Assessment of *n*-Octanol/Water Partition Coefficient: When Is the Assessment Reliable?

Vijay K. Gombar* and Kurt Enslein

Health Designs, Inc., 183 East Main Street, Rochester, New York 14604

Received March 27, 1996[⊗]

A model, VLOGP, has been developed for assessment of *n*-octanol/water partition coefficient, log *P*, of chemicals from their structures. Unlike group contribution methods, VLOGP is based on linear free energy relationship (LFER) approach and employs information-rich electrotopological structure quantifiers derived solely from molecular topology. VLOGP, a robust and cross-validated model derived from accurately measured experimental log *P* values of 6675 diverse chemicals, has a coefficient of determination, R^2, of 0.986 and a standard error of estimate of 0.20. When applied to the training set, the largest deviation observed between experimental and calculated log *P* was 0.42. VLOGP is different from other log *P* predictors in that its application domain, called Optimum Prediction Space (OPS), has been quantitatively defined, i.e., structures to which the model should not be applied for predicting log *P* can be identified. A computer-assisted implementation of this model within HDi's toxicity assessment software package, TOPKAT 3.0, automatically checks whether the submitted structure is inside the OPS or not. VLOGP was applied to a set of 113 chemicals not included in the training set. It was observed that for the structures inside the OPS the average deviation between experimental and model-calculated log *P* values is 0.27, whereas the corresponding deviation for structures outside the OPS is 1.35. This demonstrates the necessity of identifying the structures to which a model is not applicable before accepting a model-based predicted log *P* value. For a set of 47 nucleosides, the performance of VLOGP was compared with that of four published log *P* predictors; a standard deviation of 0.33 was obtained with VLOGP, whereas the standard deviation from other log *P* predictors ranged between 0.46 and 1.20.

A. INTRODUCTION

Transmembrane transport of a molecule often determines its biological properties such as cellular uptake, bioavailability, receptor affinity, protein binding, pharmacological activity, toxicity, etc. The partition coefficient, *P*, of a molecule in the *n*-octanol/water solvent system has been recognized to emulate its transport across biological membranes.[1−3] It is evident from the large number of published[4] highly significant correlations between log *P* and a variety of biological properties, particularly nonspecific ones, that transport characteristics of a chemical can be effectively modeled in terms of its *n*-octanol/water partition coefficient.

In principle, experimental measurement of *P* of a chemical is straightforward.[5] However, in the business of computer-assisted drug design and combinatorial synthesis, researchers deal with chemical structures not yet synthesized. Therefore, it is of great relevance to devise methods which can predict reliably accurate values of P, or log *P*, solely from chemical structure.

A number of methods have been developed for prediction of log *P* from chemical structure.[6−19] These methods can be classified into two broad categories, namely group contribution methods, which capitalize on the additive-constitutive nature of log *P*, and regression methods, which quantify the weights of different structure descriptors on the principle of least square deviation. Given the fact that the contributions of various chemical groups to log *P* as well as the regression equations are derived from experimentally measured log *P* values of a limited set of chemicals, the log *P* predictors developed by either of these approaches are models of closed systems. Consequently, the applicability of these models is not universal. Therefore, it is essential to ascertain whether the model is or is not applicable to a chemical of interest. Unfortunately, previously published methods[6−19] do not provide diagnostic procedures to flag those chemical structures to which the models are not applicable. The widely used, commercially available,[20] group contribution method CLOGP warns when contribution(s) of certain groups are not known, but it cannot identify *a priori* when the total contribution due to all groups needs to be "corrected" to account for the interaction(s) among constituent groups. Similarly, the regression-based log *P* predictors cited in the literature, though by design do not suffer from such limitations, lack multivariate diagnostic procedures capable of signaling when not to apply the models.

The present regression model for log *P* prediction, referred to as VLOGP, is developed from 6675 accurately measured *n*-octanol/water partition coefficients. Rigorous diagnostic procedures have been applied to assure stability of the model. Further, the application domain, called Optimum Prediction Space (OPS), of VLOGP has been quantitatively defined, i.e., the structures to which the model should not be applied for predicting log *P* can be identified. Like other log *P* predictors, VLOGP will always produce an estimate of log *P*, but the result of the OPS test determines whether the estimated value of log *P* is reliable or not.

B. EXPERIMENTAL DATA

A compilation of over 10 000 log *P* values was acquired from BioByte Corp.[20] Inorganics, organometallics, salts, mixtures, and compounds whose structures could not be numerically expressed by our molecular descriptor generation

[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1996.

software were not considered for the present work. Since uniformity of the data in the training set is essential to exclude experimental noise, not all *n*-octanol/water entries were used. Those chemicals for which the suppliers did not provide a “Starlist” value were screened at the first step. A critical examination of the database further revealed that for 113 entries there were either multiple “Starlist” values, or multiple names, or multiple Chemical Abstracts Service (CAS) registry numbers. Instead of resolving the conflicts, these chemicals were set aside as a prediction set for testing the performance of the final model. There were 8702 chemicals with acceptable log *P* values which were retained.

It is well-known that regression outliers, both in the response as well as the explanatory variable(s), cause serious influence on the least mean square analysis.[21] In order to identify any leverage points in the log *P* data, a univariate analysis was performed. For the 8702 selected chemicals the log *P* values ranged between −4.41 to 11.29 with an average of 1.81. Sixteen extremely hydrophilic (log $P < -4.0$) or lipophilic (log $P > 8.0$) chemicals deep in the tail of the univariate distribution were discarded from the learning sample. A set of 8686 chemicals with log *P* ranging between −3.7 and 7.92 (average log $P = 1.80$ and median log $P = 1.76$) was finally retained for model development.

