A New Method for the Estimation of Partition Coefficient

Nicholas Bodor,* Zoltan Gabanyi,[1] and Chu-Kuok Wong

Contribution from the University of Florida, Center for Drug Design and Delivery, J. Hillis Miller Health Center, Box J-497, Gainesville, Florida 32610. *Received October 7, 1988*

Abstract: A new highly nonlinear regressional model is presented for the estimation of 1-octanol/water partition coefficients. The molecular descriptors of the model are molecular surface, volume, weight, and charge densities on nitrogen and oxygen atoms of the molecule. All the descriptors are determined by using fully optimized structures based on AM1 calculations. The predictive power of the model is demonstrated by the accurate estimation of log *P* for complex molecules. The method is easy to use and it has general applicability.

During our recent efforts to develop computer programs[2] (expert systems) to predict novel "soft drugs"[3] and prodrugs[4] we have faced the problem of setting criteria to select the best structures among the candidates generated by the programs. While the various design rules combined with the expected properties (for example, predicted routes and rates of metabolism) provide a basis for selection, additional parameters were needed, such as the partition coefficient, which is a measure of the extent to which a solute is distributed between water and a water-immiscible liquid phase, as determined by their relative concentrations (weight per unit volume). The most frequently used octanol/water partition coefficient, *P* or its logarithm (log *P*), can have important use in predicting transmembrane transport properties, protein binding, receptor affinity, pharmacological activity, etc. of molecules. The log *P* value is generally easy to determine experimentally, but since in the design process we are dealing with predicted structures, the reliability of calculated values is important. Thus, we have examined the existing, mostly empirical methods.

In studying the effect of structural variations on log P, it was suggested that it has additive–constitutive character. Hansch[5] defined the π substituent constant in an analogous way to the definition of the well-known Hammett constant:

$$\pi_X = \log P_X - \log P_H \qquad (1)$$

where P_H is the partition coefficient for the parent compound, and P_X is the partition coefficient for an analogue in which H has been replaced with substituent X. Thus, specific substituents will have the same contribution in various molecules. This is a free energy related parameter which, however, remains constant only over slight structural modifications. It has been shown that the additivity is quite limited, e.g., it does not even hold for many benzene derivatives with two substituents.[6]

There are two widely used, essentially empirical ways for estimation of log *P*, both based on the assumed additivity: Rekker's *f* constant method,[7] and Leo and Hansch's fragment approach.[8] Rekker first defined an arbitrary set of terminal fragments using a database of about 1000 compounds with known log *P*. Linear regressional analysis was performed, where the number of different substructures was the independent variable and log *P* the dependent variables. The regressional coefficients obtained were designated group contributions. For example, the hydrophobic fragmental constant for hydrogen in Rekker's system is $f = 0.182$. There were some outliers, which were corrected by introducing a set of correction factors as integer multiples of a "magic factor" (0.289), describing some special structural features (proximity of polar groups, hydrogen atoms attached to polar groups, aryl–aryl conjugation, etc.). To estimate log *P* of a compound, one simply sums up the fragmental contributions and the applicable correction factors.

Leo and Hansch's philosophy was to determine log *P* values of a set of small molecules very accurately and calculate the fragmental values from these data. Using the concept of isolating carbon (sp^3 carbon atom with at least two bonds linked to other carbon atoms), they derived their own set of terminal fragments. This system also has a great number of correction factors (for different double bonds, multiple halogenation, polar proximity effects, etc.), which makes its application cumbersome. Many times it is difficult to decide how to divide (fragment) a molecule. A computerized version of this method is available (CLOGP program), which makes its use easier but no more reliable. Although essentially all log *P* values for the compounds included in the base set are well reproduced (due to specific corrections), there is no assurance the predicted values for complex drug molecules are reliable. This is more evident if one considers that the general fragment values cannot be used without further correction factors even for multiply substituted benzene derivatives.[9]

Molecular properties have previously been attempted to be predicted in terms of specific values representing molecular fragments. Thus, molecular heats of formation were computed as simple sums of terms derived for specific bonds from simple molecules.[10] It was soon discovered that strict additivity does not hold beyond simple substituted aliphatic and aromatic compounds: correction terms for strain energies, unsaturation, hybridization, interaction of heteroatoms, steric interaction, etc. had to be introduced.[11] It became evident that there is no assurance that heats of formation of complex molecules can be reliably calculated from these simple empirical additivity assumptions, and since the introduction of semiempirical SCF-MO calculations, molecular energies and heats of formation of large molecules are predicted on this basis. This approach, using the most advanced methods, like MNDO or AM-1 gives reliable molecular properties, including energies, conformations, ionization potentials, and dipole moments. It is noteworthy that a proposed ab initio method for large molecules using molecular fragments[12,13] was abandoned.

