J Comput Aided Mol Des (2009) 23:105–111
DOI 10.1007/s10822-008-9243-2

Parameterization of an empirical model for the prediction of *n*-octanol, alkane and cyclohexane/water as well as brain/blood partition coefficients

Mohamed Zerara · Jürgen Brickmann · Robert Kretschmer · Thomas E. Exner

Received: 31 July 2007 / Accepted: 2 September 2008 / Published online: 26 September 2008

Abstract Quantitative information of solvation and transfer free energies is often needed for the understanding of many physicochemical processes, e.g the molecular recognition phenomena, the transport and diffusion processes through biological membranes and the tertiary structure of proteins. Recently, a concept for the localization and quantification of hydrophobicity has been introduced (Jäger et al. J Chem Inf Comput Sci 43:237–247, 2003). This model is based on the assumptions that the overall hydrophobicity can be obtained as a superposition of fragment contributions. To date, all predictive models for the log*P* have been parameterized for *n*-octanol/water (logP_{oct}) solvent while very few models with poor predictive abilities are available for other solvents. In this work, we propose a parameterization of an empirical model for *n*-octanol/water, alkane/water (logP_{alk}) and cyclohexane/water (logP_{cyc}) systems. Comparison of both logP_{alk} and logP_{cyc} with the logarithms of brain/blood ratios (log*BB*) for a set of structurally diverse compounds revealed a high correlation showing their superiority over the logP_{oct} measure in this context.

Keywords Solvation and transfer free energies · Hydrophobic effect · Hydrophobic interaction · Log*P* · Brain/blood barrier · Octanol/water · Alkane/water · Cyclohexane/water

Introduction

Hydrophobic effect and interaction can be seen respectively as the interaction and the quasi-attractive interaction of non-polar molecules in the water phase. The important role played by hydrophobicity in governing pharmacokinetic processes led to an extensive use of the well known log*P* scalar quantity. The partition coefficient *P* is defined as the ratio of the concentration of a given compound in aqueous phase to its concentration in an immiscible solvent. In practice, log*P* can be measured for several combinations of more or less immiscible solvents with water such as *n*-octanol/water, the most common system, but also alkane/water, cyclohexane/water, etc. The choice of partition solvent is still controversial. However, *n*-octanol was often chosen as a simplified model of a phospholipid membrane since the work of Hansch and Leo [1] despite its failure to predict reliably brain/blood and skin penetrations.

Although numerous predictive models for log*P* have been developed for *n*-octanol/water, very few models have been reported for alkane/water or cyclohexane/water systems [2–4] One of these *n*-octanol/water models based on the three-dimensional free energy density (3D-FED) was introduced in [5, 6] and parameterized by Jäger et al. [7]. The scalar quantity 3D-FED offers a physical basis to the establishment of a predictive model with limited empirical character. Although this volume density is accessible from e.g Monte-Carlo methods, the *Grid* program [8] has proved to be suitable for a rapid evaluation of the 3D-FED for

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9243-2) contains supplementary material, which is available to authorized users.

M. Zerara · J. Brickmann
Department of Chemistry, Darmstadt University of Technology, Petersenstr. 20, 64287 Darmstadt, Germany

M. Zerara · J. Brickmann (✉)
MOLCAD GmbH, Petersenstr. 20, 64287 Darmstadt, Germany
e-mail: brickmann@molcad.de

R. Kretschmer · T. E. Exner
Fachbereich Chemie, Universität Konstanz, 78457 Konstanz, Germany

large sets of molecules. It uses an empirical force field to determine the interaction energy of a particular probe molecule (e.g. water solvent molecule) with a target molecule, the solute, for all points of a regular three-dimensional grid, in which this solute is enclosed.