C. STRUCTURE QUANTIFICATION

Information-rich explanatory variables are key to parsimonious and meaningful regression models. In models for predicting log *P*, or for that matter any property, solely from molecular structure, an effective numerical representation of molecular structure is extremely important. It is evident from the existing predictors of additive constitutive molecular property log *P* that both fragmental components and their electronic interactions[22] in a molecule determine its log *P*. For instance, Leo[20] has compiled extensive tables of features, and their contributions to log *P*, representing interactions between predefined atom and group types. Similarly, based on 1663 chemicals, Klopman[14] has augmented 68 fundamental group contributions with 30 interaction terms. In order to objectively quantify these interactions, it was decided to employ molecular descriptors which would be sensitive to small changes in bulk and electronic attributes of molecular structure.

C.1. Electrotopological State Values. We have used electrotopological state values (*E* values) as numerical quantifiers of molecular structure.[23] The *E* value of an atom, or a group of atoms, encodes information about its electron content (valence, σ, π, and lone-pair), topology, and environment. Since an *E* value is computed by taking into account the effects of both intrinsic and environmental features, it changes even with remote variations in structures; of course, the magnitude of variation depends on the severity of change.

Calculation of *E* value does not require the knowledge of molecular geometry and is extremely fast. These topology-based descriptors have a practical advantage for quantitative structure-property relationship studies, because they can be applied to large sets of molecules without significantly depleting available computational resources. The methods for calculation of *E* values have been explained in the literature.[23]

C.2. Bulk Attributes. Besides molecular weight, we used size-corrected *E* values[23] for quantification of molecular bulk. The size-corrected *E* values are computed from a rescaled count of valence electrons.[23]

C.3. Shape Attributes. Since molecular shape and molecular symmetry also influence molecular transport, we included topological shape descriptors,[24,25] ${}^{m}k$ (kappa), of orders 1−7, and seven indices of molecular symmetry for effective quantification of molecular structure.

D. MODEL DEVELOPMENT

Since we had no knowledge of the usefulness of *E* values in modeling log *P*, we decided to first develop class-specific log *P* predictors for relatively small sets of chemicals. Two chemical classes were modeled, namely aliphatic hydrocarbons ($n = 26$) and aromatic hydrocarbons ($n = 52$). Finally, a log *P* predictor based on all 8686 chemicals was developed. For developing log *P* models of any class, the following steps were taken.

D.1. Evaluation of Predictor Variables. All the structural descriptors, namely, shape indices, symmetry indices, and counts, *E* values, and size-corrected *E* values for one-atom and two-atom fragments in the training set chemicals were subjected to a frequency of occurrence check. Any variables having nonzero values for less than three chemicals were not considered as predictor variables. This was done to enhance the statistical reliability of the predictor variables.

In order to reduce problems due to possible collinearity of variables, the pairwise correlations of these variables were examined. From a pair of variables with a correlation coefficient of 0.9 or higher only one variable was retained in the descriptor set. The variable which is easier to compute, comprehend, and is more continuous (more nonzero values) was generally retained.

D.2. Regression Analysis. The method of linear multiple regression (LMR) analysis was used to obtain a tentative log *P* predictor. The goal at this stage was to select the most potent variables. The BMDP[26] procedures 2R and 9R were employed for carrying out LMR analysis.

D.3. Regression Diagnostics. It is relatively easy and straightforward to obtain a tentative regression model by using standard software packages. However, before such a model is employed for predictive purposes, it is essential that the model be subjected to a variety of diagnostics to establish that

(1) all descriptors in the model are significant,

(2) no compounds with unique compound-variable association are in the training set,

(3) no influential, leverage, or outlier compounds remain in the training set,

(4) residuals are normally distributed, and

(5) cross-validation performance is not significantly different from the performance on the training set.

Unless these characteristics are established in a model, it is not robust, and its statistical quality may be questionable.

D.4. Defining Optimum Prediction Space (OPS). A robust and statistically significant model so diligently developed still represents a closed system, because it is based on a limited training set. Therefore, it should not be expected that the model will be applicable to every chemical. Of course, given the values of the descriptor variables the value of log *P* can be estimated, but the computed value may not be meaningful unless it is ascertained that the model is not being extrapolated beyond the Optimum Prediction Space

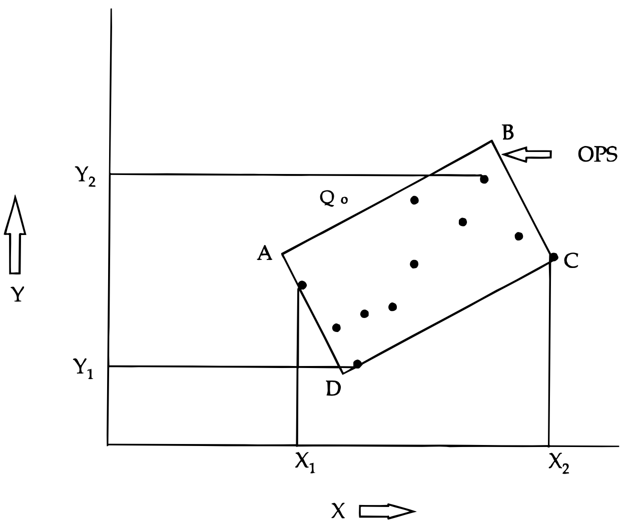

Figure 1. OPS and descriptor space for a mock two-variable model.

(OPS) associated with it. OPS is a multivariate space such that at points (chemicals) within and near the periphery of this space, the model is applicable. It is important to note that a query chemical being inside or near the periphery of the OPS does not mean that the predicted value of log *P* for that chemical will have concordance with the experimental value. All it implies is that the model is applicable to this chemical, and the probability of concordance between the predicted and the actual values is as high as that for the training set chemicals.

