

Estimation of Molecular Linear Free Energy Relationship Descriptors by a Group Contribution Approach. 2. Prediction of Partition Coefficients

James A. Platts,[†] Michael H. Abraham,^{*,†} Darko Butina,[‡] and Anne Hersey[‡]

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K., and Science Development Group, Glaxo Wellcome Research and Development, Park Road, Ware SG12 0DP, U.K.

Received May 14, 1999

A previously published method for the prediction of molecular linear free energy relationship descriptors is tested against experimentally determined partition coefficients in various solvent systems. Sets of partition data between water and octanol, cyclohexane, and chloroform were taken from the literature. For each set of partition data used, r^2 values ranged from 0.8 to 0.9 and RMS errors from 0.7 to 1.0 log unit, comparable to errors obtained with previously published models for octanol–water partition. Modified solvation equations for water–octanol and water–cyclohexane partition are presented, and their implications discussed. The possibility of applying the current approach to a wide range of solvation and transport properties is put forward.

INTRODUCTION

In a previous study¹ (hereafter referred to as paper 1) we developed a method for the estimation of molecular linear free energy relationship (LFER) descriptors. These descriptors are designed to reflect the fundamental molecular properties important in solvation-related processes, namely, size, polarity, and hydrogen bonding. Linear combinations of these properties are widely employed in modeling processes such as solubility, partition between solvents, and passive biological transport.²

A series of solvation properties, SPs, are related to descriptors via the general solvation eq 1, where the descriptors are defined as follows: R_2 is an excess molar

$$\log \text{SP} = c + rR_2 + s\pi_2^{\text{H}} + a\sum\alpha_2^{\text{H}} + b\sum\beta_2^{\text{H}} + v(\text{Vx}) \quad (1)$$

refraction. π_2^{H} is a combined dipolarity/polarizability descriptor. $\sum\alpha_2^{\text{H}}$ is the overall solute hydrogen bond acidity. $\sum\beta_2^{\text{H}}$ is the overall solute hydrogen bond basicity. Vx is McGowan's³ characteristic molecular volume ($\text{cm}^3 \text{mol}^{-1}/100$). The set of coefficients, c , r , s , a , b , and v characterize the system in which the SPs are measured. For example, if SP represents the solubility of gases and vapors in a solvent, a is the basicity of the bulk solvent (since acidic molecules will interact *via* hydrogen bonding with basic solvents), while similarly s and b are the bulk solvent's polarity/polarizability and acidity, respectively. r relates the solvent's ability to interact with n - and π -electrons, and v is a combination of an endoergic cavity term and an exoergic dispersion term. An equation analogous to eq 1, for use in processes involving gases and vapors, simply substitutes the gas–hexadecane partition coefficient, $\log L^{16}$, for Vx . In this study we

concentrate solely on processes described by eq 1; those modeled using $\log L^{16}$ will be dealt with in a separate study.

Typically, Vx and R_2 can be calculated from structure: Vx is a simple sum of atom and bond contributions,³ while R_2 is easily obtained from the refractive index or summed from fragment values. The other molecular descriptors must, however, be found from experiment. The original scales of α_2^{H} and β_2^{H} for 1:1 complexation were set up from $\log K$ values for complexation in tetrachloromethane.² These were then modified using a number of partition processes to derive the overall or summation $\sum\alpha_2^{\text{H}}$ and $\sum\beta_2^{\text{H}}$ values. Similarly, π_2^{H} was originally found from gas chromatographic measurements on non-hydrogen-bonding columns.⁴ With the establishment of these scales of π_2^{H} , $\sum\alpha_2^{\text{H}}$, and $\sum\beta_2^{\text{H}}$, their calculation from partition coefficients, chromatographic retention parameters, or other physical measurements became possible.^{2,5a}

Once these scales were established, regression against physicochemical properties, such as partition coefficients, established the coefficients c , r , s , a , b , and v for the property in question: see ref 5 for some recent reviews of the method. One such property is the water–octanol partition coefficient, $\log P(\text{oct})$, widely used as a measure of hydrophobicity in the pharmaceutical, environmental, toxicological, and chemical engineering fields.⁶ Equation 2 shows the coefficients

$$\log P(\text{oct}) = 0.088 + 0.562R_2 - 1.054\pi_2^{\text{H}} + 0.032\sum\alpha_2^{\text{H}} - 3.460\sum\beta_2^{\text{O}} + 3.814(\text{Vx}) \quad (2)$$

which characterize $\log P(\text{oct})$.^{5b} Equation 2 was found from over 600 $\log P(\text{oct})$ measurements, with a standard deviation of 0.11 log unit. $\log P(\text{oct})$ is the most important of the “wet solvents”, wherein the alternative $\sum\beta_2^{\text{O}}$ basicity scale must be used in place of $\sum\beta_2^{\text{H}}$. All other solvent systems considered in the present study employ $\sum\beta_2^{\text{H}}$.

[†] University College London.

[‡] Glaxo Wellcome Research and Development.

Water–chloroform partition can be similarly treated, resulting in the equation⁷

$$\log P(\text{chl}) = 0.321 + 0.168R_2 - 0.379\pi_2^{\text{H}} - 3.170\Sigma\alpha_2^{\text{H}} - 3.409\Sigma\beta_2^{\text{H}} + 4.149(\text{Vx}) \quad (3)$$

which yielded a standard deviation of 0.25 for 330 measurements. Another important partition process, water–cyclohexane partition or $\log P(\text{cyc})$, has also been modeled in this manner, yielding the solvation eq 4, giving a standard deviation of 0.14 log unit from 227 measurements.^{5b} Analogous equations for around 25 water–solvent partitions have been established.⁵

$$\log P(\text{cyc}) = 0.159 + 0.784R_2 - 1.678\pi_2^{\text{H}} - 3.740\Sigma\alpha_2^{\text{H}} - 4.929\Sigma\beta_2^{\text{H}} + 4.577(\text{Vx}) \quad (4)$$

To escape the reliance on experimental data for the determination of new π_2^{H} , $\Sigma\alpha_2^{\text{H}}$, and $\Sigma\beta_2^{\text{H}}$ values, we set out in paper 1 a general method for their estimation directly from structure. With a database of between 2500 and 3500 values for each descriptor, we were able to identify common substructures and, through a process of multiple linear regression, evaluate contributions of each substructure to each descriptor. Our final model used 81 fragment values for R_2 , π_2^{H} , $\Sigma\beta_2^{\text{H}}$, $\Sigma\beta_2^{\text{O}}$, and $\log L^{16}$, with a separate set of 51 fragments for calculation of $\Sigma\alpha_2^{\text{H}}$. Typically, errors of around 0.05–0.15 log unit (for values covering a range of 2–6 log units) were found.

The goal of the present study is to employ our group contribution method for sets of molecules for which partition data are available and then to apply the relevant solvation equation, e.g., eq 2, to predict the partition data. In this way, we aim to test the reliability and generality of the estimation method, and to compare it against established methods for the estimation of partition data. In addition to this validation process, this prediction should highlight fragments and atom types modeled poorly by the original approach, in effect serving to extend the effective training set of the group contribution method.

These partition processes are important in their own right; for example, $\log P(\text{oct})$ is widely used as a hydrophobicity index, and $\Delta \log P$, as $\log P(\text{oct}) - \log P(\text{cyc})$ or $\log P(\text{oct}) - \log P(\text{alkane})$, has been used to predict blood–brain distribution.^{5a} Four partition processes, $\log P(\text{oct})$, $\log P(\text{cyc})$, $\log P(\text{chl})$, and $\log P(\text{PGDP})$ (where PGDP = propylene glycol diargonate) have been put forward as a “critical quartet” of solvent systems,⁸ designed to encapsulate all relevant information about a solute’s partition and transport. However, their main relevance to this study lies in its testing of the ability of the calculations from paper 1 to reproduce well-characterized experimental data. If we can demonstrate that acceptable errors arise from these calculations, predictions of physicochemical or biological transport properties for which appropriate solvation equations have been developed⁵ can be made.

