'[N+0](=A)A', '[N+0](=A)a', '[N+0](=a)A', '[N+0](=a)a'	0.1836	1010	2.428	40
N7	3° amine	'[N+0](A)(A)A'	-0.3187	1720	1.839	99
N8	3° aromatic amine	'[N+0](a)(A)A', '[N+0](a)(a)A', '[N+0](a)(a)a'	-0.4458	492	2.819	21
N9	nitrile	'[N+0]#A'	0.01508	382	1.725	72
N10	protonated amine	'[NH3+*]', '[NH2+*]', '[NH+*]'	-1.950	189		0
N11	unprotonated aromatic	'[n+0]'	-0.3239	2819	2.202	96
N12	protonated aromatic	'[n+*]'	-1.119	104		0
N13	4° amine	'[NH0+*](A)(A)(A)A', '[NH0+*](=A)(A)A', '[NH0+*](=A)(A)a', '[NH0+*](=[#6])(=[#7])=N'	-0.3396	1075	0.2604	75
N14	other ionized nitrogen	'[N+*]#A', '[N-*]', '[N+*](=[N-*])=N'	0.2887	17	3.359	4
NS	nitrogen supplemental	'[#7]' not matching any basic N type	-0.4806	0	2.134	0
O1	aromatic	'[o]'	0.1552	413	1.080	56
O2	alcohol	'[OH]', '[OH2]'	-0.2893	2317	0.8238	526
O3	aliphatic ether	'[O](C)C', '[O](C)[A#X]', '[O]([A#X])[A#X]'	-0.0684	2376	1.085	925
O4	aromatic ether	'[O](A)a', '[O](a)a'	-0.4195	1957	1.182	130
O5	oxide	'[O]=[#8]', '[O]=[#7]', '[OX1-*][#7]'	0.0335	1272	3.367	88
O6	oxide	'[OX1-*][#16]'	-0.3339	718	0.7774	34
O7 <sup>g</sup>	oxide	'[OX1-*]![N,S)]'	-1.189	138	0.000	24
O8	aromatic carbonyl	'[O]=c'	0.1788	657	3.135	4
O9	carbonyl aliphatic	'[O]=[CH]C', '[O]=C(C)C', '[O]=C(C)[A#X]', '[O]=[CH]N', '[O]=[CH]O', '[O]=[CH2]', '[O]=[CX2]=O'	-0.1526	3163	0.000	767
O10	carbonyl aromatic	'[O]=[CH]c', '[O]=C(C)c', '[O]=C(c)c', '[O]=C(c)[a#X]', '[O]=C(c)[A#X]', '[O]=C(C)[a#X]'	0.1129	1534	0.2215	125
O11	carbonyl heteroatom	'[O]=C([A#X])[A#X]', '[O]=C([A#X])[a#X]', '[O]=C([a#X])[a#X]'	0.4833	1063	0.3890	43
O12	acid	'[O-1]C(=O)'	-1.326	187		0
OS	oxygen supplemental	'[#8]' not matching any basic O type	-0.1188	0	0.6865	0
F	fluorine	'[#9-0]'	0.4202	814	1.108	120
Cl	chlorine	'[#17-0]'	0.6895	1613	5.853	630
Br	bromine	'[#35-0]'	0.8456	366	8.927	250
I	iodine	'[#53-0]'	0.8857	137	14.02	61
Hal <sup>h</sup>	ionic halogens	'[#9-*]', '[#17-*]', '[#35-*]', '[#53-*]', '[#53+*]'	-2.996	19		0

**Table 1** (Continued)

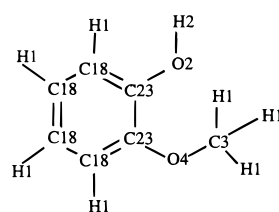
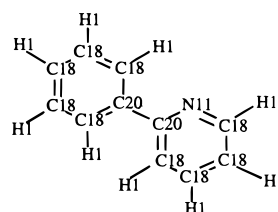
type	descriptions	SMARTS <sup>a</sup>	log <i>P</i>	obsd	MR	obsd
P	phosphorous	'[#15]'	0.8612	202	6.920	40
S1	aliphatic	'[S-0]'	0.6482	849	7.591	119
S2	ionic sulfur	'[S-*]', '[S+*]'	-0.0024	781	7.365	37
S3	aromatic	'[s]'	0.6237	295	6.691	37
Me1 <sup>i</sup>	all remaining <i>p</i> -block elements		-0.3808	40	5.754	139
Me2 <sup>j</sup>	all remaining <i>d</i> -block elements		-0.0025	29		0