The OPS of a model with *p* descriptor variables is a $p + 1$ dimensional space derived from the descriptor space, i.e., the values of the *p* independent variables of *n* observations in the training set of the model. Each of the $p + 1$ dimensions of the OPS has upper and lower bounds quantitatively defined in terms of $p + 1$ elements composed of double-transformed values of the descriptor variables. A fuller description of the method used in calculating OPS is reserved for a later publication.

It may be mentioned here that though each of the *n* observations in the training set is inside OPS, the OPS is generally smaller than the descriptor space, and, therefore, the model may not be applicable to some regions in the descriptor space itself. As a simple example, consider a model with two descriptors X and Y. The descriptor space is defined by the limits X_2 and X_1 and Y_2 and Y_1 (see Figure 1). However, the OPS, shown by ABCD, is the space jointly covered by the descriptor values of *n* training set points. It can be seen that a query point Q, though inside the descriptor space, is outside the OPS.

E. RESULTS AND DISCUSSION

E.1. Aliphatic Hydrocarbons. The names of 26 aliphatic hydrocarbons selected for testing the potential of *E* values in modeling log *P* are collected in Table 1. A four-variable model, eq 1, was obtained:

$$\log P = 0.0473 + \text{mol wt} * 0.0403\ (26.75) + X1 * 0.1267\ (6.06) - X2 * 0.1035\ (12.53) - X3 * 0.0422\ (8.64) \quad (1)$$

$$n = 26 \quad R^2 = 0.980 \quad \text{EV} = 97.6\% \quad s = 0.151 \quad F_{4,21} = 255.9$$

Table 1. Comparison of Experimental and Estimated Log *P* Values of Some Aliphatic Hydrocarbons

	chemical	log *P* value		
		eq 1	Δ_1	expt
1.	ethylene	1.18	−0.05	1.13
2.	ethane	1.77	0.04	1.81
3.	propyne	1.04	−0.10	0.94
4.	propylene	1.83	−0.06	1.77
5.	*c*-propane	1.74	−0.02	1.72
6.	propane	2.25	0.09	2.36
7.	1,3-butadiene	1.67	0.32	1.99
8.	2-butyne	1.31	0.15	1.46
9.	isobutylene	2.70	−0.36	2.34
10.	2-butene	2.31	0.00	2.31
11.	1-butene	2.36	0.04	2.40
12.	isobutane	2.77	−0.01	2.76
13.	*n*-butane	2.83	0.06	2.89
14.	1-pentyne	2.07	−0.09	1.98
15.	*c*-pentane	2.87	0.13	3.00
16.	neopentane	3.29	−0.18	3.11
17.	*n*-pentane	3.41	0.21	3.62
18.	1,4-*c*-hexadiene	2.34	−0.04	2.30
19.	1,3-*c*-hexadiene	2.45	0.02	2.47
20.	*c*-hexene	2.86	0.00	2.86
21.	1,5-hexadiene	2.65	−0.20	2.45
22.	*c*-hexane	3.44	0.00	3.44
23.	*n*-hexane	3.98	0.13	4.11
24.	2,3-dimethylbutane	3.91	−0.06	3.85
25.	2,2-dimethylbutane	3.89	−0.07	3.82
26.	*n*-octane	5.11	0.07	5.18
	mean	2.62	0.0008	2.62
	std dev	0.965	0.136	0.975
	max.	5.11	0.32	5.18
	min.	1.04	−0.36	0.94

Δ_1: log *P*(expt) − log *P* (eq 1)

where structure descriptors X1 to X3 are *E* values of C atoms involved in the following substructures:

$$X1 = \text{C-CH}_3$$

$$X2 = \text{C-C}\equiv$$

$$X3 = \text{C-CH}=$$

Large absolute *t*-values (given in parentheses) of all regression coefficients indicate that all descriptors of this model are significantly correlated with log *P*. It may be noted that for this equation the values of all structure descriptors are algorithmically calculated from a two-dimensional molecular graph. As can be seen, the model returns highly significant statistics, namely $R^2 = 0.980$, explained variance = 97.6%, standard error of estimate = 0.151, and an *F* ratio of 255.9 for degrees of freedom 4 and 21.

The log *P* values calculated from this equation have been compared (Table 1) with the experimental values. Clearly, eq 1 results in an excellent agreement with the experimental log *P* values. It may also be noted that eq 1 can distinguish between isomeric alkenes. For example, different log *P* values are computed by eq 1 for 1,4- and 1,3-cyclohexadiene.

E.2. Aromatic Hydrocarbons. Since the qualities of the log *P* model for a set of aliphatic hydrocarbons did indicate the utility of *E* values as information-rich structure descriptors, we wanted to investigate the potential of *E* values for a set of aromatic molecules before embarking on developing a general log *P* predictor. For this purpose a set of aromatic hydrocarbons ($n = 52$) was selected. The names of 52 aromatic hydrocarbons are collected in Table 2. An eight-