In the present study, an improved parameterization using the same strategy than the one described in the previous work of Jäger et al. [6, 7] is done in two steps; some model parameters were fitted to computed data (provided by the *Grid* program), while other parameters were fitted to experimental log*P* values. Additionally and even more important, the concept is extended to different solvent systems, namely alkane/water and cyclohexane/water. Finally, due to the considerable demand for methods able to calculate brain/blood partition coefficients, the new parameterized models for each solvent are in turn compared to the logarithms of the experimental brain/blood values (log*BB*) in order to find out any correlation.

The model concept and its parameterization

Following the generally admitted idea, that the solvation of nonpolar molecules can be divided into a cavity-creation and a reorganization step, this concept can be applied for the calculation of partition coefficients according to Jäger et al. [7] as:

$$\log P = -\frac{\Delta\Delta G_{solv}}{2.303RT} = \alpha \sum_{l=1}^{120} h_l \sum_{j=0}^{1} a_{lj} c_j + \beta S + \gamma V \qquad (1)$$

The first term approximates the reorganization of the solvent structure, occurring when the interactions between the solvent molecules and the solute are switched on, and is represented as a linear combination of a set of basis functions located on distinct patches of the molecular surface (SAS) [7]. c_j is a structural constant for each expansion degree j of the Legendre polynomial and h_l is the number of occurrences of a surface patch associated with atom type l. Parameters a_{lj} are determined for each atom type as defined in [9–11]. This atom-typing scheme will be called Crippen types in the following. The free energy contribution due to the formation of the cavity, assumed to be a linear function of both the molecular surface (S) and volume (V) of the solute [12], is accounted for by term 2 and 3.

The parameterization process was performed in two distinct steps [7]. First, parameters a_{lj} have been obtained from linear regression to data generated with the *Grid* program [8], which allows fast evaluations of the 3D-FED. Beneficial to our work is that the recent release provides a more reliable internal force field for the calculation of interaction energies. Subsequently, using the values for a_{lj}, parameters α, β and γ were fitted to experimental log*P* data [7]. The original approach used the second order ($j = 2$) of the Legendre polynomial as shown in equation 1. Due to the large amount of parameters a_{lj}, this expansion could easily lead to overfitting. Therefore, we will restrict ourself to a Legendre polynomial of first order in this publication. It is clear that this restriction will result in a worse model for the training data but hopefully with the benefit of a larger prediction capacity.

Using this reduced Legendre polynomial, the fit to the experimental log*P* values resulted in a bad correlation if term 2 and 3 of equation 1 are included (correlation coefficient $R^2 = 0.868$ and mean square error $\sigma = 0.992$ log units; a plot of the experimental versus the calculated log*P* values for the training set is given in the supporting material). The values of $\beta = 0.0011$ and $\gamma = 0.018$ are very small indicating that these only contribute very little to the correlation. As shown in the result section, term 1 on its own performs even better (it seems that there are some convergence problems in the parameter optimization, so that β and γ are not optimized further towards zero in the three-parameter equation). The differences to the previous publication [7] are a little bit surprising. But, they can partly be explained by the use of the recent release of the *Grid* program, which includes an new force field with an improved entropy term [13] already accounting partly for the cavity formation contributions. Therefore, we propose here a new simplified and improved model for the calculation of log*P*:

$$\log P = \alpha \sum_{l=1}^{120} h_l a_l c = \alpha' \sum_{l=1}^{120} h_l a_l \qquad (2)$$

Here α and c parameters give rise to a single parameter α'. In practice, a continuous representation is achieved by a weighted normalized summation for each surface point s

$$\log P = \frac{\alpha' \sum_s \sum_i f_i g_i(d_{is})}{\sum_i g_i(d_{is})} \qquad (3)$$

where

$$f_i = \frac{a_i}{S_i} S_i = \langle \rho^{FESD} \rangle_i S_i \qquad (4)$$

can be seen as an average FED when divided by the area of the associated surface patch i. The Gaussian-type function $g(d_{is}) = e^{-ad_{is}^2}$ with parameter a depends on the distance d_{is} along the surface between a given surface point s and the reference point of patch i. The optimal value for a has been found to be 2.0 in line with previous observations [6, 7]. This function ensures that atoms lying far from the molecular surface will have a smaller contribution to the overall log*P* than atoms, which are in contact with the molecular surface. In other words, the calculated log*P* depends on the molecular conformations and, in this way,

on molecular surfaces. Although there is no need for a surface-based representation of the parameters derived from the final model, such a representation becomes necessary when partial molecular surfaces are considered.