<sup>a</sup> The SMARTS presented here were implemented using MOE version 1998.10. A different version of MOE may result in slightly different interpretations of the SMARTS. Some SMARTS in the table have been rewritten for brevity using the NOT operator (!) and atom lists indicated by parentheses within square brackets: [(atom list)]. We expect these notations will be implemented in future versions of MOE SMARTS. Types such as C13 are meant to include all possible heteroatoms not explicitly defined in other types, not only those specifically listed in the footnote. Heteroatoms not listed in the footnotes were not present in the training set. In all cases, the complete list of actual SMARTS used is listed below.

<sup>b</sup> '[CH3][A#N]', '[CH2X4][A#N]', '[CH3][#15]', '[CH2X4][#15]', '[CH3][#16]', '[CH2X4][#16]', '[CH3][#53]', '[CH2X4][#53]', not matching '[CH2X4]a'. <sup>c</sup> '[CHX4][A#N]', '[CH0X4][A#N]', '[CHX4][#15]', '[CH0X4][#15]', '[CHX4][#16]', '[CH0X4][#16]', '[CHX4][#53]', '[CH0X4][#53]', not matching '[CHX4]a' or '[CH0X4]a'. <sup>d</sup> '[c][#5]', '[c][#14]', '[c][#15]', '[c][#33]', '[c][#34]', '[c][#50]', '[c][#80]'. <sup>e</sup> '[CX4][#X]' not matching '[CX4][#N]', '[CX4][#16]', '[CX4][#15]', '[CX4][#53]'. <sup>f</sup> '[#1]O[CX4]', '[#1]Oc', '[#1]O[#1]', '[#1]O[#5]', '[#1]O[#14]', '[#1]O[#15]', '[#1]O[#33]', '[#1]O[#50]', '[#1][#5]', '[#1][#14]', '[#1][#15]', '[#1][#16]', '[#1][#50]'. <sup>g</sup> '[OX1-\*][#15]', '[OX1-\*][#33]', '[OX1-\*][#43]', '[OX1-\*][#53]'.

<sup>h</sup> Although there were no group I or II cations in the training set, Hal is the most appropriate atom type. <sup>i</sup> Present in the training set are: B, Si, Ga, Ge, As, Se, Sn, Te, Pb, Ne, Ar, Kr, Xe, Rn. <sup>j</sup> Present in the training set are Fe, Cu, Zn, Tc, Cd, Pt, Au, Hg.

**Table 2.** Example of Atom Classification and Properties Calculation

					
type	log <i>P</i>	MR	type	log <i>P</i>	MR
C3	-0.2035	2.753	9 × C18	0.1581	3.350
4 × C18	0.1581	3.350	2 × C20	0.2713	3.904
2 × C23	0.5437	3.853	9 × H1	0.1230	1.057
7 × H1	0.1230	1.057	N11	-0.3239	2.202
H2	-0.2677	1.395	calcd	2.75	50.39
O2	-0.2893	0.8238	expt	2.63	49.67
O4	-0.4195	1.182			
calcd	1.40	34.66			
expt	1.32	34.66			

completely assigned, four supplementary types were added, one each for C, H, N, and O. These types, meant to match any atoms not otherwise classified, were not included in the fitting process and have values assigned as the average values of the basic types for each element. Additionally, to allow for portability and simple implementation of the classification system, it is presented in SMARTS<sup>13</sup> strings in Table 1. An example of the classification system is presented in Table 2.