Table 2. Comparison of Experimental and Estimated Log *P* Values of Some Aromatic Hydrocarbons

	chemical	log *P* value		
		eq 2	Δ_1	expt
1.	benzene	2.14	−0.01	2.13
2.	toluene	2.65	0.04	2.69
3.	ethynylbenzene	2.59	−0.06	2.53
4.	styrene	3.15	0.01	3.16
5.	ethylbenzene	3.10	0.05	3.15
6.	*p*-xylene	3.13	0.05	3.18
7.	*m*-xylene	3.13	0.07	3.20
8.	*o*-xylene	3.13	0.00	3.13
9.	indene	3.04	−0.12	2.92
10.	indane	3.26	0.07	3.33
11.	1-phenyl-1-propene	3.25	0.10	3.35
12.	*c*-propylbenzene	3.35	−0.08	3.27
13.	allylbenzene	3.21	0.02	3.23
14.	1,3,5-trimethylbenzene	3.60	−0.18	3.42
15.	1-ethyl-2-methylbenzene	3.56	−0.03	3.53
16.	1,2,3-trimethylbenzene	3.60	−0.05	3.55
17.	*n*-propylbenzene	3.79	−0.11	3.68
18.	isopropylbenzene	3.68	−0.02	3.66
19.	1,2,4-trimethylbenzene	3.60	0.05	3.65
20.	naphthalene	3.41	−0.06	3.35
21.	1,2,4,5-tetramethylbenzene	4.06	−0.06	4.00
22.	*tert*-butylbenzene	4.26	−0.15	4.11
23.	*p*-cymene	4.14	−0.04	4.10
24.	*n*-butylbenzene	4.17	0.09	4.26
25.	2-methylnaphthalene	3.87	−0.01	3.86
26.	1-methylnaphthalene	3.87	0.00	3.87
27.	acenaphthylene	3.96	0.07	4.03
28.	acenaphthene	3.98	−0.06	3.92
29.	biphenyl	3.99	0.10	4.09
30.	2,6-dimethylnaphthalene	4.32	−0.01	4.31
31.	2,3-dimethylnaphthalene	4.32	0.08	4.40
32.	1,8-dimethylnaphthalene	4.32	−0.06	4.26
33.	1,7-dimethylnaphthalene	4.32	0.12	4.44
34.	1,5-dimethylnaphthalene	4.32	0.06	4.38
35.	1,4-dimethylnaphthalene	4.32	0.05	4.37
36.	1,3-dimethylnaphthalene	4.32	0.10	4.42
37.	1,2-dimethylnaphthalene	4.32	−0.01	4.31
38.	2-ethylnaphthalene	4.27	0.11	4.38
39.	1-ethylnaphthalene	4.27	0.12	4.39
40.	hexamethylbenzene	c		4.31
41.	fluorene	4.18	0.00	4.18
42.	diphenylmethane	4.17	−0.03	4.14
43.	1,4,5-trimethylnaphthalene	4.77	0.13	4.90
44.	2,3,6-trimethylnaphthalene	4.76	−0.03	4.73
45.	phenanthrene	4.59	−0.02	4.57
46.	anthracene	4.59	−0.05	4.54
47.	1-methylfluorene	c		4.97
48.	stilbene	4.80	0.01	4.81
49.	9,10-dihydroanthracene	4.33	−0.08	4.25
50.	diphenylethane	4.78	0.02	4.80
51.	pyrene	5.17	0.01	5.18
52.	benz[*a*]anthracene	5.75	−0.14	5.61
	mean	3.91	0.0012	3.94
	std dev	0.697	0.075	0.704
	max.	5.75	0.13	5.61
	min.	2.14	−0.18	2.13

c: model not applicable to this compound
Δ_1: log *P*(expt) − log *P* (eq 2)

variable robust log *P* predictor, eq 2, was obtained from this set of aromatic hydrocarbons.

$$\log P = 5.2058 + X1 * 0.1517\ (53.75) + X2 * 0.0403\ (7.12) - X3 * 0.0366\ (7.42) + X4 * 0.0883\ (11.50) + X5 * 0.1337\ (7.13) - X6 * 2.4445\ (13.10) - X7 * 0.1423\ (6.10) + X8 * 0.1213\ (5.68) \quad (2)$$

$$n = 50 \quad R^2 = 0.989 \quad \mathrm{EV} = 98.6\% \quad s = 0.082 \quad F_{8,21} = 446.8$$

where structure descriptors X1 to X8 are *E* values on C atoms involved in the following substructures:

X1 = whole molecule

X2 = aliphatic C atoms

X3 = total of aliphatic C bound to aromatic C

X4 = total of -CH_3 group

X5 = ═CH- bound to aromatic C

X6 = average per atom

X7 = -CH═CH-

X8 = aliphatic C bound to aliphatic C (nonring)

Again, one can see that eq 2 yields excellent statistics. It was interesting to observe that X1, the sum of *E* values of all C atoms, was highly correlated with log *P* ($R^2 = 0.814$). X1 is not strictly additive like molecular weight but is a more sensitive and information-rich measure of molecular bulk. The values of log *P* calculated from eq 2 are compared with the experimental log *P* values in Table 2. As in the case of aliphatic hydrocarbons, the agreement is excellent; the largest absolute deviation between experimental and calculated log *P* values is 0.18. It may be mentioned that during the development of eq 2, the regression diagnostic procedures identified hexamethylbenzene and 1-methylfluorene as being influential and an outlier, respectively. These chemicals have been reported[14] to give largest deviation between experimental and calculated log *P*. According to our validation algorithms, however, these two compounds are identified to be outside the OPS associated with eq 2. Therefore, our algorithms warn about the acceptance of the computed log *P* values of these compounds. In order to have confidence in the reliability of a predicted log *P* value, it is very important to proactively identify the chemical structures to which a given model is not applicable.

E.3. General Log *P* Predictor. Encouraged by the performance of the two hydrocarbon log *P* predictors, an attempt was made to develop a general log *P* predictor based on the accurately measured log *P* values of 8686 chemicals with diverse structures. The advantage of such a large number of data points is that one can afford to set aside, during the regression diagnostic phase, many more chemicals without significantly reducing the prediction space of the model. The threshold for the Studentized Residual was set to 2.0 in order to obtain a tighter fitting model, of course, with somewhat reduced prediction space. Since this log *P* model will be used along with the OPS checking algorithms, it was opted to sacrifice prediction space in favor of better model performance. During the regression diagnostic process, a total of 2011 chemicals were set aside leaving 6675 chemicals in the training set of the final model. Hereafter, this model is referred to as VLOGP. Due to the volume of the data and the size of the 363-variable equation neither is included in the text here. However, the salient statistics of VLOGP are collected in Table 3. The values of log *P* for the training set span a wide range between −3.56 and 7.73.

