

ClogP_{alk}: a method for predicting alkane/water partition coefficient

Peter W. Kenny · Carlos A. Montanari ·
Igor M. Prokopczyk

Received: 18 March 2013 / Accepted: 18 May 2013 / Published online: 5 June 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Alkane/water partition coefficients (P_{alk}) are less familiar to the molecular design community than their 1-octanol/water equivalents and access to both data and prediction tools is much more limited. A method for predicting alkane/water partition coefficient from molecular structure is introduced. The basis for the ClogP_{alk} model is the strong ($R^2 = 0.987$) relationship between alkane/water partition coefficient and molecular surface area (MSA) that was observed for saturated hydrocarbons. The model treats a molecule as a perturbation of a saturated hydrocarbon molecule with the same MSA and uses increments defined for functional groups to quantify the extent to which $\log P_{\text{alk}}$ is perturbed by the introduction each functional group. Interactions between functional groups, such as intramolecular hydrogen bonds are also parameterized within a perturbation framework. The functional groups and interactions between them are specified substructurally in a transparent and reproducible manner using SMARTS notation. The ClogP_{alk} model was parameterized using data measured for structurally prototypical compounds that dominate the literature on alkane/water partition coefficients and then validated using an external test set of 100 alkane/water $\log P$ measurements, the majority of which were for drugs.

Keywords Alkane/water · ClogP_{alk} · Ligand efficiency · Lipophilicity · $\log P$ · $\log P_{\text{alk}}$ · Molecular surface area · Partition coefficient · SMARTS · Solvation

Introduction

Lipophilicity, usually quantified as the logarithm (to base 10) of the 1-octanol/water partition coefficient (P_{oct}) or distribution coefficient, (D_{oct}) is the most widely used physicochemical property in drug discovery and the key molecular design parameter in medicinal chemistry [1, 2]. Lipophilicity has been of interest in the physiological context for many years and it was in 1847 that von Bibra and Harless [3] suggested that the narcotic effects of *Schwefeläther* might be due to its affinity for brain lipids. The essential features of partitioning between liquid phases had already been articulated by Nernst [4] when Meyer [5] and Overton [6] published their studies relating anesthetic potency to solubility in olive oil. The 1-octanol/water partitioning system appears to have first been used by Collander [7] who also measured the partitioning of compounds between water and oleyl alcohol, a solvent recommended by the younger Meyer [8] for studies of this nature. Collander certainly recognized the influence of hydrogen bonding on partitioning and noted the existence of linear relationships between partition coefficients measured for different monohydric alcohols and water [7]. The adoption of the 1-octanol/water system for partitioning studies in drug discovery can be traced to a 1963 article by Hansch et al. [9]. With the benefit of hindsight, the original selection of 1-octanol/water as the favored partitioning system does appear arbitrary and it was also incorrectly stated that Collander had observed linear correlations between $\log P$ for systems other than alcohols and water.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9655-5) contains supplementary material, which is available to authorized users.

P. W. Kenny (✉) · C. A. Montanari · I. M. Prokopczyk
Grupo de Estudos em Química Medicinal, NEQUIMED,
Instituto de Química de São Carlos, Universidade de São Paulo,
Av. Trabalhador Sancarlense, 400, São Carlos, SP 13560-590,
Brazil
e-mail: pwk.pub.2008@gmail.com

This is not to deny that $\log P_{\text{oct}}$ has proved a useful parameter in drug discovery or that the concept of lipophilicity has shaped the thinking of medicinal chemists.

The main uses of $\log P_{\text{oct}}$ in contemporary drug discovery are for modeling the association of compounds with cell membranes that is a necessary, but not sufficient, condition for passive diffusion through membranes and as a general purpose descriptor of desolvation for modeling affinity and properties, such as aqueous solubility, that are relevant to pharmaceutical design [10–14]. A number of researchers have questioned [15–21] the suitability of 1-octanol/water partition coefficients for these purposes since 1-octanol is capable of forming hydrogen bonds with polar solutes. It can be argued that this capacity for hydrogen bonding makes 1-octanol a poor model for the hydrocarbon-like interior of a lipid bilayer or for polar-apolar contacts that can be either intermolecular (drug-target complex) or intramolecular (folded protein). It is noteworthy that while the lower polarity limit for the Rule of 5 [1] is defined in terms of lipophilicity, the upper for polarity is specified by numbers of hydrogen bonding groups. Another characteristic of the 1-octanol/water system that compromises its general applicability as a descriptor of aqueous desolvation is its relative inability to sense hydrogen bond donors in solute molecules [22]. In the context of direct measurement of partition coefficients, 1-octanol is also much wetter (2.5 M water) [23] than hydrocarbon solvents such as hexadecane (2×10^{-3} M water) [21] or cyclohexane (1.5×10^{-3} M water) [24]. Water in the organic phase stabilizes polar solutes and makes them appear more lipophilic than if the partition coefficients had been determined by the solubility ratio [26]. Self-association (e.g. for lactams) [27] in the organic phase has an analogous effect and can present difficulties when solubility ratio is used to measure partition coefficients since the experiment is performed at saturation and solute concentration cannot be varied.

Alkane/water partitioning systems also have a long history and the insights of Golumbic et al. [28] into steric effects on hydrogen bonding by phenols, gained from measuring cyclohexane/water partition coefficients (P_{chx}), actually pre-dated the Collander study [7]. More recently, $\log P_{\text{chx}}$ measurements have been used to investigate the effects of intramolecular hydrogen bonding and steric control of conformational preferences [29]. The alkane/water partition coefficient (P_{alk}) of a compound is typically lower than P_{oct} and the difference ($\Delta \log P$) between the logarithms of the two quantities reflects the potential of the compound to make electrostatic interactions, such as hydrogen bonds, with polar solvents [30]. Measured values of $\log P_{\text{alk}}$ differ in the choice of alkane (e.g. cyclohexane, hexadecane), experimental conditions (e.g. temperature) and whether or not the organic and aqueous phases

are mutually saturated when measurement is made. There is no single consistent data set that is sufficiently large and structurally diverse for developing predictive models for $\log P_{\text{alk}}$. It has been suggested [31] that $\log P$ values measured for alkanes other than cyclohexane do not differ significantly from each other and that these can be combined for analysis. Although statistically significant differences have been observed [30, 31] between $\log P$ values measured for cyclohexane and other alkanes, the differences are small and it is not clear how widely applicable the published regression equations are for inter-conversion. In the current study, no attempt was made to convert $\log P_{\text{cyc}}$ values to their equivalents in other alkane/water solvent systems and the term $\log P_{\text{alk}}$ will be taken to mean any partition coefficient that has been measured in a saturated hydrocarbon/water partitioning system.

While a number of approaches to predicting $\log P_{\text{alk}}$ have been reported, practical access to predictive tools is limited. The most extensive body of work is that of Abraham et al. who use a general solvation equation to model partition coefficients measured in a range of solvent systems [22, 24, 31]. Using the general solvation equation to predict alkane/water appears [31] to require some experimental measurements (e.g. $\log P_{\text{oct}}$) for compounds of interest and it is not clear how reliably or generally the approach can be used to make predictions from molecular structure alone. Molecular surface area [33] (MSA) and volume [34] have been used with atom-typing to model $\log P_{\text{alk}}$, as have molecular electrostatic properties [35], molecular interaction fields [36] and implicit solvent models [37]. The alkane/water partition coefficient has also been treated as a perturbation of its 1-octanol/water equivalent by making $\Delta \log P$ the target for predictive modeling [21]. The $\text{Clog} P_{\text{alk}}$ model introduced in the current study also treats $\log P_{\text{alk}}$ as a perturbation although, in this case, the reference state is a saturated hydrocarbon molecule with the same MSA as the molecule for which prediction is being made. A rationale for this approach is provided by the strong correlation of $\log P_{\text{alk}}$ with MSA that is observed for saturated hydrocarbons. The perturbation, which can be regarded as a measure of molecular polarity, is quantified for functional groups using atom types and it is the development and parameterization of this model that forms the basis for this work.