It is clear that prediction of *any* molecular property based on simple empirical or calculated fragment values has no scientific basis: fragments are generally quite differently behaving in different molecules. Thus, it is rather interesting that such a complex molecular property, the partition coefficient, which combines solute–solvent interactions in two different solvents is, even today, calculated ("predicted") on the basis of simple, empirically derived fragment values, ignoring a wide variety of specific

[1] On leave of absence from CompuDrug Ltd., Budapest, Hungary.

[2] Bodor, N.; Gabanyi, Z.; Wong, C.-K., unpublished work.

[3] Bodor, N. *CHEMTECH* **1984**, 28–38, and references cited.

[4] Bodor, N.; Kaminski, J. *Annu. Rep. Med. Chem.* **1987**, *22*, 303–313, and references cited.

[5] Fujita, T.; Isawa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.

[6] Franke, R.; Dove, S.; Kuhne, B. *Eur. J. Med. Chem.* **1977**, *14*, 363–374.

[7] Rekker, R. E. *The Hydrophobic Fragment Constant*; Elsevier: Amsterdam, 1976.

[8] Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1975**, *18*, 865–868.

[9] Leo, A. *J. Chem. Soc., Perkin Trans. 2* **1983**, 825–838.

[10] Cox, J. *Tetrahedron* **1962**, *18*, 1337–1350.

[11] Cox, J. *Tetrahedron* **1963**, *19*, 1175–1184.

[12] Christoffersen, R. E.; Genson, D. W.; Maggiora, G. M. *J. Chem. Phys.* **1971**, *54*, 239–252.

[13] Christoffersen, R. E. *J. Am. Chem. Soc.* **1971**, *93*, 4104–4111.

0002-7863/89/1511-3783$01.50/0

molecular properties (conformations, ionization, hydration, stereosimerism, ion-pair formation, keto–enol tautomerism, intra- and intermolecular H-bond formation, folding, etc.) affecting solubilities and, more so, partitioning.

Klopman and Iroff[14] used for the first time a different, molecular approach, based on quantum chemical calculations, to estimate log P. For a set of 61 simple organic compounds they determined the atomic charge densities using MINDO/3 and a Hückel-type method. A linear regression model was developed, which includes the total number of hydrogen, carbon, oxygen, and nitrogen atoms in the molecule and the sums of squared charges for carbon, nitrogen, and oxygen atoms. The latter parameters characterize the interaction of the solute and solvent molecules according to a simple electrostatic model. They also included some indicator variables to show if ester, acid, amide, or nitrile functionalities are present. Based on MINDO/3 calculations, the following model was found:

$$\log P = 0.344 + 0.2078n_H + 0.093n_C - 2.119n_N - 1.937n_O - 1.389q_C{}^2 - 17.28q_N{}^2 + 0.7316q_O{}^2 + 2.844n_A + 0.910n_T + 1.709n_M \quad (2)$$

where n_H, n_C, n_N, and n_O are the number of hydrogen, carbon, nitrogen, and oxygen atoms, $q_C{}^2$, $q_N{}^2$, and $q_O{}^2$ are the sum of squared charges on the carbon, nitrogen, and oxygen atoms or groups (group charge is the sum of charges of the heavy atom and bonded hydrogen atoms), and n_A, n_T, and n_M are indicator variables of acid/ester, nitrile, and amide functionalities. There are a number of advantages to using Klopman's method compared to the fragmental systems: the number of parameters is much less than in the fragmental methods, and it can be easily computerized and does not produce ambiguous results due to different fragmentation, or different interpretation of complex correction rules. If we take a close look at the method, however, some problems arise. The method for example is applicable only to compounds containing carbon, hydrogen, nitrogen, and oxygen atoms. Some special indicator variables for arbitrarily chosen fragments were introduced even though use of atomic contributions should suffice. This detracts from the closedness of the model, since it cannot be known with any certainty whether there is a need to involve further indicator variables. Also, the amide and nitrile fragments have low frequency in the data set, making the corresponding regression parameters unreliable. In addition, "standard" molecular geometry (representative bond length and angles) was used for MINDO/3 calculations; only the dihedral angles were optimized. It is also clear that calculated charge distribution alone is not enough to characterize the solubility of the compound. While Klopman's method is simple and unambiguous, a more comprehensive treatment including other important molecular properties is needed. Accordingly, we have also examined the contribution of such important parameters as molecular volume, weight, surface, and shape, in addition to the charge distribution and dipole, the latter a very important property in the present context, completely ignored before.