Table 3. Salient Statistical Parameters of the VLOGP Predictor

parameter	value
training set population	6675
max. log P	7.73
min. log P	−3.56
av log P	1.851
std dev of data (SD)	1.628
coeff of determination (R^2)	0.986
explained variance	98.5%
std error of estimate (SE)	0.201
SD/SE	8.10
largest dev	0.417
av squared residual (model)	0.040
av squared residual (jackknife)	0.043
deg of freedom	363, 6311
F-ratio	1188

Table 4. Distribution of Absolute Deviations between Experimental and Calculated Log P Values

	percent chemicals	
deviation range	group	cumulative
0.000−0.050	19.1	19.1
0.051−0.100	16.9	36.0
0.101−0.150	15.0	51.0
0.151−0.200	13.3	64.3
0.201−0.250	11.7	75.9
0.251−0.300	9.7	85.6
0.301−0.350	8.0	93.6
0.351−0.400	5.4	99.0
0.401−0.420	1.0	100.0

Table 5. Performance Comparison of Five Log P Predictors

	log P predictor				
parameter[a]	KLOGP[14]	ALOGP[16]	CLOGP[17]	BLOGP[11]	VLOGP[b]
		$N = 47$ (Nucleosides and Bases)			
Δ^2	9.71	12.44	39.58	66.40	4.94
corr coeff	0.885	0.838	0.711	0.398	0.901
std dev	0.46	0.52	0.93	1.20	0.33
pred R^2	0.60	0.48	−0.63	−1.74	0.79
		$N = 35$ (Nucleosides)			
Δ^2	7.37	10.54	35.25	63.32	3.20
corr coeff	0.913	0.862	0.753	0.432	0.923
std dev	0.47	0.56	1.02	1.36	0.31
pred R^2	0.63	0.47	−0.79	−2.21	0.84
		$N = 12$ (Bases)			
Δ^2	2.34	1.90	4.33	3.08	1.74
corr coeff	0.769	0.797	0.597	0.810	0.793
std dev	0.46	0.42	0.63	0.53	0.39
pred R^2	0.47	0.57	0.01	0.30	0.60

[a] $\Delta^2 = \Sigma(\log P_{calc} - \log P_{ref})^2$, corr coeff = correlation between calculated and observed values, std dev = standard deviation of ($\log P_{calc} - \log P_{ref}$), pred $R^2 = (SD - \Delta^2)/SD$, where SD is the sum of squared deviations of each measured log P value from their mean. [b] This work.

The high degree of fit of the model is indicated by a small standard error of only 0.201, and the stability of the model is illustrated by a small difference between the model average squared residual (= 0.040) and the jackknife average squared residual (= 0.043). The 363-variable model has an explained variance of 98.5% and is significant at $p < 0.0001$ with an F ratio of 1188. Once again, there are no “correction factors” involved in the VLOGP model.

The values of log P calculated by the VLOGP model were compared with the experimental values of the 6675 training set compounds. A distribution of the deviations in various ranges is shown in Table 4. It can be seen that for over half the chemicals the deviations are smaller than 0.15 and for over 75% chemicals below 0.25.

E.4. Cross-Validation. A true test of a model’s performance is its ability to accurately predict log P of chemicals not included in the training set. Two experiments were conducted to cross-validate the VLOGP model: the first with a benchmark set of 47 nucleosides and nucleoside bases, and the second with the prediction set of 113 chemicals referred to in section B above. For a convenient and fast application of the VLOGP model, it was installed in our TOPKAT 3.0 package for making these tests.

E.4.1. Nucleoside Data Set. Viswanadhan *et al.*[27] have published a list of 35 nucleosides and analogs and 12 nucleoside bases and derivatives which they used to evaluate several methods for the prediction of log P. The log P values of this set of nucleosides were computed with VLOGP. The results were compared with the experimental values and with the log P values predicted by four other methods, namely ALOGP, BLOGP, CLOGP, and KLOGP. Due to inaccessibility to other published and commercially available log P predictors and due to the limited scope of the present study, the comparison was limited to these four models. The results are summarized in Table 5. The following observations are evident from the comparative study: (a) Δ^2, the sum of squared deviations, for VLOGP is the lowest of all five log P predictors, thus indicating best agreement with the reference values, (b) the correlation coefficient between the reference and predicted log P values is the highest for the VLOGP model, (c) VLOGP produces the smallest standard deviation, and (d) the prediction R^2 value from VLOGP is positive and the highest; negative values for prediction R^2 indicate that the average log P will produce a fit better than that obtained by the respective model. It can be seen from Table 5 that these trends of better performance of VLOGP hold when comparison of different models is carried over individual subsets of 35 nucleosides and 12 nucleoside bases. It should be noted that this performance of VLOGP has been achieved without using three-dimensional structural descriptors or any additional interaction terms.

It may be conceded that this set of 47 chemicals is rather small. However, this set has appeared in the log P literature as a *de facto* benchmark for testing the performance of log P predictors. Further, since a model developed from a given training set is a representation of that set, its application is not universal. Therefore, one could certainly find chemicals where VLOGP performs worse. But with its OPS quantitatively defined, the structures for which log P should not be estimated with VLOGP can be identified.

E.4.2. Miscellaneous Data Set. The log P values of 113 diverse compounds in the prediction set were calculated using the VLOGP model. Twenty-nine chemicals were identified to be outside the OPS of VLOGP. The results are shown in Table 6. Since the reference log P values of these chemicals came from a proprietary compilation, only empirical formulas of the chemicals are given. For several of the chemicals up to four log P values were available in that compilation. They are shown as reference values v1 to v4. All reported values were compared with the log P values predicted by VLOGP. The deviations from the best and the worst agreement are collected under the columns titled “best fit” and “worst fit”, respectively.