Computational details

The molecular modeling software described in this article was either provided by OpenEye Scientific Software [38] or built using OpenEye programming toolkits. Molecular structures were encoded as isomeric SMILES [39, 40] strings and Omega [41, 42] was used to generate a single

Table 1 Atom types defined as vector bindings

Name	SMARTS	Description
Csp3	[CX4]	sp ³ carbon
Cnotsp3	[#6;!\$Csp3]	Unsaturated carbon
CDoubBond	c=[C,N,O,S]	Doubly bonded 'aromatic' carbon
ArC	[c;!\$CDoubBond]	Normal aromatic carbon
ArN	[n&D2]	Pyridine-like aromatic nitrogen
ArCN	[\$ArC,\$ArN]	Aromatic carbon or nitrogen
ArNH	[nH]1[\$ArCN][\$ArCN][\$ArCN][\$ArCN]1	Pyrrole-like aromatic nitrogen
ArNR	n1(#6)[\$ArCN][\$ArCN][\$ArCN][\$ArCN]1	Substituted pyrrole-like nitrogen
ArO	o1[\$ArCN][\$ArCN][\$ArCN][\$ArCN]1	Aromatic oxygen
ArS	s1[\$ArCN][\$ArCN][\$ArCN][\$ArCN]1	Aromatic sulfur
OS	[O,S;X1]	Doubly-bonded oxygen or sulfur
NH2	[N;H2;X3]	Amino
NHR	[N;H1;X3][\$Csp3]	Alkylamino
NR2	[NX3](\$Csp3)[\$Csp3]	Dialkylamino
GenAmino	[\$NH2,\$NHR,\$NR2]	Amino and alkylated analogues
PrimAmine	[\$NH2][\$Csp3]	Primary amine
SecAmine	[\$NHR](\$Csp3)[\$Csp3]	Secondary amine
TertAmine	[\$NR2](\$Csp3)(\$Csp3)[\$Csp3]	Tertiary amine
Amine	[\$PrimAmine,\$SecAmine,\$TertAmine]	Amine
Ether	O(\$Csp3)[\$Csp3]	Ether
Nitro1	N(=O)=O	Nitro with pentavalent nitrogen
Nitro2	[N+](O-)=O	Ylid nitro
Nitro	[\$Nitro1,\$Nitro2]	Nitro group
ArNoxid1	[nX3]=O	Aromatic N-oxide with pentavalent nitrogen
ArNoxid2	[n;+;X3][O-]	Ylid aromatic N-oxide
ArNoxid	[\$ArNoxid1,\$ArNoxid2]	Aromatic N-oxide
AlNoxid1	[NX4]=O	Aliphatic N-oxide with pentavalent nitrogen
AlNoxid2	[N;X4;+][O-]	Ylid aliphatic N-oxide
AlNoxid	[\$AlNoxid1,\$AlNoxid2]	Aliphatic N-oxide
SO2	[SX4](=O)(=O)(\$Csp3)[\$Csp3,N&X3]	Sulfones and sulfonamides
EWG1	[C,S]=[N,O,S]	Electron-withdrawing group for defining imides and related substructures
HBD	[O&H,n&H,\$([NH][C,S]=O),\$([N;H,H2]c)]	HB donor
HBD2	[\$HBD;!\$(N1CC(=O)AC1=O)]	Restrict HB donor definition
Am4Pyr	[\$ArN]1[\$ArCN][\$ArCN]c(\$GenAmino)[\$ArC][\$ArC]1	4-Aminopyridines and aza-substituted
BzCorr	[\$Amine,O&!H][\$Csp3][\$Cnotsp3]	For applying the benzyl correction
CyAmine	CyAmine \$(C1CNCC1),\$ (C1CNCCC1),\$ (C1CCNCC1)	
NOCCNO	[\$Amine,\$Ether][\$Csp3][\$Csp3][N,\$Ether]	Interactions through bonds (2)
NOCNO	[\$Amine,O&!H][\$Csp3][N,O]	Interactions through bonds (1)

conformation for each. Molecular geometries were energy-minimized (MMFF94S) [43] using the Szybki [44] molecular mechanics program. Omega failed to generate 3D coordinates for sulfur hexafluoride and the molecular model for this compound was built with octahedral symmetry using the experimentally determined bond length [45]. Molecular surface area was calculated from atomic coordinates and Bondi [46] radii using a probe radius of 1.4 Å. Molecular similarity was calculated as Tanimoto

coefficient using path-based fingerprints (1,024 bits; paths up to 5 bonds) using the GraphSim toolkit [47] The JMP software [48] was used for all data analysis.

Two pieces of software were created in the course of this study. SSProFilter was built with the OEChem [49] toolkit and combines the functionality of the previously described Filter [50] program (which should not be confused with the OpenEye product of the same name) with the ability to profile molecular structures by counting matches with

substructural patterns encoded as SMARTS [51]. The SSProfiler program was used to identify molecular structures containing only the atom types of interest for deriving parameters. The ClogP_{alk} program that is the focus of this article was built using the OEChem [49] (SMARTS matching; structure handling) and OESpicoli [52] (MSA calculation) toolkits. Measured logP_{alk} values were obtained from a number of literature sources [16, 17, 21, 22, 24, 25, 29–32, 36, 53–65] and the data set is provided with structures (in isomeric SMILES format) as supplemental material.

The ClogP_{alk} predictive model is defined by Eq. 1:

$$C \log P_{alk} = \log P_0 + s \times MSA - \sum_i \Delta \log P_{FG,i} - \sum_j \Delta \log P_{Int,j} \quad (1)$$

The first two terms of Eq. 1 relate the logP_{alk} values measured for saturated hydrocarbons to MSA and describe reference state of the model. The last two terms quantify the extent to which this saturated hydrocarbon reference state is perturbed by the introduction of non-carbon atoms and unsaturation into the molecular structure. Each functional group ($\Delta \log P_{FG,i}$) and interaction ($\Delta \log P_{Int,j}$) term is associated with a substructural target defined using the OpenEye implementation of SMARTS notation. In both cases vector bindings (e.g. \$Csp3 ≡ [CX4]) were used to make the patterns more human-readable and the vector bindings used by the ClogP_{alk} model are listed in Table 1. It can be helpful to think of the functional group and interaction terms in Eq. 1 as first and second order perturbations respectively.

Each functional group term is calculated, using Eq. 2 from the number of matches (N_i) with the relevant SMARTS target. In the context of the model a functional group is simply a substructure, which can even be a single atom (e.g. aromatic carbon), with which a perturbation is associated. It is important to note that Eq. 2, which has an intercept term ($k_{FG,i}$), is only applied if the relevant functional group actually occurs in the molecular structure for which prediction is being made and this intercept term was only rarely used (e.g. for the aromatic carbon atom type).

$$\Delta \log P_{FG,i} = f_{FG,i} \times N_i + k_{FG,i} \quad (2)$$

Atoms are typed when the functional group terms are calculated which means that each atom is linked to, at most, one SMARTS pattern when Eq. 2 is applied. For multi-atom SMARTS patterns, only the atom mapped by the first atom (the head atom) of the SMARTS string is typed and the contribution of the entire functional group (e.g. sulfonamide) is loaded onto the head atom (e.g. sulfur). The SMARTS patterns are matched in the order in which they occur in the parameter file and each atom is

linked to the last SMARTS target with which it matches. This facilitates customization of the model because a generic parameter can simply be over-written by a more specific one that might, for example, have been derived using proprietary data. Each interaction term is calculated, using Eq. 3 from the number of matches (N_j) with the relevant SMARTS target and, in contrast with the functional group terms, an atom can be associated with more than one interaction term.

$$\Delta \log P_{Int,j} = f_{Int,j} \times N_j \quad (3)$$

The ClogP_{alk} model is perturbation-based and its parameterization can be described as ordered in that parameters were not derived simultaneously. The first step in deriving a functional group parameter is to assemble logP_{alk} values measured for compounds in which only hydrogen, saturated carbon and the relevant functional group are present in the molecular structure. This creates a training set specific to the relevant functional group and parameters are derived independently of each other using Eq. 2 (the intercept term $k_{FG,i}$ is only rarely used). The interaction parameters (Eq. 3) were set in the final step of the process. The model was validated using an external test set of 100 measured logP_{alk} values.