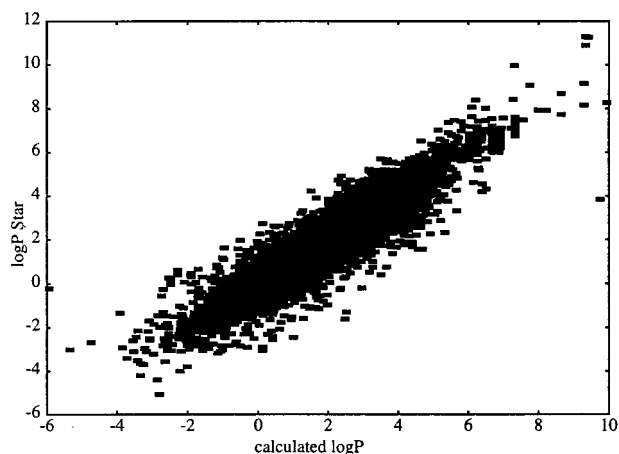
The SMARTS system used here is that included in the Molecular Operating Environment, version 1988.10 (MOE),<sup>14</sup> which has slightly different notation from original SMARTS. While SMARTS is essentially an extension of the well-known line notation SMILES, some notation will be reviewed here. In Table 1, each SMARTS is enclosed by single quotes in order to clearly separate the strings in cases where more than one string is used to define an atom type. In all cases, the first atom in the string is the only atom classified by that atom type. This is an important distinction from similar typing schemes that would match the entire SMARTS. Square brackets are used to group together descriptors of each atom specified in the string as in original SMARTS. The descriptors present in our string have the following designations in MOE SMARTS. 'A' and 'a' are used to represent any non-hydrogen aliphatic and aromatic element,

respectively. This includes carbon. The term '[#X]' is used to specify any non-hydrogen, non-carbon element, and aromaticity may be specified by including the appropriate 'A' or 'a', as in '[A#X]' or '[a#X]'. Using '[#N]' specifies any element in the subset of elements O, N, F, Cl, and Br, subject to the same aromaticity control. The common organic elements (B, C, N, O, P, S, F, Cl, Br, and I) may be specified by their symbol, where the lower case (b, c, n, o, p, s) would indicate aromaticity. Other elements may be specified using '[# atomic number]', with 'A' and 'a' again as aromaticity controls, such that '[a#7]' indicates an aromatic nitrogen, as does 'n'. The symbols -, =, #, and : indicate single, double, triple, and aromatic bonds, respectively. Absence of a specific bond symbol implies a single or aromatic bond. The number of attached hydrogens is specified using H, so '[CH4]' would match the carbon in methane. Charge is indicated with + and - such that '[#7+\*]' indicates a positively charged nitrogen. 'X' without '#' is used to specify the total number of bonds, including those to implicit hydrogens, so '[CX4]' would match the carbon in methane as well as the quaternary carbon of a *tert*-butyl group. All other notation follows that of the original SMARTS. A complete description of MOE SMARTS is available in the MOE documentation.

The Medicinal Chemistry Project<sup>15</sup> of Hansch and Leo has compiled a list of now almost 35 000 experimentally determined log *P* values and SMILES. From this, the "star list" of "very accurately" determined values was used to form a training database of 9920 molecules. These molecules were entered into MOE, and the classification system was implemented. The resulting list of atom types present in each molecule was used to determine the contribution of each atom type by linear least squares fit of the experimental data. Following this initial fit, several of the carbon types (C1, C2, C6, C8, C14, C15, C16, C17, C18, C19, C20, C21, C26, C27), not all of which originally had negative contributions, were constrained to be positive, and the data were refit in order to generate more reasonable atomic contributions.

A database of 3412 experimentally determined MR values was determined using the Clausius-Mossotti equation:

$$R_m = \frac{M}{\rho} \frac{n_r^2 - 1}{n_r^2 + 2} \quad (2)$$



**Figure 1.** Correlation plot for the fit of 9920 log *P* Star values:  $r^2 = 0.918$ ;  $\sigma = 0.677$ .

where  $n_r$  is the refractive index,  $\rho$  is the density, and  $M$  is the molar mass of the substance, with these data taken from the *CRC Handbook*.<sup>16</sup> These molecules were subjected to the same atom classification system, and the contribution of each atom type was again determined by linear least squares fit. Again several atom types (N5, N6, N11, O7, O9, O10, O11) were constrained to be positive, and the data were refit to generate more reasonable MR contributions.