Table 6. Performance of VLOGP on a Miscellaneous Prediction Set

			log *P*						
			ref						
no.	empirical formula	inside OPS?	v1	v2	v3	v4	pred.	best fit dev	worst fit dev
1.	$C_{11}H_{13}N$	N	2.08	2.13			3.020	0.890	0.940
2.	$C_9H_{11}NO_2$		1.56	1.69			1.610	0.050	0.080
3.	$C_{10}H_{10}O_3$		1.81	2.54			2.181	0.359	0.371
4.	$C_6H_{12}Br_2O_4$	N	−0.24	−0.29			−1.396	1.106	1.156
5.	$C_6H_{10}Cl_2$		3.18	3.21			2.610	0.570	0.600
6.	$C_{16}H_{25}NO_2$		2.51	2.63			2.190	0.320	0.440
7.	$C_4H_4O_4$		−0.34	0.46			0.515	0.055	0.855
8.	$C_9H_{13}N_1$		2.23	2.49			2.751	0.261	0.521
9.	$C_{10}H_{13}N_5O_3S_1$		−0.56	−0.79			−0.598	0.038	0.192
10.	$C_{14}H_{19}N_1$	N	3.13	3.3			2.390	0.740	0.910
11.	$C_{17}H_{25}NO_2$	N	2.93	3.13			3.159	0.029	0.229
12.	$C_8H_8Cl_5F_3O_1$		4.06	4.15			3.392	0.668	0.758
13.	$C_7H_9Cl_5O_1$		3.14	3.51			3.432	0.078	0.292
14.	$C_6H_7Cl_5$	N	3.37	3.53			3.727	0.197	0.357
15.	$C_{15}H_{14}O_6$		0.15	0.36			0.988	0.628	0.838
16.	$C_8H_{16}O_1$		2.10	2.37	2.38		2.640	0.260	0.540
17.	$C_{12}H_{12}F_3N_1$	N	2.91	3.21			2.038	0.872	1.172
18.	$C_{12}H_{12}F_3N_1$	N	2.85	3.19			2.053	0.797	1.137
19.	$C_{10}H_{13}N_3O_4$		−2.36	−2.45			−2.895	0.445	0.535
20.	$C_6H_8N_2O_2S_1$		−0.38				0.000	0.380	0.380
21.	$C_{14}H_{18}N_4O_2$		2.06	2.14			1.418	0.642	0.722
22.	$C_{24}H_{31}FO_6$		2.77	2.91			2.063	0.707	0.847
23.	$C_{10}H_{16}ClN_3O_4$		1.53	1.68			1.347	0.183	0.333
24.	$C_{10}H_{13}NO_1$		0.81	0.86			1.042	0.182	0.232
25.	$C_{27}H_{38}O_{10}$		0.11	0.27			0.445	0.175	0.335
26.	$C_{10}H_{11}NO_2$		1.06	1.10			1.306	0.206	0.246
27.	$C_6H_6Br_2Cl_4$		3.88	3.99			3.888	0.008	0.102
28.	$C_{10}H_{20}O_1$		3.02	3.09			3.029	0.009	0.061
29.	$C_6H_3Cl_3O_1$		4.01	4.28			3.774	0.236	0.506
30.	$C_6H_8O_4$		0.22	0.74			0.812	0.072	0.592
31.	$C_{17}H_{25}NO_2$	N	2.97	3.13			3.770	0.640	0.800
32.	$C_{10}H_{15}NO_1$		0.93	1.43			0.839	0.091	0.591
33.	$C_6H_3Cl_4N_1$		4.04	4.57			4.038	0.002	0.532
34.	$C_{21}H_{30}O_3$		2.04	2.37			2.797	0.427	0.757
35.	$C_9H_{18}O_1$		3.00	3.53			2.913	0.087	0.617
36.	$C_{27}H_{40}O_{10}$		0.15	0.44			0.099	0.051	0.341
37.	$C_{18}H_{25}NO_2$	N	2.96	2.97			3.549	0.579	0.589
38.	$C_9H_{16}ClN_3O_3$		1.34	1.75			1.403	0.063	0.347
39.	$C_{18}H_{20}FN_3O_4$		−0.28	−0.39			1.062	1.342	1.452
40.	$C_{13}H_{17}N_1$	N	2.62	2.72			1.918	0.702	0.802
41.	$C_{10}H_{18}O_1$		2.32	2.72			1.790	0.530	0.930
42.	$C_8H_7NO_5$		0.97				1.252	0.282	0.282
43.	$C_{14}H_{14}O_2$		1.56	1.91			1.900	0.010	0.340
44.	$C_{15}H_{10}Cl_2N_2O_2$		2.39	2.51			2.406	0.016	0.104
45.	$C_2H_2Cl_2$		1.86	2.09			1.253	0.607	0.837
46.	$C_6H_5Cl_5$	N	3.60	3.61	3.8	3.85	0.160	3.440	3.690
47.	$C_{21}H_{26}N_2O_3$	N	2.54	2.73	2.94		0.982	1.558	1.558
48.	$C_{10}H_9N_3O_2$		2.59	2.59			2.651	0.061	0.061
49.	$C_7H_9Cl_5S_1$		3.75	3.85			4.390	0.540	0.640
50.	$C_6H_{14}O_6$	N	−3.10				−4.215	1.115	1.115
51.	$C_{11}H_{13}NO_1$		0.59	0.91			0.470	0.120	0.440
52.	$C_{10}H_{12}FN_5O_2$		−0.12	−0.18			−0.333	0.153	0.213
53.	$C_8H_{11}Cl_3O_6$		1.02	1.12			1.105	0.015	0.085
54.	$C_{10}H_{13}N_5O_4$		−1.10	−1.11			−1.431	0.321	0.331
55.	$C_{12}H_{12}N_2O_3$		1.41	1.59			0.966	0.444	0.624
56.	$C_{24}H_{32}O_5$		1.94	2.18			2.423	0.243	0.483
57.	$C_9H_{16}ClN_3O_3$		1.00	1.11			0.908	0.092	0.202
58.	$C_{25}H_{40}O_3$		5.97	6.13			5.994	0.024	0.136
59.	$C_{11}H_{18}N_2O_3$		2.07	2.10			2.110	0.010	0.040
60.	C_6H_{10}		2.80	3.01			3.082	0.072	0.282
61.	$C_9H_{12}ClN_3O_4$		−0.71	−1.05			−1.015	0.035	0.305
62.	$C_{27}H_{31}N_2O_2$	N	−0.34	−0.55			6.429	6.769	6.979
63.	$C_{20}H_{24}N_2O_2$		2.64	2.88			3.600	0.720	0.960
64.	$C_{11}H_{18}ClN_3O_4$		1.93	1.98			2.029	0.049	0.099
65.	$C_{12}H_{15}N_1$	N	2.29	2.32			0.991	1.299	1.329
66.	$C_{13}H_{16}ClNO_1$		2.40	3.52			2.924	0.524	0.596
67.	$C_8H_{11}Cl_5O_1$		3.69	3.97			3.800	0.110	0.170
68.	$C_{12}H_{15}N_1$	N	2.32	2.41			0.473	1.847	1.937
69.	$C_9H_{13}N_3O_5$		−2.13	−2.51			−2.116	0.014	0.394
70.	$C_{11}H_{17}NO_1$		1.87	2.05			1.422	0.448	0.628
71.	$C_9H_{13}NO_1$		0.67	0.83			−0.295	0.965	1.125
72.	$C_6H_{12}O_1$		1.22	1.34			1.784	0.444	0.564