Alkane/water partition coefficients for saturated hydrocarbons

The basis for the ClogP_{alk} predictive model is the strong, linear relationship between logP_{alk} and MSA observed for a set of 48 saturated hydrocarbons that included linear, branched and cyclic species (Fig. 1; Eq. 4).

$$\log P_{alk} = 0.0340 \times MSA/\text{\AA}^2 - 0.309 \quad (4)$$

$$N = 48; R^2 = 0.987; RMSE = 0.152.$$

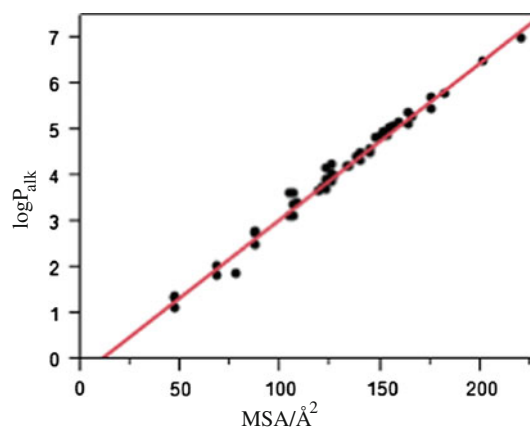


Fig. 1 Relationship between logP_{alk} and calculated molecular surface area observed for saturated hydrocarbons

Table 2 Functional group contributions for unsaturated carbon and halogen-substituted aromatic carbon

Atom type	SMARTS ^a	$f_{FG,i}^b$	$k_{FG,i}^b$	N ^c	RMSE ^d
Aromatic C	[\$Car]	0.0526	0.965	47	0.17
Alkene C	C=C	0.16	0.26	18	
Alkyne C	C#C	0.74	Not used	8	
Aromatic N (HB acceptor)	[\$Nar]	2.50 ^e	Not used	30	0.26
Aromatic N (HB donor)	[\$ArNH]	2.16 ^e	Not used	9	0.16
Aromatic N (with substituent)	[\$ArNR]	1.07 ^e	Not used	3	0.04
Aromatic N-oxide	[\$ArNoxid]	6.09 ^e	Not used	3	0.5
Aromatic oxygen	[\$OAr]	0.17 ^{e,f}	Not used		
Aromatic sulfur	[\$SAr]	0.10 ^e	Not used	2	0.10
Aromatic F	Fc	-0.13 ^e	Not used	2	0.14
Aromatic Cl	Clc	-0.15 ^e	-0.23	24	0.16
Aromatic Br	Brc	-0.20 ^e	-0.11	4	0.09
Aromatic I	Ic	-0.38 ^e	Not used	2	0.08
C with hydrogen and F	[C;X4;!H0]F	0.60 ^g	0.66 ^g		
C with hydrogen and Cl	[C;X4;!H0]Cl	0.60	0.66	24	0.15
C with hydrogen and Br	[C;X4;!H0]Br	0.60	0.49	13	0.10
C with hydrogen and I	[C;X4;!H0]I	0.83	Not used	8	0.11

^a See Table 1 for vector binding definition

^b Equation 2

^c Number of data points used to determine parameters

^d Root mean square error from parameter fitting

^e Parameter values dependent on aromatic carbon parameters

^f Parameter derived using assumption that $\log P_{alk} = \log P_{oct}$ for furan and benzofuran

^g Parameters set to chlorine equivalents

Derivation of functional group parameters

The $\text{Clog}P_{alk}$ model is perturbation-based and its parameterization can be described as ordered in that parameters were not derived simultaneously. Functional group parameters can be derived directly when the only atom types in the molecular structure are those that define the functional group, saturated carbon and hydrogen. In these situations, values of $\Delta\log P_{FG,i}$, obtained by subtracting measured $\log P_{alk}$ from the value calculated from Eq. 4, can be fit using Eq. 2. In other cases, functional group X (e.g. aromatic nitrogen) cannot be found except when another functional group Y (e.g. aromatic carbon) is also present. If parameters have already been derived for functional group Y then these can be used to adjust the predicted (Eq. 4) value of $\log P_{alk}$ prior to calculating $\Delta\log P_{FG,i}$ and fitting Eq. 2. The first step of the parameterization of the $\text{Clog}P_{alk}$ model was to investigate the need for aromatic carbon parameters since this atom type is almost ubiquitous in compounds of pharmaceutical interest. When using SMARTS notation, certain carbons (e.g. pyridone carbonyl) are considered to be aromatic and a more restrictive aromatic carbon atom type was defined using vector bindings to prevent matches with these. A number of atom

types, including the re-defined aromatic carbon [\$Car], are listed in Table 1, and these were used to build SMARTS definitions and make them more easily read by humans. Parameters for alkene, alkyne and aromatic atoms are listed in Table 2 and it should be noted that intercept terms ($k_{FG,i}$) were used for alkene and aromatic carbon but not alkyne carbon. Intercept terms were also used for chlorine and bromine bonded to either aromatic or saturated carbon. Parameters for these atom types are included in Table 2 for convenience because intercept terms are not used for any other functional groups. The value of the root mean square error (RMSE) provides an indication of the internal consistency of the data used to derive parameters for the relevant functional group.

It was possible to derive parameters for a number of functional groups that were not dependent on the values of the parameters for unsaturated carbon atom types and these are listed in Table 3. Parameters for urea and thiourea were observed to be similar and SMARTS definitions for amides were generalized to match thioamides although the measured data corresponded only to amides. Parameters for imines were generated by taking the mean of the primary amine and nitrile parameters. Parameters for functional groups that were dependent on the parameter values for

Table 3 Functional group contributions

Description	SMARTS ^a	$f_{G,i}^b$	N ^c	RMSE ^d
Primary amine	[N;H2;X3][Csp3]	4.23	20	0.15
Secondary amine	[N;H;X3]([Csp3])[Csp3]	4.17	10	0.24
Tertiary amine	[NX3]([Csp3])([Csp3])[Csp3]	3.81	7	0.22
Nitrile	C(#[NX1])[Csp3]	3.39	5	0.15
Guanidine	C(=N)([GenAmino])[GenAmino]	8.73	1	
Alcohol	[OH][Csp3]	4.48	56	0.16
Ether	O([Csp3])[Csp3]	2.96	10	0.20
Aldehyde	[CH](=[OS])[Csp3]	3.38	9	0.08
Ketone	C(=[OS])([Csp3])[Csp3]	3.72	26	0.09
Ester	C(=[OS])([Csp3])O([Csp3])	3.52	37	0.19
Carboxylic acid	O=C([OH])[Csp3]	5.52	19	0.13
Primary amide	C(=[OS])([Csp3])[NH2]	7.28	5	0.10
Secondary amide	C(=[OS])([Csp3])[NH][Csp3]	6.99	1	
Tertiary amide	C(=[OS])([Csp3])N([Csp3])[Csp3]	6.77	1	
Urea	C(=O)([GenAmino])[GenAmino]	7.80	1	
Nitro	[Nitro][Csp3]	3.21	14	0.14
Sulfoxide	[SX3](=O)([Csp3])[Csp3]	6.95	1	
Thiol	[S;X2;H][Csp3]	1.47	5	0.27
Thioether	[SX2]([Csp3])[Csp3]	1.85	5	0.50
Disulfide	[SX2][SX2]	0.72	2	0.37
Thiourea	C(=[SX1])([GenAmino])[GenAmino]	7.80	3	

These parameters were set independently of the aromatic carbon parameters

^a See Table 1 for vector binding definition

^b Equation 2

^c Number of data points used to determine parameters

^d Root mean square error from parameter fitting

aromatic carbon are listed in Table 4. A generic unsaturated carbon atom type was used in the SMARTS although in most cases the measured data correspond only to the aromatic carbon type.

Interaction parameters

Interactions between atoms within molecules such as intramolecular hydrogen bonds affect partition coefficients and are treated as secondary perturbations in the ClogP_{alk} model. While these effects can also be accounted for within the functional group framework, treating interactions in this manner allows more flexibility in parameterization and simplifies the process of specifying substructures. Values for the interaction parameters depend on the functional group parameters and their derivation represents the final step of the parameterization process. The data set available is sparse and deriving interaction parameters typically involves an element of extrapolation. The interaction parameters are listed in Table 5.