A wide range of methods for the estimation of partition data, almost exclusively $\log P(\text{oct})$, have been proposed. Among these are Rekker’s fragmentation method,⁹ Hansch and Leo’s “constitutive” CLOGP,¹⁰ Klopman’s KLOGP,¹¹ and Ghose *et al.*’s ALOGP.¹² These typically estimate \log

$P(\text{oct})$ by identifying important substructures and atom types, either predefined or from some fragmentation scheme, and assigning a contribution to $\log P(\text{oct})$ from each fragment. These contributions are found either from multiple linear regression^{9,11,12} or by comparison with measured values for fragments.¹⁰

Such fragmental methods have generally been restricted to prediction of $\log P(\text{oct})$, since large training sets of data are needed. One exception is Klopman’s¹³ treatment of aqueous solubility, wherein atomic and functional group contributions were assigned. In general, however, important physicochemical and biological properties cannot be treated in this way due to insufficient training data being available, preventing simple predictions of, e.g., water–cyclohexane or blood–brain partition being made. Another approach treats the effects of solvation as part of a quantum mechanical calculation, either explicitly or implicitly. Examples of this approach include Cramer and Truhlar’s OMNISOL package¹⁴ and the GB/SA approach.¹⁵ These calculations must be parametrized for a given solvent, usually including one or more of water, octanol, chloroform, and alkanes.

It has been proposed⁶ that a molecule’s $\log P$ values in solvent systems are correlated, such that one can calculate one $\log P$ from knowledge of another, as shown in eq 5

$$\log P_1 = a \log P_2 + b \quad (5)$$

However, Collander¹⁶ has pointed out that such a method is only valid for families of related solvent systems, such that water–decanol partition could be estimated from $\log P(\text{oct})$, but $\log P(\text{cyc})$ could not. The reason for this is clear; eqs 2–4 show very different dependencies on molecular properties. For example, $\log P(\text{oct})$ has no dependence on $\Sigma\alpha_2^{\text{H}}$, but $\log P(\text{cyc})$ has a large $\Sigma\alpha_2^{\text{H}}$ coefficient, a difference which cannot be modeled by the simple linear model of eq 5. Put another way, solvation equations such as eqs 2–4 are not parallel when considered as vectors in “descriptor space”.

Several attempts to classify the predictive ability of $\log P(\text{oct})$ models have been published; Bodor *et al.*¹⁷ and Schüürman¹⁸ have reported evaluations of several $\log P(\text{oct})$ estimation methods, comparing their performance against sets of experimentally measured data. These studies typically report that “constructive” methods such as CLOGP are rather better predictors than group contribution methods such as KLOGP. A very interesting study by Ghose *et al.*¹² compared the performance of ALOGP and CLOGP as a function of molecular size, and found CLOGP to be the best choice for smaller compounds, but ALOGP to be superior for molecules with more than 45 atoms. Also, by subdividing their entire set of $\log P(\text{oct})$ data into structural subtypes, they were able to highlight specific classes of molecules for which each method performs well or poorly, an approach we have used in the present work.

RESULTS

Several sets of data were selected from the literature: one set each for partition between water and octanol, chloroform, and cyclohexane was taken from the MedChem97 database,¹⁹ along with the water–octanol set from Buchwald and Bodor’s recent review.¹⁶ In all cases, care was taken that the data had no reported errors and, where appropriate,

Table 1. Observed and Calculated Values for Bodor's log *P*(oct) Dataset^a

name	logP(oct)		name	logP(oct)		name	logP(oct)		name	logP(oct)	
	calcd	obsd		calcd	obsd		calcd	obsd		calcd	obsd
codeine	1.71	1.14	atenolol	1.17	0.16	AcAlaLeuNH ₂	-0.72	-0.54	27-crown- 9, bz	0.77	0.23
flufenamic acid	3.84	5.25	bunitrolol	2.41	2	AcTrpValNH ₂	0.27	0.73	30-crown-10, bz	0.75	0.03
indomethacin	4.12	4.27	metipranolol	3.86	2.66	AcSerValNH ₂	-2.12	-1.53	33-crown-11, bz	0.71	-0.09
methadone	5.32	3.93	metoprolol	3.12	1.88	AcValIleGlyNH ₂	-0.9	-0.45	18-crown-6, db	2.17	2.2
morphine	1.3	0.76	oxprenolol	2.86	2.18	AcTrpGlyPheNH ₂	0.1	0.99	24-crown-8, db	2.1	2.11
phenylbutazone	4.72	3.16	penbutolol	4.95	4.15	AcLeuThrLeuNH ₂	-0.22	0.24	27-crown-9, db	2.08	1.63
aprimidine	6.25	4.86	propranolol	3.43	3.09	AcLeuSerPheNH ₂	-0.25	0.23	30-crown-10, db	2.04	1.8
asocainol	6.75	4.85	theophylline	-0.87	-0.02	AcAlaTyrLeuNH ₂	-0.37	-0.04	30-crown-10, db	2.04	1.82
carocamide	2.19	1.38	furosemide	1.53	2.03	uridine	-3.45	-1.98	33-crown-11, db	2.02	1.45
diltiazem	3.65	2.7	veralipride	0.36	1.47	dUrd	-2.46	-1.62	48-crown-16, db	1.86	0.52
disopyramide	3.79	2.58	lidocaine	2.57	2.26	ddUrd	-1.46	-0.89	PCBnoCl	4.22	3.98
mexiletine	2.9	2.15	tetracaine	2.36	3.73	ddeUrd	-1.72	-1.07	PCB, 4-	4.66	4.61
morizine	3.21	2.98	piracetam	-1.72	-1.54	FddUrd	-1.36	-0.49	PCB, 2,2'	5.1	4.73
nicainoprol	3.02	1.63	lysergide	2.18	2.95	cytidine	-3.36	-2.51	PCB, 4,4'	5.1	5.58
procainamide	0.73	0.88	tiapride	-0.19	0.9	dCyd	-2.36	-1.77	PCB, 2,4,5-	5.89	5.81
propafenone	4.4	4.63	bromopride	1.61	2.83	ddCyd	-1.36	-1.3	PCB, 2,2',5-	5.7	5.6
quinidine	2.61	2.64	phenobarbital	1.52	1.47	ddeCyd	-1.63	-1.57	PCB, 2,2',5,5'	6.32	6.09
sotalol	0.88	0.24	caffeine	-0.62	-0.07	FddCyd	-1.27	-0.92	PCB, 2,3,4,5-	6.57	6.41
verapamil	5.71	3.83	cocaine	2.42	2.3	Ado	-3.19	-1.23	PCB, 2,2',4, 5,5'	6.9	6.44
chloramphenicol	0.86	1.14	nicotine	0.92	1.17	dAdo	-2.19	-0.55	PCB, 2,3,4,5,6-	7.06	6.52
trimethoprim	0.79	0.91	diazepam	3.13	2.82	ddAdo	-1.18	-0.22	PCB, 2,2',4,4',5,5'	7.37	6.8
atropine	2.69	1.83	fluphenazine	3.38	4.36	ddeAdo	-1.48	-0.36	PCB, 3,3',4,4',5,5'	7.2	7.55
phenytoin	2.52	2.47	triflupromazine	4.71	5.19	FddAdo	-1.1	0.08	PCB, 2,2',3,3',4,4',6-	7.86	6.99
imipramine	4.73	4.8	methotrimeprazine	4.25	4.68	Guo	-3.82	-1.89	PCB, 2,2',3,3',5,5',6,6'	8.26	7.15
alizapride	0.22	1.79	meprobamate	2.24	0.7	dGuo	-2.82	-1.3	PCB, 2,2',3,3',4,5,5', 6,6'	8.74	8.16
amisulpride	-0.26	1.1	perphenazine	2.98	4.2	ddGuo	-1.85	-1.01	PCBdecachlor	9.23	8.26
sulpiride	0.06	0.62	promethazine	3.53	4.75	ddeGuo	-2.11	-1.21	PCDD, no Cl	1.93	4.37
thiethylperazine	4.09	5.4	promazine	3.53	4.55	Thd	-1.89	-1.17	PCDD, 1-	2.54	5.05
cimetidine	-1.19	0.4	sultopride	0.47	1.06	ddThd	-0.9	-0.63	PCDD, 2,7-	3.47	6.38
diphenhydramine	4.08	3.4	thioridazine	5.66	5.9	ddeThd	-1.16	-0.81	PCDD, 1,2,4-	4.43	7.47
chlorothiazide	-2.31	-0.24	trifluoperazine	3.82	5.03	FddThd	-0.76	-0.28	PCDD, 2,3,7,8-	4.98	6.42
terazosin	-0.41	-0.38	trimeprazine	4.07	4.81	15-crown-5, bz	0.89	0.91	PCDD, 1,2,3,7,8-	5.74	6.64
haloperidol	3.91	3.36	pindolol	1.68	1.75	18-crown-6, bz	0.87	0.58	PCDF no Cl	3.82	4.12
acebutolol	2.17	1.77	AcTyrPheNH ₂	0.79	0.54	21-crown-7, bz	0.83	0.57	PCDF, 2,8-	4.7	5.65
alprenolol	3.62	3.1	AcThrValNH ₂	-1.61	-1.25	24-crown-8, bz	0.81	0.45	PCDF, 1,2,7,8-	5.57	6.23

^a $n = 140$, $r^2 = 0.894$, $RMS = 0.966$, $av\ err\ 0.206$, $max\ dev\ 3.04$ (PCDD, 1,2,4-).

corresponded to the neutral species and not to some ionization product. Descriptors R_2 , π_2^H , $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$, and $\Sigma\beta_2^O$ were calculated for each compound in each set as described in paper 1, with V_x obtained as before.