## RESULTS AND DISCUSSION

**log *P*.** The set of 9920 experimental log *P* values was fit with 68 adjustable parameters by linear least squares with  $r^2 = 0.919$  and  $\sigma = 0.673$ . Figure 1 shows the correlation of the fit. There was no significant change in overall fit when the contributions of several types were constrained to be positive ( $r^2 = 0.918$ ,  $\sigma = 0.677$ ). The predictive  $r^2 = 0.914$  was determined as in ref 12, as a measure of robustness. The list of contributions for this final fit and number of molecules containing each atom type is included in Table 1.

The fit of our training set is certainly comparable to that reported for other log *P* calculation methods, both by fragment methods and by atomic contributions, and the high predictive  $r^2$  is a clear indication of the predictive power of this model and classification system. However, there are some features of certain molecules that are neglected in this system. As we calculate the property values solely from the atomic contributions, there are no correction factors or additive constants employed by this system. Since the method only matches atom connectivity and does not consider 3-D structure, we do not explicitly account for long-range interactions or intramolecular hydrogen bonds. At this time, the method also does not treat chirality, resulting in the *R*- and *S*-isomers of a molecule having the same calculated log *P* and MR. The standard deviation of 0.667 is slightly higher than reported for several other methods ( $\sim 0.4$ ), likely due to the exclusion of 3-D structure factors.

This update of the atom classification system and atom type contributions was intended to address many of the concerns regarding previous atomic methods of log *P* calculation. There is a general implication in the literature that atomic methods employ an excessive number of atom types in the classification system. In the most recent edition of ALOGP<sup>12</sup> there are 120 different atom types, 44 of which describe carbon. At first glance it may seem unreasonable

to claim 44 different kinds of carbon; however, these types must account for all possible variations in bonding and neighbors. The fundamental structure of the molecule in our method is encoded in the atom types, just as it is in the fragments present in the Rekker and CLOGP methods. By way of comparison, Rekker and co-workers<sup>8</sup> used at least 169 fragments. Unfortunately, we have been unable to locate in the recent literature the number of fragments and correction factors used in CLOGP.

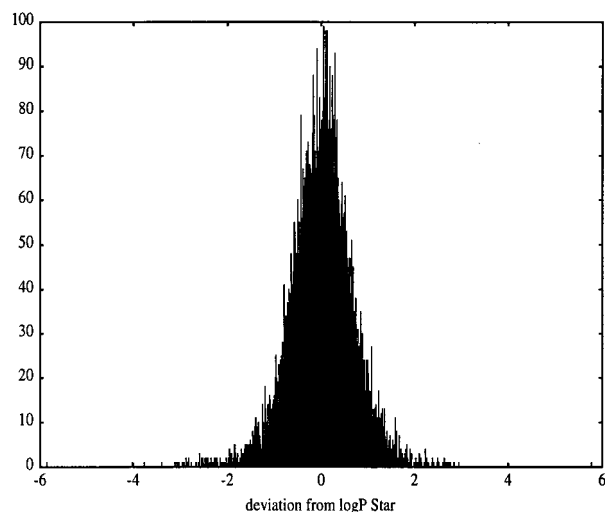
Nonetheless, the number of basic types has been reduced to 68, with only 27 types for carbon. The system is focused on the typical elements found in organic compounds (C, H, N, O, P, S, halogens), but also includes metals and noble gasses. The four supplementary types were added after the fitting process in order to ensure that all molecules can be completely assigned. It is a common problem that as missing fragments are encountered, CLOGP cannot evaluate a log *P* for the compound.<sup>12,17</sup> While we do not claim the use of such supplementary types will always produce an accurate result, they will at least provide the user with a reasonable estimate of the log *P* and should be accompanied by an appropriate warning message in any implementation of the method.

We also note that several earlier methods train a large number of adjustable parameters on a rather small data set, with up to 424 parameters for 1465 molecules.<sup>7</sup> Not only have we reduced the number of parameters, but we have chosen a large and complete data set of 9920 molecules for training.