Table 6 (Continued)

			log *P*						
			ref						
no.	empirical formula	inside OPS?	v1	v2	v3	v4	pred.	best fit dev	worst fit dev
73.	$C_6H_6Cl_6$		3.72	3.78	3.80	4.14	3.657	0.063	0.483
74.	$C_{16}H_{23}NO_2$	N	2.51	2.52			3.172	0.652	0.662
75.	$C_9H_{12}FN_3O_3$		−1.14	−1.18			−0.967	0.173	0.213
76.	$C_{12}H_7F_6N_3O_2$		5.02	5.06			5.428	0.368	0.408
77.	$C_{24}H_{40}O_3$		6.11	6.13			6.509	0.379	0.399
78.	$C_7H_7NO_1$		1.75	1.85			1.913	0.063	0.163
79.	$C_{10}H_{11}NO_1$	N	0.40	0.75			−0.185	0.585	0.935
80.	$C_{12}H_{15}N_1$	N	2.37	2.47			1.647	0.723	0.823
81.	$C_6H_4Cl_6$	N	4.12	4.31	4.34		0.470	3.650	3.650
82.	$C_{11}H_{15}NO_2$		2.57	2.80			2.714	0.086	0.144
83.	$C_{24}H_{32}O_4S_1$		2.26	2.78			2.827	0.047	0.567
84.	$C_6H_6Cl_4$		3.52	3.65	3.72		4.136	0.416	0.616
85.	$C_{12}H_{16}O_6$		−0.75	−0.89			−0.524	0.226	0.366
86.	$C_6H_{12}O_2$		0.08	0.23			−0.304	0.384	0.534
87.	$C_{13}H_{11}Cl_3N_2O_2$	N	3.90	3.90			3.182	0.718	0.718
88.	$C_8H_{12}Cl_4O_2$		2.55	2.82			3.164	0.344	0.614
89.	$C_{22}H_{29}FO_5$		1.83	1.94			1.657	0.173	0.283
90.	$C_{22}H_{19}Cl_2NO_3$		6.05	6.05			5.817	0.233	0.233
91.	$C_{10}H_{15}N_3O_4$		−2.41	−2.45			−1.861	0.549	0.589
92.	C_5H_8		2.40	2.44			2.711	0.271	0.311
93.	$C_{16}H_{23}NO_2$	N	2.43	2.83			2.609	0.179	0.221
94.	$C_{13}H_{14}F_3N_1$	N	3.41	3.65			4.922	1.272	1.512
95.	$C_{11}H_{13}N_1$	N	2.00	2.09			0.493	1.507	1.597
96.	$C_{19}H_{22}N_2O_1$		2.68	2.82			3.528	0.708	0.848
97.	$C_{16}H_{18}O_8$		−0.66	−0.78			−0.439	0.221	0.341
98.	$C_{12}H_{15}N$	N	1.63	2.04			4.511	2.471	2.881
99.	$C_{13}H_{16}BrNO_1$		2.66	3.56			3.724	0.164	1.064
100.	$C_7H_{14}O_1$		1.82	1.84			2.008	0.168	0.188
101.	$C_5H_{10}N_4O_2$		−0.25	−0.28			−0.644	0.364	0.394
102.	$C_{13}H_{14}F_3N_1$	N	3.54	3.64			5.199	1.559	1.659
103.	$C_{13}H_{14}F_3N_1$	N	3.45	3.59			5.198	1.608	1.748
104.	$C_6H_9NO_6$		−0.15	−0.40			−0.219	0.069	0.181
105.	$C_6H_6N_2O_1$		−0.34	−0.37			−0.363	0.007	0.023
106.	$C_{23}H_{30}O_3$	N	2.54	3.08			3.596	0.516	1.056
107.	$C_6H_6Cl_5I_1$		3.96	4.05			4.003	0.043	0.047
108.	$C_9H_{11}N_1$	N	1.49	1.58			2.292	0.712	0.802
109.	C_4H_8		2.31	2.33			1.957	0.353	0.373
110.	$C_6H_6BrCl_5$		3.74	3.81			3.955	0.145	0.215
111.	$C_{18}H_{18}ClNS_1$		5.18	5.18			5.481	0.301	0.301
112.	$C_6H_6Cl_4$		3.08	3.15	3.40	3.74	2.648	0.432	1.092
113.	$C_{19}H_{23}N_5O_7S_2$		2.42				1.804	0.616	0.616