(i) Prediction of log *P*(oct). Observed and calculated values (using eq 2) for 140 of Bodor's data are presented in Table 1, along with the statistics of the fit. Overall, our calculation method results in reasonable agreement with experiment, with an r^2 value of 0.894 and an RMS error of 0.966. This places our method somewhere in the middle of the methods considered by Bodor et al., with considerably better accuracy than the simplest fragmental methods, e.g., for Rekker, $r^2 = 0.761$ and $RMS = 1.316$ or, for KLOGP, $r^2 = 0.695$ and $RMS = 1.473$. As expected, however, the constitutive models fare rather better (CLOGP, $r^2 = 0.934$, $RMS = 0.684$; ACD log *P*, $r^2 = 0.965$, $RMS = 0.498$).

In addition to examining the entire set of molecules, Bodor also reported the performance of the various models for two subsets of the data, namely, drugs, peptides, and nucleosides, or D, P, and N (101 molecules) and for drugs only (68). In these cases, they found the fragmental methods came into their own, significantly increasing the accuracy of the fit, while constitutive methods noticeably declined in accuracy for these subsets. Our predictions for these subsets show for the D, P, and N set a very small increase in accuracy, with $r^2 = 0.886$ and $RMS = 0.930$, such that, in terms of r^2 at least, our model is comparable to CLOGP ($r^2 = 0.887$, RMS

$= 0.718$). For the drugs only set, the fit becomes slightly poorer than for either of the previous sets, $r^2 = 0.762$, $RMS = 0.946$.

A second set of 148 log *P*(oct) values, taken from the MedChem97 database, is presented in Table 2. This set consists mainly of drugs and agrochemicals, flagged in the database as having some activity, and also having a "starred" log *P*(oct) value. Agreement between calculated and experimental values is not as close as in Bodor's set of data; the statistics of this fit are $r^2 = 0.80$ and $RMS = 0.92$. CLOGP²⁰ for the same set gives $r^2 = 0.91$ and $RMS = 0.58$. In comparing these two models of log *P*(oct), one must bear in mind that CLOGP is to some degree trained on log *P** values from the MedChem97 database, and should indeed be more accurate for such values.

The method presented here is clearly not the most accurate predictor of log *P*(oct) currently available, and is not recommended in place of more accurate methods such as CLOGP. It should be apparent, however, that the method does deliver at least *reasonable* predictions of log *P*(oct) for a fairly wide range of complex molecules, and is even noticeably better than some commonly used models. Moreover, it is gratifying to see the accuracy obtained for molecules wholly unlike those used for training, such as the set of 10 di- and tripeptides.

The largest error in the prediction of log *P*(oct) (Tables 1 and 2) is for PCDD-1,2,4 (1,2,4-polychlorodibenzodioxin)

Table 2. Observed and Calculated log *P*(oct) Values from MedChem97^a

name	logP(oct)		name	logP(oct)		name	calcd	obsd
	calcd	obsd		calcd	obsd			
uridine	-3.5	-1.98	aprobarbital	1.26	1.37	spiperone	2.28	3.03
citric acid	-1.01	-1.72	1,2-propanediol-3,2-tolyloxy	1.53	1.41	sulindac	3.8	3.05
ascorbic acid	-2.96	-1.64	hydantoin-5-ethyl-5-phenyl	1.43	1.53	deoxycorticosteroneacetate	4.02	3.08
acyclovir	-1.96	-1.56	sulfabenz	1.72	1.55	phenyl-4-aminosalicylate	3.19	3.15
zalcitabine	-1.42	-1.3	dimethylphthalate	1.65	1.56	phenylbutazone	4.49	3.16
adenosine	-2.75	-1.23	nalidixic acid	0.65	1.59	fenbufen	3.6	3.2
thymidine	-1.9	-1.17	chromone-2-carboxylic acid	0.85	1.63	barbituricacid-5-allyl-5,1-me-butyl-2-thio	1.69	3.23
triethanolamine	-1.75	-1.0	3,4-methylenedioxyamphetamine	0.9	1.64	fenamiphos	3.68	3.23
stavudine	-1.16	-0.81	α -naphthylthiourea	1.64	1.66	2-phenylbenzimidazole	3.19	3.24
histamine	-0.47	-0.7	fludrocortisone	1.8	1.67	salicylanilide	3.36	3.27
sulfanilamide	-0.44	-0.62	ceterizine	2.86	1.7	diphenhydramine	3.93	3.27
phosphineoxidetris-1-aziridinyl	-1.3	-0.62	dichlorvos	0.23	1.7	renanlone	3.31	3.28
2-pyrazinecarboxamide	-1.59	-0.6	benzylalcohol-2-hydroxy-5-bromo	0.97	1.72	ketanserine	1.55	3.29
cefazolin	-1.2	-0.58	pindolol	1.67	1.75	azinphosethyl	1.35	3.4
busulfan	-0.38	-0.52	amphetamine	2.24	1.76	crufomate	3.27	3.42
oxamyl	-1.0	-0.47	propylgallate	1.19	1.8	alanycarb	2.05	3.43
2,5- <i>H</i> -furanone-3-chloro-4-dichloro-methyl-5-hydroxy	0.16	-0.44	3,6-diaminoacridine	1.2	1.83	2-phenethylisothiocyanate	2.8	3.47
<i>N</i> -acetylhomocysteine thiolactone	-1.1	-0.35	phentermine	2.71	1.9	propiconazole	4.09	3.5
methiazole	0.24	-0.34	methoxsalen	1.72	1.93	ibuprofen	4.66	3.5
cotinine	0.3	-0.32	strychnine	2.75	1.93	droperidol	2.87	3.5
acetazolamide	-1.85	-0.26	piroxicam	-0.67	1.98	triazophos	2.21	3.55
dacarbazine	-2.57	-0.24	podofilox	1.13	2.01	PCP	4.92	3.63
chlorothiazide	-2.36	-0.24	diethyltoluamide	2.6	2.02	propericiazine	3.24	3.65
normorphine	1.41	-0.17	dicoumarol	1.47	2.07	stanolone	3.81	3.66
thioguanine	-0.66	-0.07	prilocaine	2.3	2.11	phenthoate	3.34	3.69
sulfapyridine	0.57	0.0	benzoin	2.5	2.13	hexachlorocyclohexane	4.7	3.72
zidovudine	-0.84	0.05	carbamazepine	3.64	2.19	tetracaine	2.28	3.73
terbutaline	0.31	0.08	1,4-naphthoquinone-2-methyl	1.74	2.2	2,6-diisopropylphenol	4.31	3.79
phenicarbazide	-0.86	0.13	chlorpropamide	2.07	2.27	3-phenylpropylisothiocyanate	3.33	3.8
pentyletetrazole	1.74	0.14	<i>N</i> -methylcarbamate-1-naphthyl	2.79	2.36	azobenzene	3.96	3.82
atenolol	0.98	0.16	malathion	2.16	2.38	tomelukast	5.35	3.82
coramine	0.64	0.33	triazolam	3.59	2.42	disulfiram	3.85	3.88
thalidomide	0.04	0.33	meperidine	2.98	2.45	fentanyl	5.31	3.89
enprofylline	-0.16	0.33	benzoxazole2-amino-5-chloro	1.31	2.47	methadone	5.19	3.93
cianidol	0.53	0.36	phenytoin	2.36	2.47	phenothiazine	2.43	4.15
cycloheximide	1.65	0.55	eucalyptol	3.61	2.5	isradipine	3.12	4.18
cytoxin	1.49	0.63	2,5-dimethoxy4-bromoamphetamine	2.95	2.58	dichlorophen	3.88	4.26
ifosfamide	1.48	0.86	hycanthone	2.27	2.7	dicofol	5.36	4.28
phenacemide	0.72	0.87	altretamine	1.25	2.73	drometizole	1.85	4.31
primidone	0.31	0.91	phosmet	1.67	2.78	diclofenac	3.8	4.4
saccharin	-0.22	0.91	diazem	3.46	2.8	diflunisal	4.39	4.44
2,4-diamino-5,3,4-dimethoxybenzylpyrimidine	0.67	0.97	alphaprodine	3.48	2.83	3,5-diiodosalicylic acid	5.29	4.56
<i>di-p</i> -aminophenylsulfone	0.12	0.97	selegiline	3.14	2.9	chloroquine	3.87	4.63
<i>N</i> -desacetylcolchicine	2.41	1.1	2-phenyl-1,3-indanedione	2.25	2.9	piperonylbutoxide	5.1	4.75
propamocarb	1.22	1.12	<i>N</i> -methylcarbamate-3,5-dimethyl-4-methylthiophenyl	3.54	2.92	ethion	3.71	5.07
sulmazole	0.47	1.17	myclobutanil	3.8	2.94	mefenamid acid	3.79	5.12
vanillin	0.49	1.21	papaverine	3.93	2.95	chlorpromazine	4.22	5.19
4-iodoantipyrine	2.41	1.27	propranolol	3.23	2.98	endrin	5.1	5.2
colchicine	2.56	1.3	pyrilamine	2.52	2.98	cinnarizine	5.46	5.77
sorbic acid	1.57	1.33						