Methods

Starting geometries of the compounds were generated with CHEMCAD,[15] an interactive molecular building software program, which ran on IBM PC/AT. The created coordinate data files were converted to the format of AMPAC input files and were transferred to a MicroVAX II, where the quantum chemical computation was done. Fully optimized geometries were obtained by the AM1[16] method. Starting from the optimized molecular geometry, molecular surface and volume were determined. They are important factors of partition, because of the strong dependence between the surface of the solute molecule and the free energy of solvation.[17] Molecular volume was calculated by a numeric integration technique. A set of regular three-dimensional cubic grids is generated, the center of a grid is positioned at an atom, the edge of the cube is the diameter of the atom. Every grid point is tested as to whether it is within the atom, and special care is taken to avoid the problem of atomic overlap. A grid point is considered to be within the atom if the following conditions are satisfied:

$$g_{ijk}^{(L)} - c_L < r_L \text{ and } g_{ijk}^{(L)} - c_p > r_p \quad p = L \ldots L - 1 \qquad (3)$$

where $g_{ijk}^{(L)}$ is a point in grid L, c_L is the center of atom L, r_L is the van der Waals radius of atom L. The first condition in (3) is satisfied if the test point is within the current atom; the second condition assures that the test point does not belong to any previously considered atom. Volume contribution of an atom is estimated by the expression

$$V = {}^4\!/_3 r^3 \pi(n/n_t) \qquad (4)$$

where n is the number of grid points satisfying condition 3, n_t is the total number of grid points within the atom, and r is the van der Waals radius. Molecular volume is calculated by summing up these atomic contributions.

A similar algorithm was developed to calculate the surface area of molecules. First a set of spheric surface points is generated. Again a cubic grid is used as a starting point. The grid is centered at the origin of the coordinate system and has an edge, $a = 2$. A sphere is defined, centered at the origin, with radius $r = 1$. All grid points are selected that are near the surface of the sphere by the following criteria:

$$(x^2 + y^2 + z^2)^{1/2} - 1 < \text{eps} \qquad (5)$$

where x, y, and z are the nonnegative Cartesian coordinates of the grid point and eps is the precision threshold. If a grid point satisfies the inequality 5, then, in the general case, seven other acceptable grid points can be generated by projections. Now a point can easily be generated on the surface of an atom of the molecule:

$$p_{ik} = v_i + r_i g_k \qquad (6)$$

where p_{ik} is the kth point on the surface of atom i, v_i is the center of atom i, r_i is the radius of atom i, and g_k is the kth spheric grid point. A point p is on the surface of the molecule if it is not in the sphere of any other atom, i.e., it satisfies the following inequalities:

$$p_{ik} - v_j > r_j \quad \text{for } j = 1 \ldots n \text{ and } i \text{ is not equal to } j \qquad (7)$$

If we count the number of points satisfying the system of inequality 7, we can estimate the atomic contribution to the molecular surface:

$$S_i = 4\pi n/n_t r_i{}^2 \qquad (8)$$

where S_i is the atomic surface contribution, n_t number of all spheric trial points, and n is the number of points satisfying inequality 7.

The total molecular surface can be calculated by summing the above atomic surface contributions.

Results of the surface and volume calculations were compared to the results of Pearlman's SAVOL program;[18] the difference was less than 0.2%, which is acceptable. The precision can be improved by using a denser grid (we used 20 × 20 × 20 grid), but we do not feel that this is necessary. An additional geometric parameter was derived by using molecular surface and volume. From the volume, one can calculate the minimum surface of the molecule (ideally spheric). The ratio of the actual surface and this minimum surface gives a descriptor of ovality ($O > 1$):

$$O = S/4\pi\left(\frac{3V}{4\pi}\right)^{2/3} \qquad (9)$$

where S is the molecular surface and V is the molecular volume.