Parameterization of different solvent systems

The improved parameterization described above has been applied to three solvent systems; namely *n*-octanol/water, alkane/water and cyclohexane/water. For *n*-octanol/water, the training set consists in 400 three-dimensional molecular structures with their corresponding experimental logP_{oct} values. This set of molecules, which has already been used earlier [7], has been chosen for the current study, as we want to outline that the improvement of the parameterization is due to the methodology and not to the nature nor the size of the training set. Noteworthily, a bigger and more appropriate set will be used to establish the final model. As in the previous publication, the conformation of this and the other training and test sets used below are generated using the CORINA program [14]. Using Eq. 2 and the fitting procedure described in the previous section, the resulting correlation between experimental and calculated logP_{oct} values for the training set are shown in Fig. 1. The fit statistics for the training set show that the correlation coefficient $R^2 = 0.94$ and the mean square error $\sigma = 0.66$ log units are, despite the reduced parameter set, very good. The correlation coefficient and the mean square error for the test set, which consisted in an external set of approximately 100 molecules, are respectively $R^2 = 0.85$ and $\sigma = 0.79$ log units (see Fig. 2). These results, which are within range of experimental uncertainty, ensure the validity of the new model. Since the present parameterization uses the same reference data as in the previous work

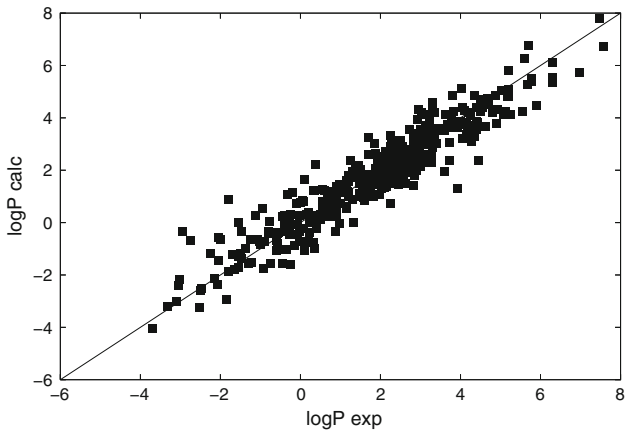

Fig. 1 Calculated logP_{oct} values plotted versus the corresponding experimental values (measured in *n*-octanol/water solvent). The correlation coefficient is $R^2 = 0.94$ and the mean error is $\sigma = 0.66$ log units for the training set

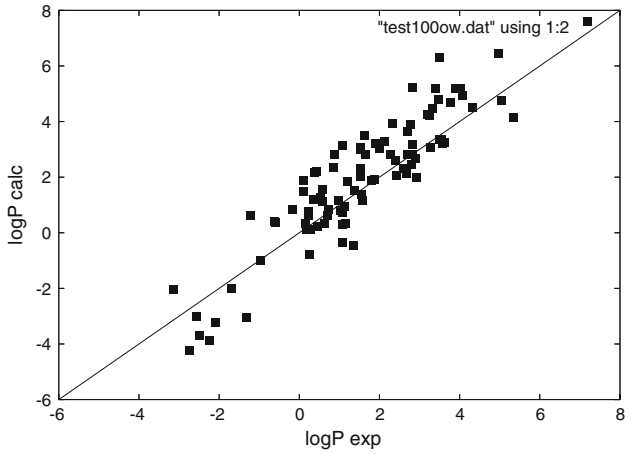

Fig. 2 Calculated logP_{oct} values plotted versus the corresponding experimental values (measured in *n*-octanol/water solvent). The correlation coefficient is $R^2 = 0.85$ and the mean error is $\sigma = 0.79$ log units for the external test set

[7] and less parameters for this model, it is legitimate to assess the better quality of the new model with respect to the former one.