The nature of the interaction parameters is best illustrated with reference to examples. The logP_{alk} value measured for Ephedrine was used to set an intramolecular (5-membered ring) hydrogen bond term for the interaction of amine nitrogen with hydroxyl. A parameter of the same value was then assigned to describe the contribution of the analogous intramolecular hydrogen bond in a 6-membered ring and the substructural specification for the hydroxyl was broadened to match other hydrogen bond donors such as amides and sulfonamides. The effect of an intramolecular hydrogen bond (6-membered ring) involving carbonyl oxygen as the acceptor was parameterized using logP_{alk} measured for 2-hydroxybenzaldehyde. A parameter for the through-bond interactions between ether oxygen atoms was derived by comparing the logP_{alk} measurement for 1,4-dioxane with the value calculated using the ether oxygen parameter in Table 3. Heteroaromatic rings with more than one heteroatom were also treated using interaction parameters. For example, the interaction between the nitrogen atoms in quinoxaline was quantified by comparing its measured logP_{alk} with the value predicted using aromatic nitrogen

Table 4 Functional group contributions

Description	SMARTS ^a	$f_{FG,i}^b$	N ^c	RMSE ^d
ArNH ₂	[NH2]c	2.59	18	0.20
ArNHR	[NH](c)[Csp3]	2.00	2	0.10
ArNR ₂	N([Csp3])([Csp3])c	1.34	3	0.15
ArC#N	C#[NX1]c	1.73	4	0.08
ArOH	[OH]c	3.22	43	0.36
ArOR	O(c)[Csp3]	0.94	6	0.11
ArCHO	[CH](=[OS])[Cnotsp3]	1.66	7	0.15
ArC(=O)R	C(=[OS])([Cnotsp3])[Csp3]	2.19	7	0.15
ArC(=O)Ar	C(=[OS])([Cnotsp3])[Cnotsp3]	1.54	4	0.26
ArCO ₂ H	O=C([OH])[Cnotsp3]	3.75	4	0.14
ArCO ₂ R	C(=[OS])([Cnotsp3])O[Csp3]	2.01	5	0.28
RCO ₂ Ar	C(=[OS])([Csp3])O[Cnotsp3]	2.65	1	
2-Pyrone	[OS]=c1occcc1	3.02	1	
4-Pyrone	[OS]=c1ccocc1	3.54	1	
ArC(=O)NH ₂	C(=[OS])([Cnotsp3])[NH2]	5.26	2	0.27
ArC(=O)NHR	C(=[OS])([Cnotsp3])[NH][Csp3]	5.57	1	
RC(=O)NHAr	C(=[OS])([Csp3])[NH][Cnotsp3]	5.37	7	0.17
ArC(=O)NR ₂	C(=[OS])([Cnotsp3])N([Csp3])[Csp3]	5.19	2	0.20
RC(=O)N(R)Ar	C(=[OS])([Csp3])N([Cnotsp3])[Csp3]	4.37	4	0.45
Imide	C(=[OS])[ND3][SEWG1]	2.0	1	
Imide	c(=[OS])[nD3]c=[OS]	2.0 ^e		
Imide with NH	C(=[OS])[NH][SEWG1]	2.6	1	
Imide with NH	c(=[OS])[nH]c=[OS]	2.6 ^e		
2-Pyridone (NH)	c(=[OS])[nH]	6.9 ^f		
2-Pyridone (NR)	c(=[OS])[nD3]	5.68	1	
4-Pyridone	c1(=[OS])[ArCN][ArCN][nX3]cc1	9.04	1	
ArNO ₂	[SNitro][Cnotsp3]	1.31	6	0.17
Sulfone	S(=O)(=O)([#6])([#6])	4.60	1	
Primary/secondary sulfonamide	S(=O)(=O)[N;H2,H1]	6.19	1	
Tertiary sulfonamide	S(=O)(=O)[N;!H;!H2]	4.21	1	
P=O	P=O	7.1	10	0.41
Thiol	[SH]c	0.45	1	
Thioether	[SX2](c)[Csp3]	0.68	1	

These parameters are all dependent on the values of the aromatic carbon parameters

^a See Table 1 for vector binding definition

^b Equation 2

^c Number of data points used to determine parameters

^d Root mean square error from parameter fitting

^e Determined from aliphatic equivalent

^f Determined from *N*-alkyl equivalent by adding 1.2 log units

parameters derived from logP_{alk} values measured for pyridine, quinolone, isoquinoline and their alkylated derivatives.

Validation

The ClogP_{alk} model is applied using 82 functional group definitions and 41 interaction definitions and it is important

to stress that the functional group parameters are either derived independently of each other or depend on the parameter(s) for at most one other atom type (e.g. aromatic carbon). In this respect it differs from a multivariate regression model in which the parameters are dependent on each other (and the extent to which descriptors are correlated with each other). One of the reviewers of the manuscript raised questions about the number of degrees of

Table 5 Interaction parameters for ClogP_{alk} model

SMARTS ^a	f _{Int,j} ^b	Description
[OH]-[#6]~[#6]-[OH]	-0.7	HB between hydroxyls
[OH]-[#6]~[#6]-O[$\text{\$Csp3}$]	-1.3	Hydroxyl/ether HB
[\$HBD]-[#6;!\$(C=[$\text{\$OS}$])~[#6]-C=O]	-3.3	Intramolecular HB (e.g. 2-hydroxybenzaldehyde)
[\$HBD]-[#6;!\$(C=[$\text{\$OS}$])~[#6]-[$\text{\$Nitro}$]]	-2.8	Intramolecular HB (e.g. 2-nitrophenol)
[\$HBD2]-[#6]~[#6]=O	-5.0	HB to carbonyl; 5-membered ring
[\$HBD2]-[#6]~[#6](=O)~[#6]-[\$HBD2]	2.5	Avoid double counting of intramolecular HB
O=[#6]~[#6]([$\text{\$HBD2}$])~[#6]=O	2.5	Avoid double counting of intramolecular HB
[\$HBD]-[#6;!\$(C=[$\text{\$OS}$])~[#6]-C=O]	-3.7	HB to carbonyl; 6-membered ring
[\$HBD]-c[aR3][$\text{\$ArN}$]	-4.0	HB to aromatic nitrogen; 5-membered ring
[\$HBD]-[$\text{\$Csp3}$;!\$CyAmine]-[$\text{\$Csp3}$]-[$\text{\$Amine}$]	-3.2	HB to amine; 5-membered ring
[\$HBD]-[$\text{\$Csp3}$;!\$CyAmine]-[$\text{\$Csp3}$]-[$\text{\$Csp3}$]-[$\text{\$Amine}$]	-3.2	HB to amine; 6-membered ring
[\$BzCorr]	-0.9	Benzyl groups tend to make amines and alcohols more lipophilic
[\$NOCCNO]	-0.75	Through bond (2) interaction between ether O or amine N
[\$NOCNO]	-0.75	Through bond (1) interaction between ether O or amine N
[\$Amine]-[$\text{\$Csp3}$]-C=O	-1.6	Through bond interaction between amine and oxygen
[\$ArNH,\$ArNR][$\text{\$ArC}$][$\text{\$ArN}$]	0.8	Heteroatomic interaction (imidazole)
[\$ArNH,\$ArNR][$\text{\$ArN}$]	-0.2	Heteroaromatic interaction (pyrazole)
[\$ArO,\$ArS][$\text{\$ArC}$][$\text{\$ArN}$]	-0.9	Heteroaromatic interaction (oxazole/thiazole)
[\$ArO,\$ArS][$\text{\$ArN}$]	-0.7	Heteroaromatic interaction (isoxazole/isothiazole)
[\$ArN][$\text{\$ArN}$]	-0.6	Heteroaromatic interaction (1,2-diaza)
[\$ArN][$\text{\$ArC}$;R2][$\text{\$ArN}$]	-0.9	Heteroaromatic interaction (1,3-diaza)
[\$ArN]1[$\text{\$ArC}$][$\text{\$ArC}$][$\text{\$ArN}$][$\text{\$ArC}$][$\text{\$ArC}$]1	-0.6	Heteroaromatic interaction (1,4-diaza)
[\$ArN]1[$\text{\$ArC}$][$\text{\$ArC}$]($\text{\$ArN}$)[$\text{\$ArC}$][$\text{\$ArC}$][$\text{\$ArC}$]1	-0.7	Heteroaromatic interaction (e.g. 1,5-naphthyridine)
[\$ArN]1[$\text{\$ArCN}$][$\text{\$ArCN}$][$\text{\$ArN}$][$\text{\$ArO}$, $\text{\$ArC}$, $\text{\$ArS}$,n&X3]1	-1.2	Heteroaromatic interaction (1,4-diaza in 5-membered ring)
[\$ArN;r5][aR3][aR3][$\text{\$ArN}$ &r6,(c=O)]	-1.4	Purines
[\$OS]=c1[nX3][$\text{\$ArC}$][$\text{\$ArN}$][$\text{\$ArC}$][$\text{\$ArC}$]1	-1.3	Pyrimidones
[\$OS]=c1[$\text{\$ArN}$][$\text{\$ArC}$][nX3][$\text{\$ArC}$][$\text{\$ArC}$]1	-4.6	Pyrimidones
[\$OS]=c1[nX3][$\text{\$ArC}$][$\text{\$ArC}$][$\text{\$ArC}$][$\text{\$ArN}$]1	-3.2	Pyrimidones
n1([$\text{\$Csp3}$])[$\text{\$ArC}$][$\text{\$ArN}$][$\text{\$ArC}$][$\text{\$ArC}$][$\text{\$ArC}$]1[$\text{\$ArN}$]	3.75	3-Alkylpurine
[\$Am4Pyr]	2.4	4-Aminopyridines and aza analogues
O=[C;!R]c(c[$\text{\$Csp3}$])c[$\text{\$Csp3}$]	0.7	Carbonyl flanked by alkyls
O=C[N;X3;!R]c(c[$\text{\$Csp3}$])c[$\text{\$Csp3}$]	0.5	Amide flanked by alkyls
[C;R](=[$\text{\$OS}$])[O;R]	1.4	Lactones
[\$([CR][$\text{\$Csp3}$]),\$([CH])](=[$\text{\$OS}$))C=C	1.5	Unsaturated aldehyde or cyclic ester
c(=[$\text{\$OS}$])([nX3])[aR3]	-1.4	Carbonyl next to ring fusion
c1(=[$\text{\$OS}$])([aR3])[$\text{\$ArCN}$][$\text{\$ArCN}$][nX3][$\text{\$ArC}$][$\text{\$ArC}$]1	-1.6	Carbonyl next to ring fusion
[\$SO2]	1.8	Aliphatic sulfone/sulfonamide correction