^a $n = 148$, $r^2 = 0.795$, RMS = 0.919, av err 0.134, max dev 2.65 (piroxicam).

which is predicted to have a log *P*(oct) of 4.43, compared to the experimental value of 7.47, an error of 3.04 log units. This is very slightly smaller than the largest error in CLOGP, i.e., 3.09 for terazosin, and in general the very largest errors are similar in size for both methods. CLOGP improves upon our method in the number of errors falling between 1 and 2 log units, with just 26 falling in this range compared with 48 using our method.

When observed log *P*(oct) is plotted against calculated log *P*(oct) for Bodor's test set (Figure 1), it is immediately

apparent that the slope and intercept of the line of the best fit, with values of 0.862 ± 0.025 and 0.484 ± 0.089 , are significantly different from their ideal values of 1.0 and 0.0, respectively. One can envisage two possible reasons for this: firstly, an insufficient range of compounds may have been used in developing eq 2 for log *P*(oct). Secondly, there may be some systematic errors in the calculations, giving rise to the nonideal slope and intercept.

We can check the former possibility by developing a new equation for log *P*(oct) based on a much wider range of

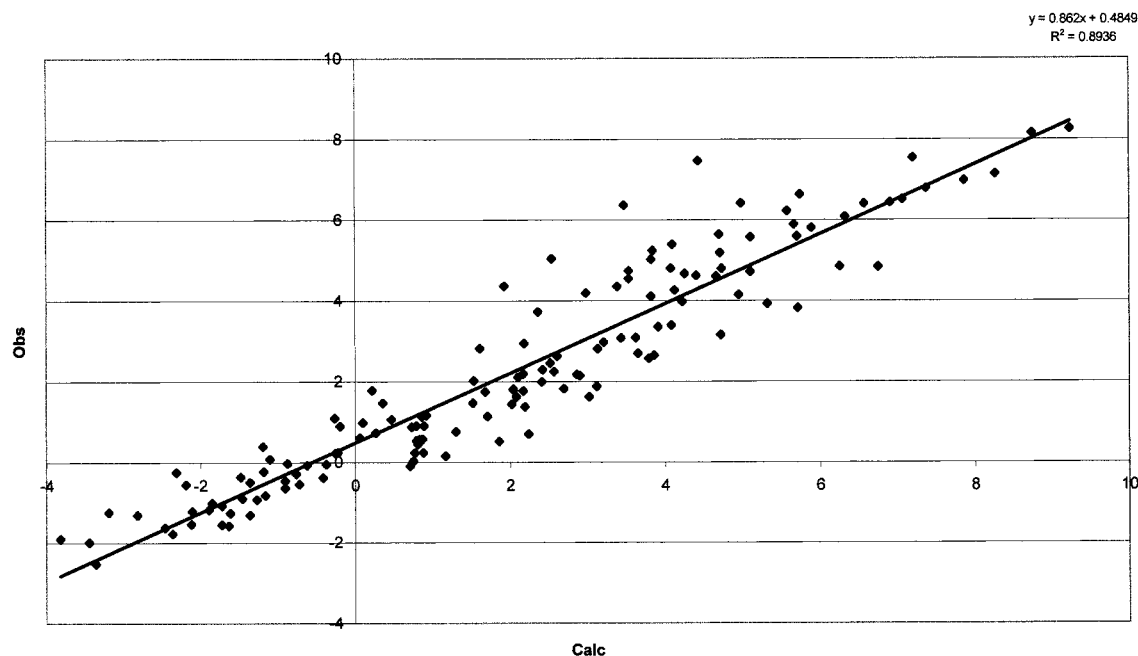


Figure 1. Observed vs calculated log $P(\text{oct})$ for Bodor's set of 140 molecules.

Table 3. Coefficients and Standard Deviations for Eqns 2 and 6

	eq 2		eq 6	
	coeff	sd	coeff	sd
c	0.088	0.015	0.319	0.023
r	0.562	0.014	0.972	0.021
s	-1.054	0.021	-0.870	0.023
a	0.032	0.021	-0.240	0.027
b	-3.460	0.026	-2.499	0.023
v	3.814	0.015	2.674	0.020

compounds. The entire set of starred log $P(\text{oct})$ values was extracted from the MedChem97 database. Species known to be calculated wrongly, such as metal complexes, charged molecules, and disconnected structures (e.g., ion pairs), were excluded from this set, as were any with warnings regarding the quality of the log $P(\text{oct})$ measurement. Regression of the remaining 8844 log $P(\text{oct})$ values against the five calculated descriptors R_2 , π_2^H , $\Sigma\alpha_2^H$, $\Sigma\beta_2^O$, and V_x resulted in the following solvation equation:

$$\log P(\text{oct}) = 0.315 + 0.926R_2 - 0.841\pi_2^H - 0.241\Sigma\alpha_2^H - 2.506\Sigma\beta_2^O + 2.674(V_x) \quad (6)$$

$$n = 8844, r^2 = 0.826, r_{CV}^2 = 0.826, \text{RMS} = 0.674, F = 8416$$

Equations 2 and 6 are qualitatively very similar (see Table 3), with a large positive coefficient of V_x and a large negative contribution from $\Sigma\beta_2^H$, and smaller contributions from R_2 , π_2^H , and $\Sigma\alpha_2^H$. This is pleasing to note, as it indicates eq 6 to be a realistic model of log $P(\text{oct})$, in keeping with several previous studies.²¹ The statistics of eq 6 are reasonable, though again not as impressive as, say, those of CLOGP¹² ($r^2 = 0.922$ and $\text{RMS} = 0.50$ for 8141 compounds), but on par with those of other fragmental models (e.g., ALOGP-(old),¹² $r^2 = 0.828$ and $\text{RMS} = 0.73$ for 8364 compounds). The RMS error of 0.701 from eq 6 is also a substantial improvement over the 0.966 found for Bodor's test set. Thus, for this large set of data we can restate that the two major

Table 4. Range of Descriptors Used in Eq 6

	min	max	mean
R_2	-0.50	5.01	1.42
π_2^H	-0.13	6.53	1.69
$\Sigma\alpha_2^H$	0.0	2.72	0.33
$\Sigma\beta_2^H$	0.0	5.93	1.023
V_x	0.18	7.08	1.657

factors which determine a molecule's log $P(\text{oct})$ value are its size, which increases log P , and its hydrogen bond basicity, which reduces log P . Smaller contributions come from molecular polarity (reducing log P) and the presence of n - and π -electron pairs (increasing log P).

The small but nonzero contribution of $\Sigma\alpha_2^H$ in eq 6 is worthy of further discussion. A considerably larger range of $\Sigma\alpha_2^H$ was employed in this regression (see Table 4) than in eq 2. In addition, $\Sigma\alpha_2^H$ was one of the most accurately calculated descriptors in paper 1, and also the largest intercorrelation with $\Sigma\alpha_2^H$ in eq 6 has just $r = 0.42$. We therefore have some confidence in the coefficient of -0.240 reported, despite eq 2 indicating that hydrogen bond acidity is insignificant in determining log $P(\text{oct})$.

Applying eq 6 to the 140 molecules of Bodor's set in Table 1 results in only a small increase in r^2 to 0.903 compared with 0.894 from eq 2. However, the rms error is lowered by over 0.10 log unit to 0.862 from 0.966 previously. The slope and intercept of the best-fit line are now 1.124 ± 0.031 and -0.322 ± 0.099 , respectively. These values are a considerable improvement on those found using eq 2, and seem to indicate that some, if not all, of the discrepancy of slope and intercept can be assigned to shortcomings in the original log $P(\text{oct})$ equation.