The calculated logP_{oct} values for some molecules of the test set are reported in Table 1. Full lists of all parameters and all molecules are provided as supporting material. One can see the excellent agreement between calculated and experimental values for most of the molecules. However, for some molecules larger discrepancies between predicted and experimental values can be observed. Besides the problems in the choice of the atom-typing scheme and the training set, the conformation dependence of the approach could be a reason for this behavior. The molecules of the training set were chosen as small and rigid. By doing so, we minimized the influence of the conformation on the calculated log*P* values. But for the larger test set molecules, the input molecular conformation can be different from the actual conformation in solution. As described

Table 1 Experimental and calculated logP_{oct} values using Eq. 2 for some molecules of the test set

Compound	Experimental LogP_{oct}	Predicted LogP_{oct}
Pipaneperone	2.02	1.52
Acetone oxime	0.12	0.18
Butyric acid	0.79	1.02
Diphenylether	4.21	3.24
Chalcone	3.08	3.27
Pentopyranose	−3.02	−2.56
Acrylic acid	0.35	0.63
Benzenesulfonamide	1.09	0.31
Hexylcarbamate	1.85	1.21
Quinuclidine	1.38	1.55

above, unlike most fragment based predictive models of log*P*, our model calculates log*P* values that depend on the area and shape of molecular surfaces and in turn, on the corresponding molecular conformations. Therefore, instead of using a single conformation, all possible or at least most stable molecular conformations should be introduced in the calculation of an average log*P*, which can easily be determined by calculating a log*P* value for each conformation (virtual log*P*) weighted by its population. More precisely, the Boltzmann partition function can be used to derive the population of each conformer that is present in solution. For instance, if one considers *N* conformations of the same molecule and $\log P^i$ as the contribution of conformer *i*, the average $\log P^m$ can be obtained from the following equation

$$\log P^m = \frac{\sum_i \exp((E_i - E_0/kT))\log P^i}{\sum_i \exp(E_i - E_0/kT)} \quad (5)$$

where E_i is the total energy of conformer *i*, *k* the Boltzmann constant and *T* the temperature of the system, that is the temperature at which the measurement is done. Unfortunately, despite its consistency, this approach could not be applied in most cases because a remaining assumption is that the conformations are the same in both the water and the hydrophobic solvent. As a consequence, it is not possible to apply this approach for calculating average log*P*s in many practical situations unless one shows that molecular conformers and their populations are similar in both solvents or adapt it to solvent-dependent ensambles, which is out of the scope of this publication. Obviously, the present parameterization of our model and other existing ones, which take into account the molecular geometry, are still valid for small or quite rigid molecules or molecules with a well-defined structure.

Besides *n*-octanol, the most commonly used solvent for the determination of experimental partition coefficients, other solvents are also used for measuring the partition coefficients. Among these solvents, alkane and cyclohexane are sometimes chosen as the apolar solvent in combination with water. On one hand, these solvents are almost totally immiscible with water (compared to octanol, which dissolves up to 25 mol-% water). Therefore, they are, to our opinion, better suited to represent the inner, lipophilic part of a biological membrane. On the other hand, the growing interest in the logarithms of these partition coefficients $\log P_{alk}$ and $\log P_{cyc}$ is partly due to the use in difference measures between these values and $\log P_{oct}$:

$$\Delta\log P_{alk} = \log P_{oct} - \log P_{alk} \quad (6)$$

and

$$\Delta\log P_{cyc} = \log P_{oct} - \log P_{cyc} \quad (7)$$

It is generally admitted that the first difference $\Delta\log P_{alk}$ reflects the hydrogen bonding capacities of the solutes [3, 15], which are related to the ability of a drug to penetrate the skin barrier. This is in line with the fact that these solvents principally differ in their hydrogen bonding abilities; *n*-octanol is both donor and acceptor of hydrogen bonds while alkane and cyclohexane are inert. Compounds with high log*P*, independently from the nature of the partition solvent ($\log P_{oct}$, $\log P_{alk}$ or $\log P_{cyc}$) and low hydrogen capacity can enter the skin membrane. The second difference mentioned above, $\Delta\log P_{cyc}$, has been shown to correlate inversely with log*BB*, the logarithm of the brain/blood partition coefficient, for a series of H_2-receptor antagonists [16].

In Figs. 3 and 4, calculated versus experimental log*P* values are depicted for alkane/water ($\log P_{alk}$) and cyclohexane/water ($\log P_{cyc}$), respectively. Because of the limited amount of experimental data available to us and to

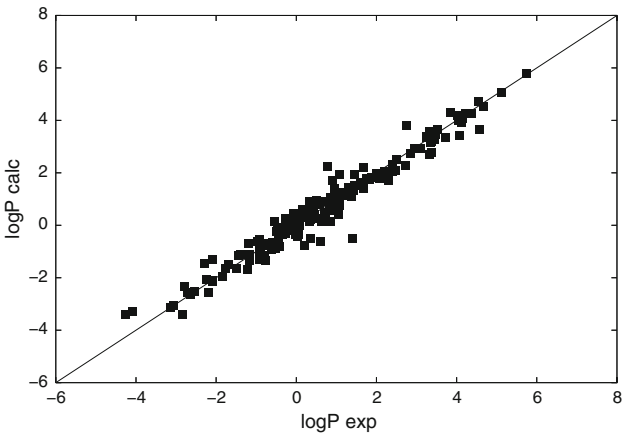

Fig. 3 Calculated $\log P_{alk}$ values plotted versus the corresponding experimental values. The correlation coefficient is $R^2 = 0.98$ and the mean error is $\sigma = 0.40$ log units for the training set

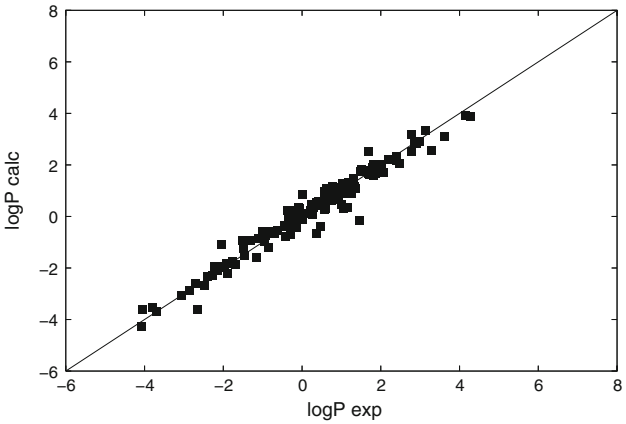

Fig. 4 Calculated $\log P_{cyc}$ values plotted versus the corresponding experimental values. The correlation coefficient is $R^2 = 0.98$ and the mean error is $\sigma = 0.34$ log units for the training set

avoid overfitting, we concentrated us on the most important 50 atom types. Two training sets of 167 molecules and 139 molecules with their experimental logP_{alk} and logP_{cyc} values, respectively, have been used to parameterize Eq. 2. The resulting parameters and calculated log$P$ values are available as supporting material. The correlation coefficients are excellent, $R^2 = 0.98$ for both logP_{alk} and logP_{cyc} with mean errors of 0.4 and 0.34 log units, respectively. The models have then been applied to predict log$P$s of small external sets of molecules (10 molecules for alkane/water and 10 other molecules for cyclohexane/eater) to validate the model. The results for the two test sets are also excellent: $\sigma = 0.47$ log units with $R^2 = 0.92$ for logP_{alk} and $\sigma = 0.49$ log units with $R^2 = 0.91$ for logP_{cyc}. As a second test, we have been able to calculate ΔlogP_{alk} and ΔlogP_{cyc} for a series of compounds, despite the limited number of Crippen types (50), for which a_l parameters (Eq. 2) were obtained. The results are reported in Table 2. In particular, one can see a reasonable agreement between experimental Δlog$P$(exp) and calculated values using the new parameterization.