^a See Table 1 for vector binding definition

^b Equation 3

freedom in the model. In responding to this question we note that a separate training set was generated for each functional group so the relevant parameters were derived independently of each other. Each training set was specific to the functional group being parameterized so that there was only a single degree of freedom in the fitting process unless an intercept term was used, in which case there were two. The number of logP_{alk} measurements used to derive

each parameter is given in Tables 2, 3 and the values of RMSE listed in those tables indicate how well each parameter fits its training set. Interaction terms were typically set using single pairs of logP_{alk} measurements. Although orthogonal training sets were used to fit parameters, they can be combined to provide a summary of how well the model fits the data used to train it. The combined training sets consisted of a total of 743 logP_{alk}

Table 6 Distribution of $\log P_{\text{alk}} - \text{Clog}P_{\text{alk}}$ for training set and external test set

Quantile($\log P_{\text{alk}} - \text{Clog}P_{\text{alk}}$)	Training set (N = 743)	External test set (N = 100)
0 % (minimum)	-1.604	-2.762
10 %	-0.214	-1.079
25 % (lower quartile)	-0.099	-0.313
50 % (median)	0.001	0.038
75 % (upper quartile)	0.102	0.591
90 %	0.252	1.407
100 % (maximum)	1.097	4.647

measurements for 515 compounds. The distribution of residuals ($\log P_{\text{alk}} - \text{Clog}P_{\text{alk}}$) is summarized as percentiles in Table 6 and the histogram in Fig. 2.

The $\text{Clog}P_{\text{alk}}$ model was validated using an external test set of 100 $\log P_{\text{alk}}$ measurements for 91 compounds, most of which were drugs (Fig. 3). Residual ($\log P_{\text{alk}} - \text{Clog}P_{\text{alk}}$) values are listed in Table 7 and their distribution is summarized as percentiles in Table 6. The interquartile range for these residuals is 0.90 log units. The most negative (over-predicted) residual is the value of -2.76 that was observed for one of $\log P_{\text{alk}}$ measurements for atropine although the residual (-0.69) corresponding to the other measurement for this compound is smaller in magnitude. The two most positive residuals calculated for the test data correspond to the steroids cortisone (4.65) and hydrocortisone (3.86) although the residuals for other steroids such as cortexolone (0.55), estradiol (-0.06), estrone (-0.03) and testosterone (-0.02 and 0.02) are much smaller in magnitude. Cortisone ($\log P_{\text{alk}} = -0.55$) and hydrocortisone ($\log P_{\text{alk}} = -2.04$) are respectively keto and hydroxy derivatives of cortexolone ($\log P_{\text{alk}} = -1.00$) and the observation that the introduction of ketone functionality into the latter appears to increase lipophilicity raises questions about the consistency of the data. The $\log P_{\text{hxd}}$ values [22] measured for cyclohexane (3.91), cyclohexanone (0.19) and cyclohexanol (-0.25) suggest that cortisone and hydrocortisone would be

expected to be at least 3.5 log units less lipophilic than cortexolone in alkane/water systems. Although self-association in the alkane might be invoked to explain these observations, one would still need a convincing argument for why it was not an issue for cortexolone.

Another suggestion made by a reviewer of the manuscript was that it would be helpful to illustrate the similarity of the training and external test sets as a histogram of nearest neighbor similarities. This addresses the concern that a model predicts by locating neighbors rather than by capturing the response of the property to the descriptors. Although this is less of an issue for a perturbation-based model like $\text{Clog}P_{\text{alk}}$ than it would be for a multivariate regression model, the information is still useful. The quantity of interest here is the similarity between each molecular structure in the external test set and its closest analog amongst the structures used to train the model. A histogram of nearest neighbor similarities is shown in Fig. 4 and the maximum value found was 0.74.

Discussion

The primary objective of this study was to investigate how far what is an essentially cheminformatic approach to prediction of $\log P_{\text{alk}}$ could be taken. The performance of the $\text{Clog}P_{\text{alk}}$ model against an external test suggests that it represents a tractable approach to predicting alkane/water partition coefficients. The parameters are completely external to the software so the method could also be applied to the prediction of $\log P$ for other partitioning systems. It should be explicitly stated that the atom-typing scheme used in this study is unlikely to be of much use for scoring of poses generated by docking because the perturbation due to a functional group is loaded onto a single atom within the functional group that has typically been chosen to facilitate coding of the SMARTS.

Availability of data is always an issue when developing predictive models like $\text{Clog}P_{\text{alk}}$ that are based on

Fig. 2 Distributions of residuals for training set (a) and external test set (b)

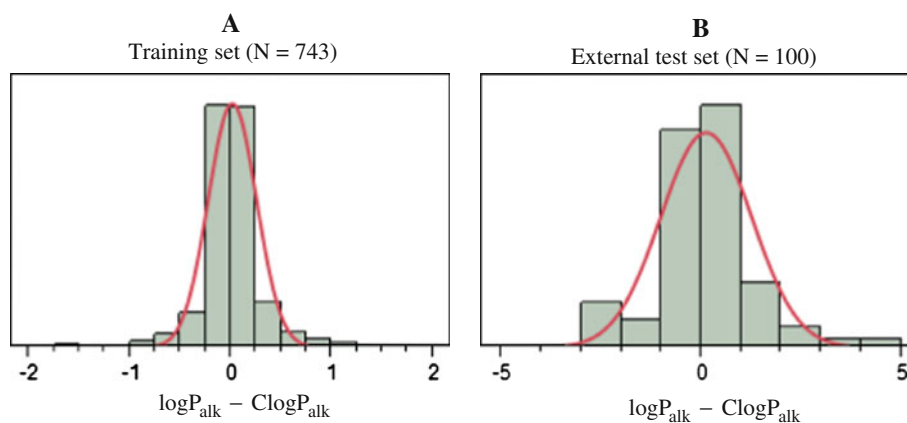


Table 7 Measured $\log P_{\text{alk}}$ values and residuals for external test set (N = 100)