A related comment on previous atomic methods is that the contributions of some atom types seem unrealistic. In the most recent version of ALOGP,<sup>12</sup> for example, the carbon of methane is found to be hydrophilic ( $-1.56$ ) while the hydrogens are hydrophobic ( $+0.73$ ). While the overall log *P* works out to a reasonable 1.38 (vs 1.09 experimental), the contributions of the atom types are clearly suspect. While there are undoubtedly some real instances where carbon may be thought of as hydrophilic, the concerns of unrealistic contributions are valid. Such situations are a direct result of the least squares fitting method, which is unconcerned with the specific values of individual contributions. When considering the whole molecule, these contributions cancel, allowing for an accurate calculation of molecular log *P*. If the goal is only the final log *P* value, unrealistic individual contributions need not be a concern. However, if one needs to estimate the lipophilicity of a part of a molecule, it is important the individual atomic contributions be more reasonable while maintaining an accurate overall calculated log *P*. To address this, the contributions of several aliphatic carbon types (C1, C2, C6, C8, C14, C15, C16, C17, C18, C19, C20, C21, C26, C27) were constrained to be positive (hydrophobic) during the fitting process. By imposing this limit on only a few carbon types, it was possible to generate more reasonable contribution values for the remaining carbon types as well as the other elements without sacrificing overall log *P* accuracy ( $r^2 = 0.917$ ,  $\sigma = 0.678$  constrained;  $r^2 = 0.918$ ,  $\sigma = 0.674$  unconstrained).

Rekker<sup>18</sup> notes a preferential underestimation of log *P* by atomic methods. While his test set is admittedly small, this trend has also been noticed by other authors. In our large training database, the average experimental log *P* is 1.878. Using our current classifications and contributions, the average calculated value is 1.870. Figure 2 is a histogram





**Figure 2.** Histogram of the deviation of the calculated value from log *P* Star: mean =  $-0.009$ ; skew =  $0.17$ .

of the deviation of the calculated values from the experimental results, with a mean of  $-0.009$  and skew =  $0.17$ . This hardly constitutes a consistent underestimation.

Previous methods have used diagrams of various forms and tables with assorted special notation to describe their fragments or their atomic classifications. These may be a useful summary for publication but can often lead to confusion in the implementation of method. Leo notes that "the greatest difficulty may arise from some ambiguity in the classification."<sup>19</sup> We have presented our atomic classification system as SMARTS (Table 1) to allow for easier portability and reduced ambiguity. Additionally, the rules and therefore the SMARTS were designed to be comprehensive, so that all molecules can be treated, and also unique, in the sense that each atom in any molecule will match one and only one atom type. This specification is important for future work as well as helping to remove ambiguity in the classification system.

**Molar Refractivity.** The set of 3412 experimental MR values was fit with 68 adjustable parameters from the same classification system by linear least squares with  $r^2 = 0.997$  and  $\sigma = 1.42$ . Again there was no significant change in overall fit when the contributions of several types were constrained to be positive ( $r^2 = 0.997$ ,  $\sigma = 1.43$ ). A high predictive  $r^2$  value of 0.994 again shows the predictive power of the method and classification system. A MR contribution could not be determined for several atom types (N12, O12, Ha1, Me2) as these types were not present in the data. However, any atoms that do not match one of the basic types present may be classified as one of the supplemental types (Ha1 and Me2 atoms may be classified as Me1), and a reasonable MR will most likely be determined.

**Accuracy of the Data.** It is easily recognized that substantial errors or bias in the training data set can lead to difficulty in fitting. For this reason we have chosen what we believe to be the most accurate and complete sets of log *P* and MR data commonly available. However, careful inspection of the data revealed that these collections are not perfect. All errors found in the training data sets were corrected.

Interpretation of SMILES strings used for data input proved to be hazardous. There were several difficulties, the

most notable being correct identification of aromaticity, especially that involving heteroatoms. We suspect this problem is common to many users, but not always identified.