Following Ghose et al., we have broken down the set of 8844 molecules into common structural groups, as set out in Table 5. In this manner, it is possible to identify classes of compounds which are particularly well or poorly modeled using eq 6, giving some idea of the strengths and weaknesses of our approach. By far the worst class of molecules in this set are the N -oxides, with $r^2 = 0.327$ and $\text{RMS} = 0.847$.

Table 5. Comparison of log $P(\text{oct})$ Prediction from Eq 6 by Compound Class

type	r^2	RMS dev	max dev	max log P	min log P	no. of data
carboxylic acid	0.776	0.607	2.022	6.3	-2.57	426
alcohol	0.882	0.692	2.016	8.42	-3.7	1084
aldehyde	0.808	0.507	1.554	3.71	-2.05	97
aliphatic primary amine	0.878	0.67	1.78	8.38	2.82	140
aliphatic secondary amine	0.766	0.749	1.989	4.9	-1.5	196
aliphatic tertiary amine	0.781	0.731	2.316	7.57	-1.47	504
aromatic primary amine	0.796	0.629	1.944	5.08	-2.51	660
aromatic secondary amine	0.841	0.606	1.673	5.41	-1.89	176
aromatic tertiary amine	0.703	0.679	1.731	5.44	-1.9	157
carboxamide	0.778	0.695	2.856	5.85	-3.09	1643
imidazole	0.863	0.77	2.448	6.3	-2.47	600
ketone	0.77	0.655	1.877	9.07	-1.84	588
<i>N</i> -oxide	0.327	0.847	2.465	2.93	0.45	9
nitro	0.761	0.691	2.748	5.44	-1.59	925
nucleoside	0.752	0.566	1.492	1.35	-2.51	122
phenol	0.812	0.661	2.901	8.42	-2.11	577
pyridine	0.764	0.609	2.141	7	-1.58	493
pyrimidine	0.702	0.709	2.656	5.7	-1.42	304
pyrrole	0.814	0.624	1.605	6.4	0.38	168

Table 6. log $P(\text{oct})$ Predictions for *ortho*-Substituted Phenols

substituent	obsd	calcd	substituent	obsd	calcd
Me	1.95	1.70	NH ₂	0.62	0.93
F	1.71	1.09	NO ₂	1.79	1.53
Cl	2.15	2.09	CO ₂ H	2.26	2.23
Br	2.35	2.34	OH	0.88	0.55
I	2.65	2.77	CO ₂ Me	2.34	2.62
OMe	1.32	1.43	CONH ₂	1.28	1.30
COMe	1.92	2.44	NHCOMe	0.72	0.91
CN	1.61	0.90			

Poor performance for this class of compounds was also found by Ghose et al. using both CLOGP and ALOGP. For the other 18 classes, r^2 values range from 0.70 for pyrimidines and aromatic tertiary amines to 0.88 for alcohols and aliphatic primary amines. In terms of RMS values, aldehydes (RMS = 0.51) and nucleosides (0.57) stand out as the best modeled. In contrast, imidazoles (0.77), aliphatic tertiary (0.73) and secondary amines (0.75), and pyrimidines (0.71) are notably poorly modeled. Fairly large maximum errors are observed, most notably for phenols, carboxamides, and pyrimidines. There seems to be little or no correlation among r^2 , RMS, and maximum deviations, with only pyrimidines and *N*-oxides giving poor results in all three statistics, suggesting that our method should be considered risky for such molecules.

Any method's ability to deal with intramolecular interactions is crucial, as many such interactions may be present in large molecules. In paper 1 we used *ortho*-substituted phenols to demonstrate this: we have now predicted log $P(\text{oct})$ for the same set of molecules to test the method further. Table 6 contains observed and calculated log $P(\text{oct})$ values for 15 *ortho*-substituted phenols. Overall the agreement is good, with an average error of 0.05 and RMS = 0.331, considerably better than for the more complex molecules in Tables 1 and 2, as might be expected. Two fairly large exceptions are found, for cyano and fluoro substituents, indicating that some fine-tuning of the intramolecular hydrogen bond definitions might be necessary. In general, however, we can conclude that such intramolecular interactions are well handled by the current method.

(ii) Prediction of log $P(\text{chl})$ and log $P(\text{cyc})$. It is clear that a great many methods have been developed for the prediction of log $P(\text{oct})$, reflecting its importance throughout chemistry. Considerably less attention, however, has been directed toward other partition processes. It is the single biggest advantage of the method described here and in paper 1 that one calculation of descriptors can yield predictions for a wide range of partition and related processes and, in principle at least, can predict values for any appropriate transport process. To demonstrate this, sets of data have been taken from the MedChem97 database for water–chloroform, log $P(\text{chl})$, and water–cyclohexane, log $P(\text{cyc})$, both of which, unlike log $P(\text{oct})$, have a strong dependence on $\Sigma\alpha_2^H$, the solute hydrogen bond acidity.

Observed and calculated values for 107 log $P(\text{chl})$ values are reported in Table 7. From eq 3 it is apparent that log $P(\text{chl})$ depends largely on $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$, and V_x , with only small contributions from R_2 and π_2^H . Overall, the accuracy found for this dataset is similar to that for the two log $P(\text{oct})$ sets in Tables 1 and 2, with an r^2 value of 0.864 and an RMS error of 0.908. To the best of our knowledge, this is the first example of such a calculation of log $P(\text{chl})$ from structure, and so no comparisons can be made with other methods. Cramer and Truhlar's OMNISOL method¹⁴ appears capable of predicting such partition data, but to date only training sets of relatively simple molecules have been published. Thus, we can state that the current method is the best, if indeed the only, method for the prediction of log $P(\text{chl})$ directly from structure.

Very large errors, above 2 log units, are found for just two log $P(\text{chl})$ values, those for methaqualone (obsd, 1.12; calcd, 4.75) and bicyclo-2,2,1-hepta-2,5-dien-7-olbenzoate (BHD) (obsd, 2.23; calcd 5.11). Neither of these is easy to explain away, especially as both have reasonable log $P(\text{oct})$ predictions ((methaqualone) obsd, 2.50; calcd, 3.11; (BHD) obsd, 3.03; calcd, 3.70). Both molecules have no acidic hydrogens, and hence $\Sigma\alpha_2^H = 0.0$, so errors in acidity prediction cannot explain the discrepancy. It may indeed be that the values taken from the MedChem97 database are in error. Omitting these two values results in $r^2 = 0.889$ and RMS = 0.797. As with log $P(\text{oct})$, a substantial number of errors, 23 in all, lie in the range of 1–2 log units. The use of test sets of log $P(\text{oct})$ above could be criticized for failing to test the prediction of $\Sigma\alpha_2^H$, due to the very small a coefficient in any model of log $P(\text{oct})$. However, the similar accuracy found for log $P(\text{oct})$ and log $P(\text{chl})$ supports the findings of paper 1, in which the correlation of $\Sigma\alpha_2^H$ was of rather better accuracy than that of π_2^H or $\Sigma\beta_2^H$.

Table 8 contains observed and calculated log $P(\text{cyc})$ values for 132 compounds, once more taken from the MedChem97 database. Of the three water–solvent partitions examined here, this is the most stringent test of the prediction method; as eq 4 shows, log $P(\text{cyc})$ has large contributions from all five descriptors. This is perhaps reflected in the fit statistics, which are the worst from the four datasets examined here. With $r^2 = 0.813$ and RMS = 0.968, these results are rather worse than for log $P(\text{chl})$. Unlike the log $P(\text{chl})$ set, however, the worst predictions cannot be explained away by the molecules belonging to a class of molecules, such as zwitterions, for which errors are expected. The very worst error, 3.62 log units, is found for the relatively simple sulfanilacetamide. Why this molecule is so poorly predicted is hard to explain,