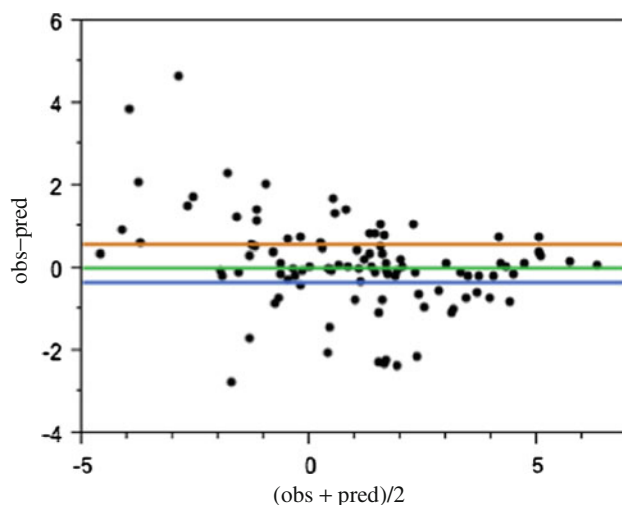
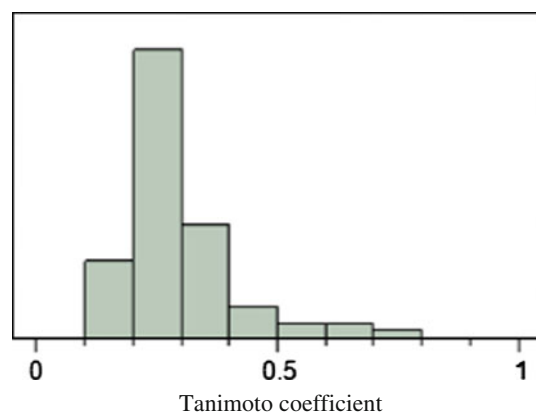
Compound	$\log P_{\text{alk}}$	$\log P_{\text{alk}} - \text{Clog} P_{\text{alk}}$	Reference
17- α -Methyltestosterone	0.84	0.02	30
3-Acetylpyridine	-1.18	0.30	22
4-Acetylpyridine	-0.93	0.55	22
5,5-DiethylbarbituricAcid	-2.00	-0.05	30
5-Ethyl,5-(s-Pentyl)barbituricAcid	-1.16	-0.83	30
Aldosterone	-2.70	2.10	60
Alprenolol	1.30	0.23	36
Aminopyrine	-0.59	-0.26	30
Anabesine	-0.58	0.41	30
Androstene-3-17-dione	1.09	0.00	30
Antipyrine	-2.15	-1.68	30
Aspirin	-2.00	-0.17	32
Astemizole	0.95	-0.33	36
Atropine	-1.02	-0.69	32
Atropine	-3.09	-2.76	30
Azatadine	1.82	0.52	36
Benzoquinone	-0.40	-0.16	58
Chlorphenamine	2.09	0.07	36
Chlorpromazine	5.24	0.33	55
Cortexolone	-1.00	0.56	60
Corticosterone	-1.62	-0.09	60
Cortisone	-0.55	4.65	60
Deoxycorticosterone	0.48	-2.30	60
Deoxycorticosterone	0.56	-2.22	30
Deprenyl	2.81	1.07	32
Desipramine	3.38	-0.17	32
Diclofenac	1.88	-0.06	32
Diltiazem	1.37	1.68	36
Dimetindene	1.65	-0.06	36
Diphenhydramine	2.56	-1.08	36
Diphenhydramine	2.67	-0.97	36
Epinastine	1.76	0.86	36
Eserine	-0.46	1.42	36
Estradiol	-0.2	-0.06	60
Estriol	-0.64	2.32	60
Estrone	0.48	-0.03	60
Fluoroxene	1.30	-2.14	24
Fluoxetine	3.62	-0.17	32
Fluoxetine	3.10	-0.69	36
Fluphenazine	3.28	-0.11	55
Flurbiprofen	1.49	0.33	36
Furosemide	-0.60	0.12	36
Haloperidol	1.53	1.42	36
Halothane	2.10	0.21	22
Hydrocortisone	-2.04	3.86	60
Hydroxyprogesterone	0.40	-2.25	60
Hydroxyzine	1.25	0.43	36

Table 7 continued

Compound	$\log P_{\text{alk}}$	$\log P_{\text{alk}} - \text{Clog} P_{\text{alk}}$	Reference
Ibuprofen	2.08	1.06	36
Ibuprofen	1.88	0.86	32
Imipramine	3.60	-0.72	36
Indomethacin	0.98	-1.05	36
Isoflurane	1.65	-0.15	22
Lidocaine	1.23	1.31	32
Methoxyflurane	2.04	-0.95	32
Metoprolol	-0.38	-0.02	36
Miconazole	3.39	-0.58	32
<i>N,N</i> -Dimethyl-4-chlorocinnamide	0.70	0.10	54
<i>N</i> -Acetyltryptamine	-1.93	1.52	17
Naproxen	0.53	0.60	36
<i>N</i> -Ethyl-4-methoxycinnamide	-0.70	-0.11	54
Nicotine	0.20	0.76	30
<i>N</i> -Pentylcinnamide	1.38	0.03	54
Papaverine	0.75	-2.34	36
Papaverine	2.56	-0.53	32
Pecazine	4.80	0.15	55
Penbutolol	3.06	0.11	36
Phenothiazine	3.97	-0.82	30
Phenytoin	-0.30	-1.44	36
Pregnenolone	0.62	-0.74	60
Procaine	-0.13	0.69	56
Prochlorperazine	4.28	0.13	55
Progesterone	1.23	-0.75	60
Promazine	3.96	-0.20	30
Promethazine	4.41	-0.14	55
Propranolol	-0.64	-2.05	32
Propranolol	1.75	0.34	36
Pyrene	5.22	0.40	65
Quinine	-0.39	-0.39	32
Quinine	0.04	0.04	36
Sulfamethizole	-3.66	0.95	30
Sulfathiazole	-4.42	0.35	30
Sulfisoxazole	-3.40	0.62	30
SulfurHexafluoride	2.11	-0.62	24
Teflurane	1.74	0.14	22
Testosterone	0.41	-0.02	60
Testosterone	0.45	0.02	30
Tetracaine	2.04	0.82	32
Tetraethylsilane	6.36	0.10	24
Tetramethylsilane	4.37	0.07	24
Theophylline	-1.70	1.74	30
Thiopental	0.52	0.49	36
Thioridazine	5.81	0.18	55
Thymol	1.40	-0.08	36
Timolol	-1.00	1.22	36

Table 7 continued

Compound	$\log P_{\text{alk}}$	$\log P_{\text{alk}} - \text{Clog} P_{\text{alk}}$	Reference
Tramadol	1.82	-0.16	36
Trifluoperazine	4.53	0.75	55
Triflupromazine	5.42	0.77	55
Tryptamine	-0.60	1.15	32
Verapamil	2.27	-0.09	36
Warfarin	0.05	2.04	36

**Fig. 3** Plot showing relationship between observed ($\log P_{\text{alk}}$) and predicted ($\text{Clog} P_{\text{alk}}$) values of alkane/water partition coefficient for external test set. Median and quartile values for $(\text{obs} - \text{pred})$ are shown as horizontal lines**Fig. 4** Histogram showing distribution of nearest neighbor (external test versus training data) similarity for external test set structures

substructural rules and derivation of parameters typically requires assumptions to be made. For example, a number of substructural patterns include match sp^2 or sp carbon although there may only have been measured data when the relevant atom is aromatic carbon. The urea parameters

were observed to be similar to those derived for thioureas and substructural patterns used to identify amides were extended to match the corresponding thioamides. Interactions, either direct or through bonds, between polar atoms present the main challenge when deriving parameters for substructurally-based lipophilicity models and these interaction terms are likely to represent the weak link in the current parameterization of the model. Experimental data for compounds in which these interactions are most pronounced would have the greatest value for developing models like $\text{Clog} P_{\text{alk}}$. Heterocycles in which a high proportion of the atoms in the ring are heteroatoms and compounds, such as tetrazole and 4-dimethylaminopyridine that can present particularly strong hydrogen bond donors or acceptors to solvent would be of particular interest in this regard. While intramolecular hydrogen bonding does present a challenge for prediction methodology, it should be noted that measurement of $\log P_{\text{alk}}$ is actually a powerful tool for studying intramolecular hydrogen bonding and investigating conformational effects on solvation [29]. Measured $\log P_{\text{alk}}$ is also a more widely accessible property than gas to water transfer free energy that is normally used when developing models for aqueous solvation. For example, a data set [66] that has been widely used for parameterization of aqueous solvation models consists of thermodynamic measurements for structurally prototypical compounds that are largely monofunctional. Models derived from this data set are then used to predict solvation energies for polyfunctional compounds and occasionally regions of protein molecular surfaces.

The parameters for the $\text{Clog} P_{\text{alk}}$ model are read from a text file and visible to the user. The model has been explicitly designed to facilitate customization by users since we believe this adds value to predictive modeling software. The most likely customization scenario is one in which proprietary data is available for compounds with substructures that are inadequately parameterized in the current model and there are two ways in which parameters can be added. Typically the substructure in question will be a variation of a substructure that has already been parameterized and an interaction term can be used to encode the variation as a perturbation of the existing substructure. Alternatively, the substructure can be included as a new functional group and inserted into the parameter file after the existing functional groups, ensuring that the more specific definition takes priority over the more general one in the existing model.