Prediction of brain/blood barrier partitioning

The brain/blood barrier (BBB) is in essence a mechanism for preventing unnecessary or toxic molecules in the blood stream to penetrate the central nervous system (CNS) and thus plays an important role in drug design. The partitioning of a drug between the brain and blood is often represented by the logarithm of the corresponding partition coefficients logBB = C_{brain}/C_{blood}, where C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively. Experimental measurement of BBB partitioning is tedious and costly and not amenable to high-throughput screening (HTS). As a result, there has been recently a surge in computational efforts to develop new predictive models for BBB partitioning. Unfortunately, the rarity of available experimental data has been the biggest obstacle to the development of such predictive models. Despite the limited predictive capabilities of our models for alkane/water and cyclohexane/water systems due to the incomplete number of atom types (approximately 50 Crippen types for these two solvents), log$P$s have been calculated for a dataset, for which experimental logBB are available. As expected, it was not possible to calculate logP_{alk} and logP_{cyc} for all molecules. Nevertheless, it was possible to calculate log$P$s for about 40 from our set of 80 molecules and strong linear correlations have been found between logBB and both logP_{alk} and logP_{cyc} but not for logP_{oct} (see Fig. 5). The correlation between logBB and logP_{cyc} is in line with the conclusions of Abraham [17, 18] and Young [19, 20], who showed for a series of 20 structurally diverse compounds that logBB does not correlate with logP_{oct}, but does correlate with both logP_{alk} and ΔlogP_{cyc}. The latter correlation was also shown by Abraham et al. [17, 18], thus confirming the use of Δlog$P$ as a good predictor of logBB. One can conclude that these findings support the idea that alkane/water and cyclohexane/water partitioning are good imitations of the brain/blood barrier, while *n*-octanol/water fails to describe this mechanism. In ongoing work, we use our calculated logP_{alk} and logP_{cyc} values in combination with other parameters to come up with a reliable logBB prediction system. We hope that less parameters are needed compared to the work of other groups starting from logP_{oct} [21–25]. Additionally, despite the quality of these results, one might expect that better correlations could be found, if more experimental data had been used for the parameterization of log$P$s (for alkane and cyclohexane). An important effort to collect data for a parameterization of logP_{alk} and logP_{cyc} is currently ongoing.

Table 2 Experimental Δlog$P$(exp) measured in the binary system octanol/cyclohexane and calculated log$P$ for the three solvent systems: octanol/water (oct), alkane/water (alk) and cyclohexane/water (cyc)

Compound	LogP_{oct}(exp)	LogP_{oct}	LogP_{alk}	LogP_{cyc}	ΔlogP(exp)	ΔlogP(calc)
Terfenadine	5.7	5.6	2.1	1.4	3.0	4.1
Azatadine	3.6	3.4	2.3	2.0	1.8	1.4
Epinastine	3.6	3.5	×	×	1.7	×
Imipramine	4.4	4.3	4.4	4.2	0.4	0.1
Chlorpheniramine	3.2	3.3	3.7	1.9	1.1	1.4
Diphenhydramine	3.4	3.8	3.9	3.8	0.4	0.0
Mepyramine	3.1	3.3	×	×	0.9	×
Hydroxyzine	3.1	3.9	2.8	2.8	1.9	1.1
Dimethindene	2.8	3.4	×	×	1.1	×

The calculated Δlog$P$ for the binary system octanol/cyclohexane (oc) are given in the last column