We have derived a great number of other parameters as well. By use of Klopman's method we tested all possible sum of squared charges for given elements, namely C, H, N, O, F, and Cl. We have generated similar parameters summing up the absolute values of charges. For all parameters above, derived parameters were generated: all squared and square-rooted descriptions. These were thought important because of the expected nonlinearity of the model. Molecular weight was also included. We generated all possible variables giving the occurrence of atoms of an element in the molecule, dipole moment, heat of formation, ionization potential, HOMO and LUMO orbital energies and their difference, number of hydrogen atoms bonded to nitrogen or oxygen atoms (possible parameters describing hydrogen bonding), sum of total squared charges, and some indicator variables describing whether the molecule is an alkane (C_nH_{2n+2}). Necessary variable transformations were carried out by using a special drug design oriented spreadsheet program, DRUGIDEA.[19] The

[14] Klopman, G.; Iroff, L. D. *J. Comput. Chem.* **1981**, *2*, 157–160.

[15] Kuhn, D. R. C. Graph. Corp., ChemCAD program, Austin, TX, 1985.

[16] Program AMPAC, Dewar Group, 1986.

[17] Pearlman, R. S. Molecular Surface Area and Volume. In *Partition Coefficient Determination and Estimation*; Dunn, W. J., Ed.; Pergamon Press: New York, 1986; pp 3–20.

[18] Pearlman, R. S., personal communication on SAVOL program, 1987.

[19] CompuDrug Ltd. Drugidea user's manual, Budapest, 1987.