When considering how a model like $\text{Clog} P_{\text{alk}}$ might be used in drug discovery, it is important to stress that $\log P_{\text{oct}}$ is likely to remain an important design parameter in lead optimization which is often carried out within the framework of structural series. At the same time, it should be remembered that the original selection of 1-octanol/water

for partitioning studies was arbitrary and not based on an objective comparison of different solvent systems. Often polar functionality is relatively conserved within a structural series and logP values measured using different partitioning systems will tend to differ by a constant when this is the case. Provided that the properties of interest respond linearly to logP, the choice of partitioning system will not usually be critical when optimizing within a single lead series.

The situation is different in lead identification where lipophilicity measures are used to prioritize compounds and structural series for hit-to-lead chemistry and further optimization. The difference, $\Delta\log P$, between $\log P_{\text{oct}}$ and $\log P_{\text{alk}}$ can be considered as a measure of the hydrogen bonding capacity of a compound and, outside structural series, cut off values cannot simply be shifted by a constant to account for differences between partitioning systems. Lipophilicity is also used to create efficiency metrics [67] (e.g. $\text{pIC}_{50} - \log P$) which can be used to compare compounds and structural series. Offsetting potency or affinity by lipophilicity in this manner has the effect (at least for neutral compounds) of shifting the reference state for the binding equilibrium from the aqueous to an organic phase. Affinity can also be scaled by molecular size and the original measure of ligand efficiency [68] was obtained by dividing the standard Gibbs free energy of binding (ΔG°) by number of non-hydrogen atoms. Reference states also need to be considered carefully when affinity is scaled because relative values of ligand efficiency for compounds differing in molecular size depend on the standard concentration used to define ΔG° [69]. Whether one scales or offsets affinity or potency, one is implicitly assuming that the relationship with the relevant physicochemical or molecular property is linear. When biological activity is offset by lipophilicity one assumes a unit slope in the linear relationship while scaling by molecular size implies an assumption that K_d and IC_{50} values in all assays will tend to the same concentration (usually 1 M) in the limit of zero molecular size. An alternative to using efficiency metrics for evaluating compounds that differ in their activity and physicochemical characteristics is to fit affinity or potency to measures of lipophilicity and/or molecular size and use the residuals to quantify the extent to which compounds beat (or are beaten by) the underlying trend in the data. One advantage of analyzing measured biological activity in this manner is that the results are invariant with respect to standard concentration. This is not the case when ligand efficiency is ‘corrected’ for molecular size [70].

The $\text{ClogP}_{\text{alk}}$ model calculates $\log P_{\text{alk}}$ as the sum of two components, one of which can be taken as a measure of molecular size and the other of polarity. This may prove useful in gaining a better understanding the nature of

drug-likeness although the assumption that marketed drugs represent a separate and contiguous region of chemical space can be challenged. One issue that must be addressed when using lipophilicity in modeling is whether logP or logD is more relevant to phenomena of interest. Typically, logP will be more relevant when compounds bind to their targets (and anti-targets) in their ionized forms while logD is more likely to be the measure of choice when the concentration of neutral form is a limiting factor as would normally be the case for aqueous solubility and passive permeation through membranes. It is instructive to write logD as a function of logP and the fraction of compound in its neutral form (F_{neut}):

$$\log D(\text{pH}) = \log P + \log F_{\text{neut}}(\text{pH}) \quad (5)$$

When logD is used to model phenomena that are influenced by ionization, the responses to partition coefficient and ionization in the resulting models are constrained to be equal in magnitude even though the two components of logD describe completely different aspects of a compound’s behavior. In drug discovery programs, logD (at a single pH) is usually the only lipophilicity measurement that is available on a routine basis and options are limited if one needs to use an experimentally measured lipophilicity value. If, however, logD is calculated from predicted values of logP and pK_a , there is no longer any need to use the composite parameter and the response of the property of interest to each component of logD can be modeled independently. One advantage of $\text{ClogP}_{\text{alk}}$ over predicted $\log P_{\text{oct}}$ in this type of analysis is the ease with which it can be resolved into molecular size and polarity components. These could in turn be combined with predicted (possibly at more than one pH) $\log F_{\text{neut}}$ to create a basis for describing physicochemical characteristics of compounds in the context of drug discovery.

We have described a method for predicting alkane/water partition coefficient from molecular structure and make the source code publicly available so that the $\text{ClogP}_{\text{alk}}$ model can be evaluated more widely. This could also have the benefit of drawing more data into the public domain. Our study also shows how views of chemistry can be imposed automatically on chemical structures using systems of SMARTS definitions in a manner that is both reproducible and transparent. We hope that the model will prove useful in pharmaceutical research and that our study will stimulate debate about the relative merits of different partitioning systems.

Acknowledgments We are grateful to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and the Conselho Nacional de Pesquisa (CNPq) for financial support. We also thank OpenEye Scientific Software for providing access to software and the anonymous reviewers of the manuscript for their constructive comments.