Table 7. Observed and Calculated log *P*(chl) Values from MedChem97^a

name	log <i>P</i> (chl)		name	log <i>P</i> (chl)		name	log <i>P</i> (chl)	
	calcd	obsd		calcd	obsd		calcd	obsd
1,1-DiMe-3- <i>m</i> -CF ₃ -phenyl-urea	2.5	2.29	<i>N</i> -methylcarbamate-1-naphthyl	3.29	2.44	dibenzylphosphoric acid	1.52	1.09
methimazole	-0.01	-0.57	sulfathiazole	0.36	-0.73	2,2,6,6-tetramethyl-3,5-heptanedione	4.32	4.02
thenoyltrifluoroacetone	3.03	2.28	mercaptobenzothiazole	3.21	2.2	di-Me-1-OH-2,2,2-tri-Cl-ethyl-phosphonate	-0.21	-0.1
quinazoline-2,4-dione	-0.64	-0.88	1,4-naphthoquinone-2-hydroxy	1.18	2.26	phenoxymethylpenicillin	1.07	1.73
hydantoin-5,5-dimethyl	-1.27	-1.41	guaifenesin	0.6	0.4	2,2-thiazolylazo-4-methylphenol	3.21	3.83
trifluoroacetylacetone	1.56	0.3	homatropine	3.58	2.86	chloramphenicol	0.87	-0.48
benzanilide	2.84	2.81	benzotriazole	-0.62	-0.08	11B,17A,20A,21-tetrahydroxy-4-pregene-3-one	0.56	0.04
nicotinamide	-1.02	-1.08	benzotriazole-2-hydroxy	-1.56	-1.03	metepa	1.57	1.0
niacin	-0.84	-2.05	isoprenaline	-1.83	-1.6	1,3-propanediamine	-1.54	-1.45
2-furaldehyde-5-hydroxy-methyl	-0.61	-0.22	5-fluorouracil	-1.92	-1.92	dimethylphosphate	-3.41	-3.82
6-dimethylaminopurine	-0.2	-1.05	anabasin	1.46	0.82	2-naphthoyltrifluoroacetone	4.81	3.78
amitrole	-2.95	-3.0	<i>o</i> -phenanthroline hydrate	3.28	3.0	thiaminepropylidysulfide	1.78	1.56
8-quinolinol-2-methyl	2.39	3.22	α -bromopropionic acid	-0.44	-0.42	benzilic acid	1.64	1.08
oxazepam	2.07	1.33	phenethylamine- <i>N</i> - α -dimethyl	3.39	2.64	mevinphos	1.57	1.91
diazepam	4.93	4.45	prochlorperazine	5.54	5.13	2-acetyl-4,5,6,7-tetrachloro-1,3-indandione	4.28	5.05
2-piperidinoethanol	0.58	1.22	6-phenylcaproic acid	3.59	2.98	cyclohexane-1,3-dione-5,5-dimethyl	2.17	0.76
<i>N,N</i> -dimethyltryptamine-5-methoxy	1.92	0.52	1-nitroso-2-naphthol	1.84	3.0	anthranilic acid- <i>N</i> -methyl	-0.05	0.81
sulfamer	-0.55	0.02	3,3-dibenzo-1,8-crown-6-ether	5.10	3.90	sulfisoxazole	0.93	0.74
atenolol	1.4	-0.13	primidone	-0.07	-0.22	benzenesulfonic acid methyl ester	2.32	2.98
piperolylic acid	-0.46	0.7	sulfamethoxazole	0.43	0.56	cyclobarbitol	1.95	0.34
picolinic acid	-0.84	-1.64	amitriptyline	7.59	5.78	1-pyrroline-2-methyl	1.03	1.09
<i>p</i> -aminobenzoic acid butyl ester	3.44	2.38	<i>o</i> -chlorobenzenesulfonamide	0.57	0.46	6B,11B,17A,21-tetrahydroxy-4-pregene-3200-dione	0.3	-1.3
succinodinitrile	0.36	-0.23	phentermine	3.23	3.66	sotalol	0.42	-1.24
4-aminoantipyrine	1.33	1.17	butyronitrile-2,4-chlorophenyl-3-methyl	5.1	3.77	sulfamethoxy pyridazine	-0.52	0.47
phosphorodithiotic acid diisopropyl	2.52	2.71	chloroxine	3.75	3.86	medazepam	5.33	4.79
xanthine-3-methyl	-2.09	-1.43	α,α -aminobutyric acid- <i>N</i> -acetyl-D,L	-1.09	-1.6	sebacic acid	0.72	0.04
methaqualone	4.75	1.12	1,3-diphenyl-2-thiourea	1.63	2.76	acrylonitrile-3-dimethylamino	0.6	1.38
<i>N</i> -ethyl-2,3-dioxopiperazine	-1.72	-1.7	di- <i>p</i> -aminophenylsulfone	1.27	0.78	β -iodopropionic acid	0.4	-0.36
diethylphosphate	-2.17	-2.56	fluphenazine	4.9	5.51	monolinuron	1.94	2.68
3-hydroxypyridine	-1.25	-1.34	<i>N</i> -4-acetylsulfanilamide	-1.52	-1.52	levulinic acid	-0.72	-1.32
cytarabine	-4.19	-3.4	sulfapyridine	0.19	0.04	benzoylphenolhydroxylamine	3.53	2.33
5-nitrosalicylic acid	0.33	0.72	norephedrine	1.1	0.32	bicyclo-2,2,1-hepta-2,5-dien-7-ol benzoate	5.11	2.23
chloral hydrate	-1.95	-0.96	β -resorcylamide	-2.14	-1.32	pyridine-2-chloro-6-hydroxy	-0.71	-0.02
chlorimipramine	7.19	5.87	norepinephrine	-3.71	-2.15	3-sulfanilamido-6-chloropyridazine	-0.06	-1.23
<i>N</i> -maleoyl-3-amino-propionic acid	-0.65	-2.0	1,1-dimethyl-3,3,4-dichloro-phenylurea	2.34	2.54	clioquinol	4.4	3.9
phenethylamine- <i>N</i> -acetyl	2.18	1.41	α -bromobutyric acid	0.18	0.08			

^a $n = 107$, $r^2 = 0.864$, RMS = 0.908, av err -0.290, max dev -3.63 (methaqualone).

especially as again it has a reasonable log *P*(oct) prediction (obsd, -0.96; calcd, -0.06), and that a similar molecule in the set, sulfamethazine, which contains very similar functional groups, is predicted to within 0.9 log unit. This may suggest the measured value is in error; without this molecule the statistics improve to $r^2 = 0.819$ and RMS = 0.915.

In a fashion similar to that of eq 6, a modified solvation equation for log *P*(cyc) using calculated descriptors was developed. This time, however, the same set of data as used in the construction of eq 4 was used, to provide a direct comparison of experimental and calculated descriptors.

$$\log P(\text{cyc}) = 1.225R_2 - 1.602\pi_2^H - 3.165\Sigma\alpha_2^H - 4.853\Sigma\beta_2^H + 4.045(Vx) \quad (7)$$

$$n = 234, r^2 = 0.895, \text{sd} = 0.604, F = 387$$

This equation is very similar to eq 4, much closer than eq 6 is to eq 2. Significant differences (see Table 9 for standard deviations of coefficients) exist among the R_2 , $\Sigma\alpha_2^H$, and Vx terms, but these are still relatively small, and can be accounted for by errors in the descriptor calculations, and essentially the two equations are the same. This indicates that calculated descriptors may be used for the construction of new solvation equations, provided suitable training sets are available. Again, this new equation was tested on the set of log *P*(cyc) values in Table 8. In contrast to the log *P*(oct) values in Table 1, little improvement in the fit statistics results from this, with $r^2 = 0.785$ and RMS = 0.910 (after removal of sulfanilacetamide), with the slope and intercept no closer to the ideal values of 1.0 and 0.0. Thus, it seems the size and diversity of the training set are the major source of nonideal slopes and intercepts in the current work.