In some cases, there are also errors in the entered data. For example, the *CRC Handbook*<sup>16</sup> lists the density of benzoic anhydride as 1.99 rather than the actual density of 1.199. Clearly, this is simply a typographical error, but as we were able to identify 35 similar errors from a small sample of our 3412 compound test set, such errors do not seem to be uncommon. We found substantially fewer errors in the MedChem Project Star Database,<sup>15</sup> but an inspection of the literature references included with the database show roughly a 5–10% error rate for a small random sample of  $\sim 100$  molecules. We also note that a fair portion of the Star Database, roughly 16%, is referenced to "unpublished results".

## CONCLUSIONS

We have presented a redesigned atomic classification system and contributions for calculation of log *P* and MR. Calculation of both properties is comparable to or better than that of other methods, and the fit, completed with only 68 adjustable parameters, is shown to be robust. The training set of 9920 molecules for log *P* calculation gave  $r^2 = 0.918$ ,  $\sigma = 0.677$ , and predictive  $r^2 = 0.914$ . For MR calculation, a 3412 molecule data set gave  $r^2 = 0.997$  and  $\sigma = 1.43$  with a predictive  $r^2 = 0.994$ . The classification rules are presented in SMARTS in order to be unambiguous and allow for simple portability. We have addressed several important concerns regarding atomic methods presented in the literature, with a larger training set and fewer adjustable parameters, but without sacrificing accuracy.

## ACKNOWLEDGMENT

S.A.W. was supported in part by the Lyons, Arbour Henry, and Blake Fellowships of the College of Pharmacy, University of Michigan. This work was supported by the Chemical Computing Group Inc., National Science Foundation Grant DBI-9614074, and the Vahlteich Research Award Fund (College of Pharmacy, University of Michigan).

## REFERENCES AND NOTES

- (1) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525.
- (2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Freeny, P. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3.
- (3) Carrupt, P.-A.; Testa, B.; Gaillard, P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley: New York, 1997; Vol. 11, Chapter 5, pp 241–315.
- (4) Nys, G. G.; Rekker, R. F. *Chim. Ther.* **1973**, *8*, 521. Nys, G. G.; Rekker, R. F. *Eur. J. Med. Chem.* **1974**, *9*, 361.
- (5) Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. Calculation of Hydrophobic Constant (log *P*) from  $\pi$  and  $f$  constants. *J. Med. Chem.* **1975**, *18*, 865.
- (6) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log *P* Calculations Based on an Extended Group Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752.
- (7) Suzuki, T.; Kudo, Y. Automatic log*P* Estimation Based on Combined Additive Modelling Methods. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 155.
- (8) Mannhold, R.; Rekker, R. F.; Dross, K.; Greetje, B.; de Vries, G. The lipophilic behaviour of organic compounds: 1. An updating of the hydrophobic fragmental constant approach. *QSAR* **1998**, *17*, 517.

- (9) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem.* **1984**, *19*, 71.
- (10) 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**, *4*, 565. Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. II. Modelling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21. Ghose, A.; 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.
- (11) Viswanadhan, V.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. IV. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163-172.
- (12) Ghose, A. K.; Viswanadhan, V.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. B* **1998**, *102*, 3762.
- (13) Weininger, D. SMILES, a Chemical Language and Information System I. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
- (14) *Molecular Operating Environment*, Version 1998.10; Chemical Computing Group: Montreal, 1998.
- (15) Biobyte Corp., Pomona, CA, 1998.
- (16) *CRC Handbook of Chemical and Physics*, 72nd ed.; Lide, D. R., Ed.; CRC Press: Boston; 1991.
- (17) Martin, Y. C.; Duban, M.-E.; Bures, M. G. Poster presented at Computational Methods in Toxicology, April, 1998, Dayton, OH.
- (18) Mannhold, R.; Rekker, R. F.; Sonntag, C.; ter Laak, A. M.; Dross, K.; Polymeropoulos, E. E. Comparative Evaluation of the Predictive Power of Calculation Procedures for Molecular Lipophilicity. *J. Pharm. Sci.* **1995**, *84* (12), 1410.
- (19) Leo, A. J. Calculating log *P* from Structures. *Chem. Rev.* **1993**, *93*, 1281.

CI990307L