References

- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 23:3–25
- Van de Waterbeemd H, Smith DA, Jones BC (2001) Lipophilicity in PK design: methyl, ethyl, futile. *J Comput-Aided Mol Des* 15:273–286
- Harless E, von Bibra E (1847) Die Ergebnisse der Versuche über die Wirkung des Schwefeläthers, Erlangen
- Nernst W (1891) Verteilung eines Stoffes zwischen zwei Lösungsmitteln und zwischen Lösungsmittel und Dampfraum. *Z Phys Chem* 8:110–139
- Meyer H (1899) Zur Theorie der Alkoholnarkose. *Arch Exp Pathol Pharmacol* 42:109–118
- Overton CE (1901) Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie. Gustav Fischer, Jena
- Collander R (1951) Partition of organic compounds between higher alcohols and water. *Acta Chem Scand* 5:774–780
- Meyer KH (1937) The theory of narcosis. *Trans Far Soc* 33:1062–1068
- Hansch C, Muir RM, Fujita T, Maloney PP, Geiger F, Streich M (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J Am Chem Soc* 85:2817–2824
- Banks WA, Kastin A (1985) Peptides and the blood-brain barrier: lipophilicity as a predictor of permeability. *Brain Res Bull* 15:287–292
- Kellogg GE, Burnett JC, Abraham DJ (2001) Very empirical treatment of solvation and entropy: a force field derived from Log Po/w. *J Comput-Aid Mol Des* 15:381–393
- Schneider N, Lange G, Hindle S, Klein R, Rarey MA (2013) A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. *J Comput-Aid Mol Des* 27:15–29
- Delaney JS (2005) Predicting aqueous solubility from structure. *Drug Discov Today* 10:289–295
- Waring MJ, Johnstone C (2007) A quantitative assessment of hERG liability as a function of lipophilicity. *Bioorg Med Chem Lett* 17:1759–1764
- Rytting JH, Davis SS, Higuchi T (1972) Suggested thermodynamic standard state for comparing drug molecules in structure-activity studies. *J Pharm Sci* 61:816–818
- Finkelstein A (1976) Water and nonelectrolyte permeability of lipid bilayer membranes. *J Gen Physiol* 68:127–135
- Radzicka A, Wolfenden R (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochem* 27:1664–1670
- Lambert WJ, Wright LA (1989) Prediction of alkane-water partition coefficients using a C₁₈ derivatized polystyrene-divinylbenzene stationary phase. *J Chromat* 464:400–404
- Xiang T-X, Anderson BD (1994) Substituent contributions to the transport of substituted p-toluic acids across lipid bilayer membranes. *J Pharm Sci* 83:1511–1518
- Mayer PT, Anderson BD (2002) Transport across 1, 9-decadiene precisely mimics the chemical selectivity of the barrier domain in egg lecithin bilayers. *J Pharm Sci* 91:640–646
- Toulmin A, Wood JM, Kenny PW (2008) Toward prediction of alkane/water partition coefficients. *J Med Chem* 51:3720–3730
- Abraham MH, Chadha HS, Whiting GS, Mitchell RC (1994) Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the $\Delta\log P$ parameter of Seiler. *J Pharm Sci* 83:1085–1100
- Dallas AJ, Carr PW (1992) A thermodynamic and solvatochromic investigation of the effect of water on the phase-transfer properties of octan-1-ol. *J Chem Soc Perkin Trans* 2:2155–2161
- Abraham MH, Whiting GS, Fuchs R, Chambers EJ (1990) Thermodynamics of solute transfer from water to hexadecane. *J Chem Soc Perk Trans* 2:291–300
- Wolfenden R, Radzicka A (1994) On the probability of finding a water molecule in a nonpolar cavity. *Science* 265:936–937
- Tsai R-S, Fan W, El Tayar N, Carrupt P-A, Testa B, Kier LB (1993) Solute-water interactions in the organic phase of a biphasic system. 1. Structural influence of organic solutes on the “water-dragging” effect. *J Am Chem Soc* 115:9632–9639
- Prokopenko NA, Bethea IA, Clemens CJ, Klimek A, Wargo K, Spivey C, Waziri K, Grushow A (2002) The effect of structure on hydrogen bonding: hydrogen bonded lactam dimers in CCl₄. *Phys Chem Chem Phys* 4:490–495
- Golumbic C, Orchin M, Weller S (1949) Partition studies on phenols. I. Relation between partition coefficient and ionization constant. *J Am Chem Soc* 71:2624–2627
- Dearden JC, Bresnen GM (2005) Thermodynamics of water-octanol and water-cyclohexane partitioning of some aromatic compounds. *Int J Mol Sci* 6:119–129
- Seiler P (1974) Interconversion of lipophilicities from hydrocarbon/water systems into the octanol/water system. *Eur J Med Chem* 9:473–479
- Leahy DE, Morris JJ, Taylor PJ, Wait AR (1992) Model solvent systems for QSAR. Part 2. Fragment values (f-values) for the critical quartet. *J Chem Soc Perkin Trans* 2:723–731
- Zissimos AM, Abraham MH, Barker MC, Box KJ, Tam KY (2002) Calculation of Abraham descriptors from solvent-water partition coefficients in four different systems; evaluation of different methods of calculation. *J Chem Soc Perkin Trans* 2:470–477
- Saunders RA, Platts JA (2004) Scaled polar surface area descriptors: development and application to three sets of partition coefficients. *New J Chem* 28:166–172
- Zerara M, Brickmann J, Kretschmer R, Exner TE (2008) Parameterization of an empirical model for the prediction of n-octanol, alkane and cyclohexane/water as well as brain/blood partition coefficients. *J Comput-Aided Mol Des* 23:105–111
- Lamarche O, Platts JA, Hersey A (2004) Theoretical prediction of partition coefficients via molecular electrostatic and electronic properties. *J Chem Inf Comp Sci* 44:848–855
- Caron G, Ermondi G (2005) Calculating virtual log P in the alkane/water system $\log P_{\text{alk}}^N$ and its derived parameters $\Delta \log P_{\text{oct-alk}}^N$ and $\log D_{\text{alk}}^{\text{pH}}$. *J Med Chem* 48:3269–3279
- Wittekindt C, Klamt A (2009) COSMO-RS as a predictive tool for lipophilicity. *QSAR Comb Sci* 28:874–877
- OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508. <http://www.eyesopen.com>. Accessed 28 Feb 2013
- Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comp Sci* 28:31–36
- Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comp Sci* 29:97–101
- OMEGA. OpenEye Scientific Software. <http://www.eyesopen.com/omega>. Accessed 28 Feb 2013
- Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *J Chem Inf Model* 50:572–584
- Halgren TA (1999) MMFF VI. MMFF94S option for energy minimization studies. *J Comp Chem* 20:720–729

44. SZYBKI. OpenEye Scientific Software. <http://www.eyesopen.com/szybki>. Accessed 28 Feb 2013
45. Ewing VC, Sutton LE (1963) Investigation by electron diffraction of the molecular structures of sulphur hexafluoride, sulphur tetrafluoride, selenium hexafluoride and selenium tetrafluoride. *Trans Faraday Soc* 59:1241–1247
46. Bondi A (1964) van der Waals volumes and radii. *J Phys Chem* 68:441–451
47. GraphSim Toolkit. OpenEye Scientific Software. http://www.eyesopen.com/docs/toolkits/current/html/GraphSim_TK-c++/index.html. Accessed 16 May 2013
48. JMP version 10.0, SAS Institute, Cary, NC 27513. <http://www.jmp.com>. Accessed 28 Feb 2013
49. OEChem Toolkit. OpenEye Scientific Software. http://www.eyesopen.com/docs/toolkits/current/html/OEChem_TK-c++/index.html. Accessed 28 Feb 2013
50. Blomberg N, Cosgrove DA, Kenny PW, Kolmodin K (2009) Design of compound libraries for fragment screening. *J Comput-Aid Mol Des* 23:513–525
51. SMARTS Theory Manual. Daylight chemical information systems. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed 18 Feb 2013
52. Spicoli Toolkit. OpenEye Scientific Software. http://www.eyesopen.com/docs/toolkits/current/html/Spicoli_TK-c++/index.html. Accessed 28 Feb 2013
53. Currie DJ, Lough CE, Silver RF, Holmes HL (1966) Partition coefficients of some conjugated heterocyclic compounds and 1, 4-naphthoquinones. *Can J Chem* 44:1035–1043
54. Delaney AD, Currie DJ, Holmes HL (1969) Partition coefficients of some *N*-alkyl and *N,N*-dialkyl derivatives of some cinnamides and benzalcyanoacetamides in the system cyclohexane-water. *Can J Chem* 47:3273–3277
55. Vezin WR, Florence A (1979) The determination of dissociation constants and partition coefficients of phenothiazine derivatives. *Int J Pharm* 3:231–237
56. Okada S, Nakahara H, Yomota C, Mochida K (1985) The role of solvent in the partition of procaine and *p*-aminobenzoic acid between organic solvent and water. *Chem Pharm Bull* 33:4916–4922
57. Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffiths R, Jones M, Rana KK, Saunders D, Smith IR, Sore NE, Wilks TJ (1988) Development of a new physicochemical model for brain penetration and its application to design of centrally acting H₂ receptor histamine antagonists. *J Med Chem* 31:656–671
58. Rich PR, Harper R (1990) Partition coefficients of quinones and hydroquinones and their relation to biochemical activity. *FEBS* 269:139–144
59. Gibbs PR, Radzicka A, Wolfenden R (1991) *J Am Chem Soc* 113:4714–4715
60. El Tayar N, Tsai R-S, Testa B, Carrupt P-A, Hansch C, Leo A (1991) Percutaneous penetration of drugs: a quantitative structure-permeability relationship study. *J Pharm Sci* 80:744–749
61. Guardado P, Balon M, Carmona C, Muñoz MA, Domene C (1997) Partition coefficients of indoles and betacarbolines. *J Pharm Sci* 86:106–109
62. Shih P, Pedersen LG, Gibbs PR, Wolfenden R (1998) Hydrophobicities of the nucleic acid bases: distribution coefficients from water to cyclohexane. *J Mol Biol* 280:421–430
63. Habgood MD, Liu ZD, Dehkordi LS, Khodr HH, Abbott J, Hider RC (1999) Investigation into the correlation between structure of hydroxypyridones and blood-brain barrier permeability. *Biochem Pharmacol* 57:1305–1310
64. Acree WE, Abraham MH (2001) Solubility predictions for crystalline nonelectrolyte solutes dissolved in organic solvents based on the Abraham general solvation model. *Can J Chem* 79:1466–1476
65. Acree WE, Abraham MH (2002) Solubility predictions for crystalline polycyclic aromatic hydrocarbons (PAHs) dissolved in organic solvents based upon the Abraham general solvation model. *Fluid Phase Equil* 201:245–258
66. Cabani S, Gianni P, Mollica V, Lepori L (1981) Group contributions to the thermodynamic properties of nonionic organic solutes in dilute aqueous solution. *J Solut Chem* 10:563–595
67. Albert JS, Blomberg N, Breeze AL, Brown AJH, Burrows JN, Edwards PD, HA FolmerR, Geschwindner S, Griffen EJ, Kenny PW, Nowak T, Olsson L, Sanganee H, Shapiro AB (2007) An integrated approach to fragment-based lead generation: philosophy, strategy and case studies from AstraZeneca's drug discovery programmes. *Curr Top Med Chem* 7:1600–1629
68. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* 9:430–431
69. Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* 72:1047–1069
70. Reynolds CH, Tounge BA, Bembenek SD (2008) Ligand binding efficiency: trends, physical basis, and implications. *J Med Chem* 51:2432–2438