Table I. Experimental and Estimated log *P* Values

		log *P*		CLOGP				log *P*		CLOGP	
no.	compd	expt[a]	est[b]	expt[c]	est[d]	no.	compd	expt[a]	est[b]	expt[c]	est[d]
1	propane	2.36	2.20	2.31	2.281	**60**	*N*,*N*-dimethylacetamide	−0.77	−0.26	−0.77	−0.802
2	isobutane	2.76	2.75	2.76	2.680	**61**	butyramide	−0.21	−0.25	−0.21	−0.176
3	pentane	3.31	3.28	3.39	3.339	**62**	furan	1.34	1.14	1.34	1.348
4	neopentane	3.11	3.30	3.11	3.079	**63**	pyrrole	0.75	0.36	0.75	0.758
5	cyclohexane	3.44	2.64	3.44	3.354	**64**	pyrrolidine	0.46	0.44	0.46	0.004
6	benzene	2.10	2.43	2.13	2.142	**65**	pyridine	0.67	0.89	0.65	0.665
7	toluene	2.74	2.73	2.73	2.791	**66**	chloroform	1.96	2.20	1.97	1.952
8	ethylbenzene	3.15	3.15	3.15	3.320	**67**	dichloromethane	1.25	1.50	1.25	1.249
9	propylbenzene	3.63	3.55	3.72	3.849	**68**	difluoromethane	0.20	0.53	0.20	0.369
10	methanol	−0.71	−0.69	−0.74	−0.764	**69**	methyl chloride	0.91	0.89	0.91	0.936
11	ethanol	−0.28	−0.22	−0.31	−0.235	**70**	methyl fluoride	0.51	0.39	0.51	0.496
12	propanol	0.28	0.29	0.25	0.294	**71**	nitromethane	−0.34	−0.74	−0.35	−0.284
13	butanol	0.88	0.84	0.88	0.823	**72**	ethylene oxide	−0.30	−0.26	−0.30	−0.792
14	isobutyl alcohol	0.75	0.87	0.76	0.693	**73**	ethyl chloride	1.43	1.20	1.43	1.465
15	*sec*-butyl alcohol	0.61	0.84	0.61	0.603	**74**	carbon tetrachloride	2.73	3.00	2.83	2.875
16	*tert*-butyl alcohol	0.36	0.83	0.35	0.473	**75**	crotonic acid	0.72	0.41	0.72	0.690
17	pentanol	1.48	1.34	1.56	1.352	**76**	adenine	−0.13	−0.36	n/a	−0.561
18	isopentyl alcohol	1.29	1.39	1.42	1.222	**77**	2-aminopyridine	0.52	0.55	0.49	0.345
19	neopentyl alcohol	1.34	1.38	3.11	3.079	**78**	2,4-dinitrophenol	1.52	1.54	1.54	1.915
20	*tert*-amyl alcohol	0.89	1.40	0.89	1.002	**79**	*m*-chlorophenol	2.50	1.78	2.50	2.485
21	cyclohexanol	1.23	1.67	0.81	0.805	**80**	nitrobenzene	1.84	1.60	1.85	1.885
22	1-hexanol	2.03	1.88	2.03	1.881	**81**	*m*-nitroaniline	1.37	1.35	2.45	2.534
23	1-octanol	3.15	2.80	2.97	2.030	**82**	phenol	1.49	1.36	1.46	1.475
24	dimethyl ether	0.10	−0.27	0.10	−0.188	**83**	hydroquinone	0.55	0.90	0.59	0.808
25	diethyl ether	0.83	0.80	0.89	0.870	**84**	aniline	0.90	1.12	0.90	0.915
26	dipropyl ether	2.03	1.86	2.03	1.928	**85**	*m*-aminophenol	0.18	0.29	0.17	0.248
27	butyl ethyl ether	2.03	1.89	2.03	1.928	**86**	*o*-aminophenol	0.62	0.18	0.62	0.648
28	methylamine	−0.57	−0.80	−0.57	−0.664	**87**	*p*-aminophenol	0.04	0.38	0.62	0.648
29	isopropylamine	−0.03	0.14	0.26	0.174	**88**	benzonitrile	1.56	2.04	1.56	1.575
30	butylamine	0.87	0.77	0.97	0.923	**89**	benzimidazole	1.37	0.92	1.46	1.547
31	*tert*-butylamine	0.40	0.70	0.40	0.573	**90**	benzaldehyde	1.45	1.46	1.48	1.495
32	cyclohexylamine	1.49	1.52	1.49	1.367	**91**	benzoic acid	1.95	1.57	1.87	1.885
33	diethylamine	0.53	0.74	0.58	0.540	**92**	2-acetylpyridine	0.84	0.73	0.85	0.438
34	piperidine	0.76	0.94	0.84	0.555	**93**	*p*-aminobenzoic acid	0.77	0.42	0.83	1.004
35	butylmethylamine	1.33	1.18	1.33	1.069	**94**	phenylurea	0.87	1.26	0.83	0.845
36	dipropylamine	1.62	1.77	1.67	1.598	**95**	anisole	2.08	1.81	2.11	2.061
37	dibutylamine	2.76	2.68	2.83	2.656	**96**	*o*-methoxyphenol	1.33	1.43	1.32	1.294
38	trimethylamine	0.22	0.08	0.16	0.048	**97**	*m*-toluidine	1.42	1.49	1.32	1.564
39	butyldimethylamine	1.70	1.70	n/a	0.946	**98**	2-(trifluoromethyl)benzimidazole	2.39	2.05	2.67	2.677
40	triethylamine	1.45	1.74	1.45	1.395	**99**	acetophenone	1.66	1.87	1.58	1.581
41	tripropylamine	2.79	3.10	2.79	2.822	**100**	phenylacetic acid	1.46	2.05	1.41	1.414
42	acetone	−0.24	−0.07	−0.24	−0.268	**101**	vanillin	1.26	1.33	1.21	1.354
43	2-butanone	0.35	0.45	0.29	0.261	**102**	phenoxyacetic acid	1.29	1.72	1.34	1.326
44	2-hexanone	1.78	1.55	1.38	1.319	**103**	*o*-vanillin	1.35	1.37	1.37	1.984
45	cyclohexanone	0.81	1.32	0.81	0.805	**104**	acetanilide	1.21	1.03	1.16	1.161
46	formic acid	−0.54	−0.67	−0.54	error	**105**	ethyl nicotinate	1.34	1.24	1.32	1.497
47	acetic acid	−0.24	−0.37	−0.17	−0.234	**106**	caffeine	−0.02	0.45	−0.07	0.260
48	propionic acid	0.29	0.19	0.33	0.295	**107**	quinoline	2.04	2.22	2.03	2.049
49	butyric acid	0.79	0.70	0.79	0.824	**108**	2-phenylimidazole	1.88	1.55	1.88	2.051
50	hexanoic acid	1.90	1.76	1.92	1.882	**109**	propionanilide	1.63	1.37	1.61	1.690
51	methyl acetate	0.18	0.05	0.18	0.142	**110**	naphthalene	3.35	3.66	3.30	3.316
52	ethyl acetate	0.70	0.59	0.73	0.671	**111**	2-ethylphenoxyacetic acid	2.59	2.48	2.42	2.749
53	propyl formate	0.83	0.61	0.83	0.794	**112**	4-phenylpyridine	2.55	2.69	2.59	2.553
54	ethyl propionate	1.21	1.15	1.21	1.200	**113**	biphenyl	4.06	4.26	4.09	4.030
55	isobutylene	2.37	1.61	2.34	2.136	**114**	diphenylamine	3.45	2.91	3.50	3.620
56	cyclohexene	2.86	2.51	2.86	2.810	**115**	diazepam	2.80	2.52	n/a	2.466
57	acetonitrile	−0.34	0.06	−0.34	−0.394	**116**	atropine	1.81	2.04	1.83	1.319
58	propionitrile	0.16	0.10	0.16	0.135	**117**	methadone	2.43	2.38	2.93	2.969
59	*N*-methylacetamide	−1.05	−0.61	−1.05	−1.078	**118**	tetracycline	−1.31	−1.29	n/a	−2.711