Table 8. Observed and Calculated log *P*(cyc) Values from MedChem97^a

name	log P(cyc)		name	log P(cyc)		name	log P(cyc)	
	calcd	obsd		calcd	obsd		calcd	obsd
1,3-indandione	-0.68	-0.12	1,4-naphthaquinone-2-methyl-3-methoxy	1.9	2.31	benzoquinone-2,6-dimethoxy	-0.74	-1.51
resorcinol-5-heptyl	-0.07	0.37	tetrahydrocarboline	-2.3	-0.5	durohydroquinone	-0.62	-0.25
<i>m</i> -methoxyaniline	-0.13	-0.13	2,6-dichloro-1,4-benzoquinone	0.99	0.88	5,6-dehydroisoandrosterone	1.55	1.63
2-trifluoromethyl-5-methylbenzimidazole	0.97	-0.48	5,6,7,8-tetrahydro-2-naphthol	0.87	0.88	fluphenazine	0.92	1.85
niacin	-3.09	-2.4	hydroquinone-2,6-dimethoxy	-1.52	-2.25	benzalcyanoacetamide- <i>N,N</i> -tetramethylene	1.17	0.7
1,4-naphthoquinone-2-methyl-3-bromo	3.06	2.85	hydroquinonemethoxy	-3.25	-2.63	diethylmalonatecinamal	4.15	2.67
β -3-methoxy-4-OH-phenylpropionic acid ethyl ester	1.23	0.86	indole-2,3-dimethyl	1.32	1.83	benzoquinone-2-methyl-6-bromo	1.5	1.1
2-OH-4,6-bisopropylamino- <i>s</i> -triazine	-1.49	-1.1	<i>p</i> -phenylenediamine	-2.11	-2.81	7-methyl-4-indanol	0.92	1.06
atrazine	1.23	0.81	<i>o</i> -phenylenediamine	-2.43	-1.65	sparteine	1.35	3.41
2-methoxy-4-Et-amino-6-isopropylamino- <i>s</i> -triazine	0.37	0.42	atropine	0.29	-1.02	androsterone acetate	4.05	3.33
indane-1,3-dione-2-benzal	2.02	3.4	ubiquinone-0	0.01	0.39	2-Me-thio-4,6-bisethylamino- <i>s</i> -triazine	0.69	0.99
<i>o</i> -aminobenzoic acid	-2.58	-2	methyl styrylketone	2.31	1.5	<i>N,N</i> -diethylcinnamamide	2.27	1.28
sulfamethazine	-3.9	-3	quinoline-3-amino	-0.63	1.28	2,2,6,6-tetramethyl-3,5-heptanedione	2.35	3.9
papaverine	3.2	2.56	ethylcyanoacetate-4-methylbenzal	3.37	3.62	DiMe-1-OH-2,2,2-tri-Cl-ethyl-phosphonate	-3.86	-1.7
<i>N</i> 2- <i>tert</i> -butyl- <i>N</i> 1-benzene-sulfonyl urea	-2.68	-0.64	malononitrile-3-methoxy-4-hydroxybenzal	-0.14	0.3	ethylcyanoacetate-2-chloro-benzal	3.16	2.97
chlorpropamide	-2.48	-0.74	malononitrile-4-methoxybenzal	1.86	1.46	2-methoxy-4,3-oxobutylphenol	0.36	-0.21
malononitrile-2-fluorobenzal	2.22	1.55	2-methyl-5-isopropylphenol	1.36	1.3	diisopropylphosphate	-4.13	-2.92
malononitrile- α -phenylbenzal	4.63	3.52	naphthalene-2,6-diacetoxy	2.38	1.75	benzaldehyde-3-ethoxy-4-hydroxy	-1.85	0.03
isopropylcinnamate	3.84	3.19	naphthalene-1,5-diacetoxy	2.52	1.61	4-chlorophenacylchloride	2.23	1.89
diethylmalonatebenzal	3.2	3.26	α -cyanobenzalacetophenone	3.47	3.15	naphthalene-2,7-diol	-3.53	-2.63
pyran-2,3-dihydro	1.3	1.73	betahistine	-0.59	-1.08	cyanoacetamide-3,4-dimethoxybenzal	-2.41	-1.63
1,4-naphthoquinone-2-chloro-3-dimethylamino	0.31	1.24	anabesine	-0.74	-0.58	1,1,1-trifluorobenzoylacetone	1.24	2.28
cinnamic acid methyl ester	2.58	2.44	cotinine	-0.99	-3.13	α -bromoacetophenone- <i>p</i> -chloro	2.53	2.5
warfarin	0.21	-0.74	coumarin-3-carboxylic acid ethyl ester	1.13	0.45	benzyl diethylmalonate	3.46	3.15
ethylcyanoacetate-4-methoxybenzal	2.29	2.47	pinacolone	1.5	1.12	2,2-epoxy-1,4-naphthaquinone	-0.32	0.8
ethylcyanoacetate-2,4-dimethoxybenzal	2.2	2.5	cyanoacetamide-4-methoxybenzal	-1.97	-1.1	α -bromoacetophenone- <i>p</i> -bromo	2.83	2.64
cinnamanilide	1.37	1.13	cyanoacetamide- <i>o</i> -chlorobenzal	-1.12	-0.56	sulfanilacetamide	-5.62	-2.00
5-bromosalicylic acid	-0.57	0.53	benzalacetophenone	4.05	3.94	diphenhydramine	3.75	3.48
ethyl- β -benzoylacrylate	2.53	2.38	testosterone propionate	4.42	4.38	phenylacetone-nitrile- <i>m</i> -phenoxy	2.43	2.33
benzaldehyde-3,4-dimethoxy	-0.15	0.32	prochlorperazine	2.15	2.5	diethylmalonate-4-methoxybenzal	2.84	2.98
coumarin	1.14	0.48	2-nitroso-1-naphthol	-0.02	0.3	<i>p</i> -methoxy- <i>N</i> -methylaniline	0.16	0.52
2-cyclohexylphenol	2.09	3.38	diethylmalonate-4-chlorobenzal	3.7	3.52	1,4-naphthoquinone-2-anilino	1.28	2.63
<i>N</i> -methylpyrrolidine	-0.38	0.42	carbazole	1.3	2.24	dimethylmalonatebenzal	1.89	1.93
3,5-dimethoxycinnamic acid ethyl ester	2.75	3.44	pyrantetrahydro-2-methoxy	1.21	1.05	<i>N</i> -allylnormorphine	-0.7	-1.7
3-methoxy-4-hydroxycinnamic acid methyl ester	0.22	0.54	2-methyl-4,6-dinitrophenol	0.66	-0.61	diethylmalonate-3,4-dimethoxybenzal	2.45	2.33
5-nitrosalicylic acid	-1.95	-0.36	nortriptyline	5.69	3.91	5-pyrazolone-1-phenyl-3-methyl-4-benzoyl	-0.16	2.14
phenylbutazone	4.22	2.96	<i>p</i> -nitrosophenol	-2.85	-1.1	4,6-nonanedione	1.34	2.36
phenethylamine- <i>N</i> -acetyl	-0.14	-1.01	<i>p</i> -methoxycinnamic acid methyl ester	2.21	2.35	ethylacetate-4-methylbenzal	3.03	2.35
quinine	0.27	0.69	testosterone-17 α -methyl	2.13	1.81	ethyl- α -benzoyl- <i>p</i> -methoxycinnamate	3.66	3.56
naphthalene-2,3-diol	-1.91	-1.6	butyronitrile-3-methyl-2-phenyl	2.97	2.26	α -cyano- <i>p</i> -methoxybenzalacetophenone	3.12	3.12
naphthalene-2,6-diol	-3.53	-2.73	butyronitrile-2,4-chlorophenol-3-methyl	3.47	3.01	α -bromoacetophenone-3-methoxy	1.71	1.97
2,3,5-trimethyl-1,4-benzoquinone	2.14	1.58	chloroxine	2.24	0.3	α -chloroacetophenone- <i>p</i> -bromo	2.53	1.94
2,6-dimethyl-1,4-benzoquinone	1.44	0.99	2-methyl-1,4-diacetoxy-naphthalene	3.32	2.06	nitrofurazone	-4.1	-2.22
1,4-naphthaquinone-2-methyl-3-hydroxy	0.48	0.79	coumarin-3-cyano	0.57	-0.56	<i>N,N</i> -CH ₂ -4,3,4-methylenedioxybenzalcyanoacetamide	-0.78	0.87

^a $n = 132$, $r^2 = 0.813$, RMS = 0.968, av err 0.139, max dev 3.62 (sulfanilacetamide).

Table 9: Coefficients and Standard Deviations for Eqs 4 and 7

	eq 4		eq 7	
	coeff	sd	coeff	sd
<i>c</i>	0.159	0.032	0.038	0.117
<i>r</i>	0.784	0.037	1.225	0.165
<i>s</i>	-1.678	0.040	-1.602	0.215
<i>a</i>	-3.740	0.037	-3.165	0.151
<i>b</i>	-4.929	0.049	-4.853	0.244
<i>v</i>	4.577	0.042	4.045	0.152

DISCUSSION

The main point to be drawn from this study is that our group contribution method for the estimation of LFER descriptors, as described in paper 1, is capable of reproducing experimental measurements on water–solvent partition coefficients. For two of the sets of data reported here, $\log P(\text{oct})$ values were predicted with r^2 values of 0.88 and 0.79 and RMS errors of 1.0 and 0.92. This represents the data rather better than some group contribution schemes previously published, such as Rekker's or Klopman's, though rather worse than more sophisticated methods such as CLOGP or ALOGP.

As a method for the estimation of $\log P(\text{oct})$, ours simply adds to the plethora of published methods, and little would be gained if no other properties could be estimated. We have shown, however, that the same calculation of descriptors is capable of reproducing $\log P(\text{chl})$ and $\log P(\text{cyc})$ with accuracy similar to that of $\log P(\text{oct})$, despite their dependence on $\Sigma\alpha_2^H$ as well as π_2^H and $\Sigma\beta_2^H$. This is a most encouraging result, as we believe this is the first time such properties have been predicted by such an approach. The ability to predict these properties directly from structure is in itself important, as they are often used as model properties for biological or environmental properties.