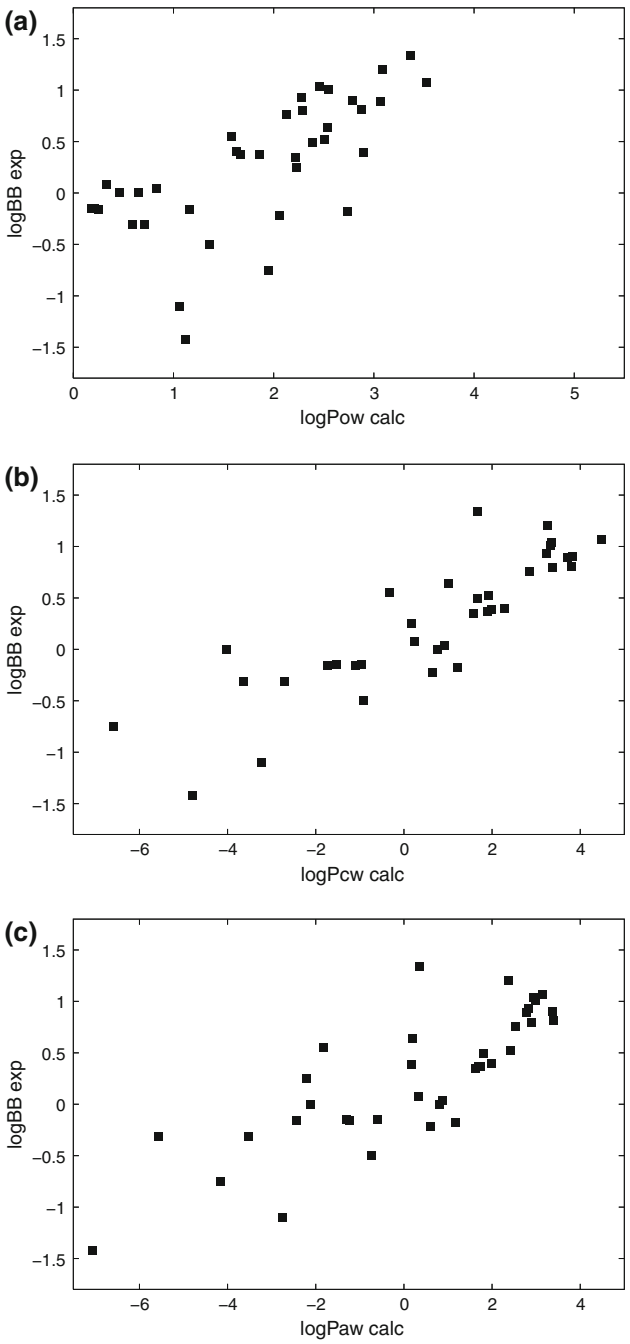

Fig. 5 Correlation between experimental log*BB* and logP_{oct}, logP_{cyc} and logP_{alk}, respectively. One can see that no significant correlation between logP_{oct} and log*BB* has been found ($R^2 = 0.23$ and $\sigma = 0.40$). (**a**) In contrast, better correlations have been found for logP_{cyc} and logP_{alk} and log*BB*. The best model equations have been found to be $\log BB = 0.087 + 0.161 \log P_{cyc}$ and $\log BB = 0.152 + 0.161 \log P_{alk}$. The correlation coefficients are $R^2 = 0.78$ and $\sigma = 0.46$ for cyclohexane/water (**b**) and $R^2 = 0.78$ and $\sigma = 0.49$ for alkane/water (**c**)