[a] From ref 21. There are very slight differences from the CLOGP experimental set in some cases. [b] Present method. [c] CLOGP file. [d] Calculated CLOGP.

regression analysis was performed by the stepwise linear regression program SLREGR, a member of software package LABSWARE.[20] A program LOGP, based on our model, was thus developed, which estimates the partition coefficient by using the results file of AMPAC.[16]

Results and Discussion

We have extended Klopman's original set of 61 compounds with an additional 57 compounds taken from a published collection.[21] The extended set of compounds comprises some basic heterocycles, halogenated compounds (F, Cl), multiple substituted benzene derivatives, and many well-known drug molecules. To select the latter compounds we examined the extensive collection of log *P* values[21] looking for compounds that have two or more 1-octanol/water log *P* values from different sources. A compound was accepted for the series if the standard deviation of the experimental log *P* values was less than 0.20. Sometimes the set of experimental log *P* values from different sources show high variability of log *P* determination due to lack of standardized conditions. Only values determined in comparable conditions were selected. By

[20] CompuDrug Ltd. Labsware user's manual, Budapest, 1986.

[21] Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

Table II, Basic Statistics of Variables in the Data Set

name	min	max	av	SD
log P	−1.31	4.06	1.27	1.12
S, Å^2	54.29	421.84	142.17	54.05
V, Å^3	32.88	373.81	106.80	48.79
O	1.09	1.68	1.30	0.10
I_{alkane}	0.00	1.00	0.03	0.18
M_w	31.05	444.44	105.89	56.23
D	0.00	6.28	1.92	1.28
Q_{ON}	0.00	3.07	0.43	0.46
Q_N	0.00	0.82	0.13	0.18
Q_O	0.00	0.86	0.13	0.18

using stepwise linear regression, a great number of different models were tested and the following regression model was found the best:

$$\log P = -1.167 \times 10^{-4}S^2 - 6.106 \times 10^{-2}S + 14.87O^2 - 43.67O + 0.9986I_{\text{alkane}} + 9.57 \times 10^{-3}M_w - 0.1300D - 4.929Q_{ON} - 12.17Q_N{}^4 + 26.81Q_N{}^2 - 7.416Q_N - 4.551Q_O{}^4 + 17.92Q_O{}^2 - 4.03Q_O + 27.273 \quad (10)$$

$$n = 118,\ F^2 = 115.1,\ r = 0.9388,\ \text{SE} = 0.296,\ \text{dof} = 104$$

where S is the molecular surface, O is the ovality of the molecule 9, I_{alkane} is the indicator variable for alkanes (its value is 1 if the molecule is an alkane, otherwise 0), M_w is the molecular weight, D is the calculated dipole moment, Q_{ON} is the sum of absolute values of atomic charges on nitrogen and oxygen atoms, while Q_N is the square root of sum of squared charges on nitrogen atoms; and Q_O is the square root of sum of squared charges on oxygen atoms.

Detailed results are given in Table I. All other variables were found statistically insignificant, so they were omitted from our model. It is interesting that the inclusion of variables for atomic occurrences do not improve the statistics of the model, although they had an important role in Klopman's model.