Moreover, this suggests that any transport process modeled by the solvation eq 1 can be predicted by our method, especially since the coefficients of eq 4 for $\log P(\text{cyc})$ are as large as for any process studied, reflecting the large difference in properties between bulk water and cyclohexane. Processes for which the coefficients of eq 1 have been established include a wide range of solvent–solvent partitions, vapor solubilities in various solvents, chromatographic retention parameters, and biological transport such as blood–brain distribution, brain perfusion, and skin permeation. A modified version of eq 1 has also been developed for general water solubility,²¹ including that of solids, opening up the possibility of prediction of $\log Sw$ from structure. Applications of the calculation method to such processes will follow in a subsequent study.

Two of the data sets, $\log P(\text{oct})$ and $\log P(\text{cyc})$, have been subjected to a new LFER analysis, using calculated rather than experimental descriptors as the independent variables. In both cases the new equation was at least qualitatively similar to the experimental equations, though substantial differences in the values of the coefficients were observed. This demonstrates that, for a reliable data set, the calculations reported in paper 1 are capable of generating a reliable, physically meaningful solvation equation. It should therefore prove possible to treat more complex processes, including important biochemical ones, in the same manner. However, where an experimentally derived equation for a given process

is available, no improvement in the fit is observed when a new equation based on the calculated descriptors is employed.

An important aspect of the calculations reported here is their speed. In paper 1, we estimated that the group contribution method could calculate descriptors for approximately 500 molecules per minute. However, the molecules in question there were somewhat smaller than those in Tables 1–4, which can be taken as more like a “real” application of the method. For the largest set of molecules, the 140 $\log P(\text{oct})$ values from Bodor's review, descriptors were calculated in 21 s on an SGI O2,²² equivalent to 400 per minute or around 6×10^5 per day. This is certainly sufficient for application in high throughput screening, wherein a library of 10 thousand molecules of a similar size could be treated in just 25 min. It should be emphasized that the arithmetical step to obtain a solvation property from the descriptor calculation is trivial, and that a single descriptor calculation is sufficient to predict a wide range of solvation properties.

Some shortcomings in our present model should be pointed out. Several of the largest errors are found for molecules containing cyclic sulfonamide fragments, such as piroxicam and chlorothiazide. In contrast, noncyclic sulfonamide groups, as found in sulfanilamide for example, are predicted very well. Our method, along with most fragmental models, treats these cyclic and noncyclic fragments as being identical, which is clearly not the case. Also, no data are available on descriptor values for charged species, notably zwitterions: all predictions are for the neutral form of a molecule. These are essentially shortcomings in the training set used in paper 1, which did not include such groups. We will, in a later study, develop descriptors for these and other groups, and incorporate them into the calculation of descriptors.

CONCLUSIONS

We have demonstrated that the calculation of LFER descriptors described in paper 1 is of sufficient accuracy to reproduce experimental partition data. Two sets of $\log P(\text{oct})$ values, and one each of $\log P(\text{chl})$ and $\log P(\text{cyc})$, 528 values in total, are predicted with a typical RMS error of between 0.90 and 1.0 log unit. This is an accuracy similar to that previously reported for similar prediction methods for $\log P(\text{oct})$, but represents the first application of such an approach to other water–solvent partitions. Further, we have demonstrated that the descriptors used can correlate almost 9000 starred $\log P(\text{oct})$ values with an RMS error of 0.71, with the model taking a form similar to those of previously published models of $\log P(\text{oct})$. A new solvation equation for $\log P(\text{cyc})$ was also developed, and found to be both statistically valid and quantitatively similar to a previously published equation, indicating the possibility of developing equations for new processes. However, no improvement was observed in the prediction of the above test sets with the new equations, suggesting that experimental equations, where available, should be preferred.

ACKNOWLEDGMENT

J.A.P. is grateful to Glaxo Wellcome for a postdoctoral fellowship.

Supporting Information Available: Tables A1–A4 of descriptors used in Bodor logP(oct) set, in MedChem97 logP(oct) set, in logP(chl) set, and in logP(cyc) set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular LFER Descriptors using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835.
- Abraham, M. H. Scales of Solute Hydrogen Bonding – their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Soc. Rev.* **1993**, *22*, 73–83.
- Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid Chromatography. *Chromatographia* **1987**, *23*, 243–246.
- Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen Bonding. 16. A New Solute Solvation Parameter, π_2^H , from Gas-Chromatographic Data. *J. Chromatogr.* **1991**, *587*, 213–228.
- (a) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.* **1994**, *83*, 1257–1268. (b) Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. Hydrogen Bonding. 32. An Analysis of water-octanol and water-alkane partitioning, and the $\Delta\log P$ parameter of Seiler. *J. Pharm. Sci.* **1994**, *83*, 1085–1100. (c) Abraham, M. H.; Chadha, H. In *Lipophilicity in Drug Action and Toxicology*; Pliska, V., Testa, B., van der Waterbeemd, H., Eds.; VCH: Weinheim, 1996.
- See for example: (a) Hansch, C. Quantitative Structure–Activity Relations and the Unnamed Science. *Acc. Chem. Res.* **1993**, *20*, 147. (b) Pliska, V., Testa, B., van der Waterbeemd, H., Eds. *Lipophilicity in Drug Action and Toxicology*; VCH: Weinheim, 1996.
- Abraham, M. H.; Platts, J. A.; Hersey, A.; Leo, A. J.; Taft, R. W. The Correlation and Estimation of Gas-Chloroform and Water-Chloroform Partition Coefficients by an LFER Method. *J. Pharm. Sci.* **1999**, *88*, 670.
- Leahy, D. E.; Morris, J. J.; Taylor, P. J.; Wait, A. R. Model Solvent Systems for QSAR. Part 3. An LSER Analysis of the “Critical Quartet”. New Light on Hydrogen Bond Strength and Directionality. *J. Chem. Soc., Perkin Trans. 2* **1992**, 705–722.
- Rekker, R. F. *The Hydrophobic Fragmental Constant*; Elsevier: New York, 1977.
- Leo, A. J. Calculating logP(oct) from Structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
- Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated logP Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.
- (a) Ghose, A. K.; Viswanadhan, V. N.; 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. A* **1998**, *102*, 3762–3772. (b) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure Directed Quantitative Structure–Activity Relationships. 1. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.
- Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.
- Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. OMNISOL: Fast Prediction of Free Energies of Solvation and Partition Coefficients. *J. Org. Chem.* **1998**, *63*, 4305–4313.
- Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semi-Analytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- Collander, R. On “Lipoid Solubility”. *Acta Physiol. Scand.* **1947**, *13*, 363–381.
- Buchwald, P. and Bodor, N. Octanol–Water Partition: Searching for Predictive Models. *Curr. Med. Chem.* **1998**, *5*, 353–380.
- Schüürman, G.; Kühne, R.; Ebert, R.-U.; Kleint, F. Multivariate Error Analysis of Increment Methods for Calculating the Octanol-Water Partition Coefficient. *Fresenius Environ. Bull.* **1995**, *4*, 13–18.
- Leo, A. J. Masterfile 1997 from MedChem software, BioByte Corp., P.O. Box 517, Claremont, CA 91711-0157.
- ClogP for Windows, v2.0, Biobyte Corp, P.O. Box 517, Claremont, CA 91711-0157.
- (a) Taft, R. W.; Abraham, M. H.; Famini, G. R.; Doherty, R. M.; Abboud, J.-L. M.; Kamlet, M. J. Solubility Properties in Polymers and Biological Media 5: An Analysis of the Physicochemical Properties which Influence Octanol-Water Partition Coefficients of Aliphatic and Aromatic Solutes. *J. Pharm. Sci.* **1985**, *74*, 807–814. (b) Leahy, D. E. Intrinsic Molecular Volume as a Measure of the Cavity Term in Linear Solvation Energy Analysis: Octanol-Water Partition Coefficients and Aqueous Solubilities. *J. Pharm. Sci.* **1991**, *80*, 590–598. (c) Steyaert, G.; Lisa, G.; Gaillard, P.; Boss, G.; Reymond, F.; Girault, H. H.; Carrupt, P.-A.; Testa, B. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 401–406.
- Le, J.; Abraham, M. H. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *89*, 868–880.
- Full details: single 180 MHz R5000 processor, with 96 Mb of RAM and 512 K cache.

CI990427T