Conclusion

An improved parameterization of a previously introduced concept [5–7] was described to allow reliable predictions of log*P* for three different partition solvent systems: *n*-octanol/water, alkane/water and cyclohexane/water. The application of this new model necessitated the evaluation of the volume energy density (3D-FED) using the molecular interaction field implemented in the *Grid* program [8]. This information was used with the experimental log*P* data to determine the parameters of our final model. The physical basis of our surface-based model and its ability to determine local contributions to overall log*P*s by projecting them onto the molecular surface provides a powerful tool for designing molecules, for which undesirable fragments can be substituted by other fragments. The strong linear correlations found between the logarithms of the brain/blood barrier (BBB) and the predicted logP_{alk} and logP_{cyc} is an important step towards the development of highly predictive tools for log*BB*. Most important, the advantage of our approach is that it becomes possible to predict both logP_{alk} and logP_{cyc} and in turn log*BB* without the need of experimental data. Unfortunately, an obstacle resides in the scarcity of experimental data available to us, which are necessary for the parameterization and the derivation of more reliable values for the parameters used in our models. More molecules with their measured log*P* are required for especially the two solvents alkane and cyclohexane. In this sense, we are on the way to collect bigger training sets (e.g. about 4000 molecules for logP_{oct} and more, whenever new data becomes available) with high chemical diversity, hopefully covering all Crippen types, to derive the updated parameters.

Supporting material

The training and test sets with their experimental and calculated log*P* values as well as the new parameters used in the calculation presented here are available as supporting material from the publisher's web site.

References

1. Leo A, Hansch C, Elkins D (1971) Chem Rev 6:71
2. Caron G, Ermondi G (2005) J Med Chem 48:3269
3. El Tayar N, Tsai RS, Testa B, Carrupt PA, Leo A (1991) J Pharm Sci 80:590
4. Abraham MH, Chadha HS, Whiting GS, Mitchell RC (1994) J Pharm Sci 83:1085
5. Pixner P, Heiden W, Merx H, Möller A, Moeckel G, Brickmann J (1994) J Chem Inf Comput Sci 34:1309–1319
6. Jäger R, Schmidt F, Schilling B, Brickmann J (2000) J Computer-Aided Mol Des 14:631–646
7. Jäger R, Kast SM, Brickmann J (2003) J Chem Inf Comput Sci 43:237–247
8. Goodford PJ (1996) Chemometrics 10:107
9. Ghose AK, Crippen GM (1986) J Comp Chem 7:565–577

10. Viswanadan VN, Ghose AK, Revankar GR, Robins RK (1989) J Chem Inf Comput Sci 29:163–172
11. Ghose AK, Viswanadan VN, Wendoloski JJ (1998) J Phys Chem A 102:3762–3772
12. No KT, Kim SG, Cho KH, Scheraga HA (1999) Biophys Chem 78:127
13. Cruciani G (2006) Molecular interaction fields. Wiley-VCH
14. Gasteiger J, Rudolph C, Sadowski J (1992) Tetrahedron Comput Meth 3:537
15. El Tayar N, Tsai RS, Testa B, Carrupt PA, Hansch C, Leo A (1991) J Pharm Sci 80:744
16 Ganellin CR (1991) Uses of partition coefficients by brain penetration applied to the design of H_2-receptor histamine antagonists. Elsevier
17. Abraham MH, Chadha HS, Mitchell RC (1994) J Pharm Sci 83:1257
18. Chadha HS, Abraham MH, Mitchell RC (1994) Bioorg Med Chem Lett 2511:1085
19. Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffiths R, Jones M, Rana KK, Saunders D, Smith IR, Sore NE, Wilks TJ (1988) J Med Chem 31:656
20. Young RC, Ganellin CR, Griffiths R, Mitchell RC, Parsons ME, Saunders D, Sore NE (1993) J Med Chem 28:201
21. Wichmann K, Diedenhofen M, Klamt A (2007) J Chem Inf Model 47:228
22. Platts JA, Abraham MH, Zhao YH, Hersey A, Ijaz L, Butima D (2001) Eur J Med Chem 36:719
23. Subramanian G, Kitchen DB (2003) J Computer-Aided Mol Des 17:643
24. Fu X, Song Z, Liang W (2005) Internet Electron J Mol Des 4:737
25. Rishton GM, LaBonte K, Williams AJ, Kassam K, Kolovanov E (2006) Curr Opin Drug Dis Dev 9:303