Overall molecular parameters (volume, surface, and weight) successfully replace them in the model. Higher powers of the included parameters (third, fourth) were also found insignificant, showing that our model adequately describes the nonlinearity of the data set. There was no need to include specific parameters for every element (C, H, F, and Cl); this can be explained by the special role of nitrogen and oxygen atoms in hydrogen bonding. All regression coefficients are statistically significant, their F-to-remove values vary between 105.7 and 3.7, and all are significant at probability level 0.05. We tried to fit many other models, including Klopman's original model (using AM-1 optimized values) as well, but all other models were inferior to model 10.

The new model contains 15 regression parameters, providing sufficient data to avoid the danger of finding chance correlation, as 105 degrees of freedom exist. It does not contain indicator variables for arbitrarily selected substructures. The need to include indicator variable I_{alkane} may arise from the different nature of partition of alkanes, for they cannot participate in any special interaction (hydrogen bonding, electrostatic effect) with the surrounding solvent molecules; the alkanes cannot therefore account for a quasi-structured hydrate environment. In fact, alkanes should be irrelevent for the drug molecules. The set of compounds is larger than the Klopman's set and is richer in different and complex compounds of pharmacological interest. It is important that the parameters vary at sufficiently large range in the data set (Table II), so that nonlinearities can be identified.

The geometrical parameters, surface and, indirectly, the volume are the most significant descriptors in our model. It means that the most important factor of the partition is the creation of a hole in the structure of water. The water molecules near the solute molecule are in a state of higher free energy, mainly on account of their lower entropy, than the other water molecules. In a classical physical approach one can think that putting a solute molecule in the solvent is similar to having a bubble in a liquid that has a surface tension energy linearly related to the surface of the bubble. Of course, the partition is a much more complex process, as the inclusion of ovality, O, and quadratic terms demonstrates. Molecular weight is also a volume-related parameter. These parameters, however, cannot fully describe the wide variation of log P. Dipole moment is an overall descriptor of the electronic interaction among the solvent and solute molecules. All the remaining parameters are derived from computed charge densities of nitrogen and oxygen atoms of the molecule, these elements being capable of forming hydrogen bonding with the solvent molecules. It is interesting that Q_N and Q_O have statistically equal regressional coefficients enabling their combination. The sign structure of the additional terms for nitrogen and oxygen are the same. It means the type of nonlinearity is the same for both elements. The difference in the values of corresponding regressional coefficients may come from the dissimilar extent of the hydrogen bonding of nitrogen and oxygen.

Table III, Predictive Power of the Model 10

		log P			CLOGP log P	
		expt	est[a]	est[b]	expt	est
1	testosterone	3.31	3.49	4.08	3.32	3.349
2	prednisone	1.46	1.71	0.12	1.46	0.582
3	progesterone	3.78	3.12	4.57	3.87	3.845
4	hydrocortisone	1.67	1.92	1.16	1.61	0.658
5	penicillin	1.83	1.68		1.83	1.683
6	phenytoin		2.52		2.47	2.085
7	prostaglandin	2.00	1.25		n/a	2.151
8	triamcinolone		1.71		1.16	−0.314
9	dexamethasone	1.99	1.79		1.83	1.406
10	betamethasone		1.73		1.83	1.406

[a] Estimation of log P using model 10. [b] Estimation of log P using Rekker's fragment analysis.

Table I also compares the calculated values by the empirical method, CLOGP. Since most of the compounds were included in the CLOGP basis set, the calculated values (including their respective corrections) fit well with the experimental ones. When a compound is not in the CLOGP basis set, the estimated CLOGP values deviate significantly from the experimental ones in three out of four cases (**39**, **76**, and **118**).

We tested the predictive power of our model 10 estimating the log P values of complex drug molecules, i.e., steroids and other compounds. Results of the estimation are given in Table III.

The calculated values using the proposed model, which does not need any fragment and correction values, are very good in all cases, obviously better than either fragmental method. The values estimated with CLOGP for prednisone, hydrocortisone, and triamcinolone are strikingly (1 log unit) different, despite having these molecules in the basis set. The Rekker method, used only for the first four molecules, shows even greater difference for prednisone, while the average error is significantly greater for both the Rekker method (0.93) and CLOGP (0.49) than the one calculated for the present method (0.31).

In conclusion, a new predictive model for partition coefficients was developed based on MO calculations performed on the whole molecule, which has better predictive value than previously known methods. The model includes seven geometrical and quantum chemical descriptors and is based on 118 organic compounds. The model is highly nonlinear. The predictive capability of the model is demonstrated on the example of different steroids and drug molecules. The method is easy to use and it has general applicability.