As mentioned above, out of a total of 113 compounds 29 were identified to be outside the OPS associated with VLOGP. Though the application ratio of 74.3% seems low, the capability of the OPS algorithms to proactively warn about the chemicals to which the model is not applicable is in fact a boon. It was found that for the compounds inside the OPS the average deviation for the "best fit" and the "worst fit" was only 0.272 and 0.446, respectively. However, the corresponding statistics for the compounds outside the OPS were, respectively, 1.336 and 1.503, indicating that the likelihood of predicting a log *P* closer to the reference log *P* value is greater for compounds in the OPS. It can, thus, be inferred that not only is the model not applicable to the compounds outside the OPS, but in the absence of a validation procedure like OPS, it is impossible for the user to discriminate between "good" and "bad" predictions.

F. CONCLUSIONS

A log *P* predictor has been developed using information-rich *E* values as structure descriptors. These structure descriptors are sensitive to small structural changes, and, thus, can objectively account for interaction terms explained by special "factors" in group contribution methods for predicting log *P*. The rigorous application of regression diagnostic techniques makes VLOGP a robust model, which, coupled with the estimate validation algorithms of OPS, provides reliable calculation of log *P* values solely from molecular structure. The capabilities to check whether a submitted structure is outside the OPS gives users a tool to decide when not to accept a predicted value of log *P*.

REFERENCES AND NOTES

(1) Collander, R. Permeability of Plant Cells. *Ann. Rev. Plant Physiol.* **1957**, *8*, 335−348.

(2) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The Linear Free Energy Relationship between Partition Coefficients and the Aqueous Solubility of Organic Liquids. *J. Org. Chem.* **1968**, *33*, 347−350.

(3) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and their Uses. *Chem. Rev.* **1971**, *71*, 525−616.

(4) Hansch, C.; Leo, A. *Exploring QSAR: ACS Professional Reference Book*; American Chemical Society: Washington, DC, 1995.

(5) Purcell, W. P.; Bass, G. E.; Clayton, J. M. *Strategy of Drug Design, A Guide to Biological Activity*; Wiley: New York, 1973; Appendix I.

(6) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.

(7) Rekker, R. *The Hydrophobic Fragmental Constant*; Elsevier Scientific Publishing: Amsterdam, 1976.

(8) Hopfinger, A. J.; Battershell, R. D. Application of SCAP to Drug Design. 1. Prediction of Octanol−Water Partition Coefficient Using Solvent-Dependent Conformational Analysis. *J. Med. Chem.* **1976**, *19*, 569−573.

(9) Kohler, M. G.; Grigoras, S.; Dunn III, W. J. The Relationship between Chemical Structure and the Logarithm of the Partition Coefficient. *Quant. Struct.-Act. Relat.* **1988**, 150−159.

(10) Kasai, K.; Umeyama, H.; Tomonaga, A. The Study of Partition Coefficients. The Prediction of Log *P* Value Based on Molecular Structure. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 2701−2706.

(11) Bodor, N.; Gabayani, Z.; Wong, C. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.

(12) Chou, J. T.; Jurs, P. C. Computer-Assisted Computation of Partition Coefficients from Molecular Structure Using Fragment Constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172−178.

(13) Sasaki, Y.; Kubodera, H.; Matuszaki, T.; Umeyama, H. Prediction of Octanol/Water Partition Coefficients Using Parameters Derived from Molecular Structures. *J. Pharmacobio-Dyn.* **1991**, *14*, 207−214.

(14) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated Log *P* Calculation Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.

(15) Maylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.

(16) Ghose, A.; 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*, 565−577.

(17) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley Interscience: New York, 1979; p 19.

(18) Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.-Chim. Ther.* **1984**, *19*, 71−78.

(19) Partition Coefficient Determination and Estimation; Dunn III, W. J., Block, J. H., Pearlman, R. S., Eds.; Pergamon Press: New York, 1986.

(20) BioByte Corp., P.O. Box 517, Claremont, CA 91711, U.S.A.

(21) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression & Outlier Detection*; John Wiley & Sons: New York, 1987.

(22) Leo, A. The Octanol−Water Partition Coefficient of Aromatic Solutes: The Effect of Electronic Interactions, Alkyl Chains, Hydrogen Bonds, and ortho-Substitution. *J. Chem. Soc., Perkin Trans.* **1983**, *2*, 825−838.

(23) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(24) Kier, L. B. Shape Indices of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

(25) Gombar, V. K.; Jain, D. V. S. Quantification of Molecular Shape and Its Correlation with Physicochemical Properties. *Indian J. Chem.* **1987**, *26A*, 554−555.

(26) BMDP Statistical Software Manual; Dixon, W. J., Chief Ed.; University of California Press: Los Angeles, CA, 1988.

(27) Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. Assessment of Methods Used for Predicting Lipophilicity: Application to Nucleosides and Nucleoside Bases. *J. Comput. Chem.* **1993**, *14*, 1019−1026.

CI960028N