Journal of Molecular Graphics and Modelling 29 (2011) 985–995

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms

Hongming Chen[a,*], Susanne Winiwarter[b], Markus Fridén[b,c], Madeleine Antonsson[b], Ola Engkvist[a]

[a] DECS GCS Computational Chemistry, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden
[b] Discovery DMPK, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden
[c] Department of Pharmaceutical Bioscience, Division of Pharmacokinetics and Drug Therapy, Uppsala University, Box 591, SE-75124 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 21 February 2011
Received in revised form 8 April 2011
Accepted 12 April 2011
Available online 27 April 2011

Keywords:
Quantitative structure–activity relationship (QSAR)
Unbound brain-to-plasma concentration ratio
Total brain-to-plasma concentration ratio
Random forest (RF)
Support vector machine (SVM)
Partial least squares (PLS)

ABSTRACT

Distribution over the blood–brain barrier (BBB) is an important parameter to consider for compounds that will be synthesized in a drug discovery project. Drugs that aim at targets in the central nervous system (CNS) must pass the BBB. In contrast, drugs that act peripherally are often optimised to minimize the risk of CNS side effects by restricting their potential to reach the brain. Historically, most prediction methods have focused on the total compound distribution between the blood plasma and the brain. However, recently it has been proposed that the unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$) is more relevant. In the current study, quantitative $K_{p,uu,brain}$ prediction models have been built on a set of 173 in-house compounds by using various machine learning algorithms. The best model was shown to be reasonably predictive for the test set of 73 compounds ($R^2 = 0.58$). When used for qualitative prediction the model shows an accuracy of 0.85 (Kappa = 0.68). An additional external test set containing 111 marketed CNS active drugs was also classified with the model and 89% of these drugs were correctly predicted as having high brain exposure.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

The blood–brain barrier (BBB) constitutes the primary system to protect the brain from exposure to potentially hazardous xenobiotics. The most important physical structure of the BBB is the brain capillary endothelium, a very tight membrane that hinders paracellular permeation. Transcellular permeation is restricted by the high levels of efflux transporters present in the endothelial cells, like P-glycoprotein (Pgp) and multi-drug resistance protein (MRP) transporters. Thus, the BBB is well suited to protect the brain against xenobiotics. This defence mechanism can be utilised by designing peripherally acting drugs with low risk of central side effects. However, for drugs targeting proteins in the central nervous system (CNS) brain exposure may be the biggest hurdle to overcome in the drug discovery process [1].

The two main mechanisms which determine the level of drug exposure in the brain in relation to the exposure in the blood are passive diffusion and active influx/efflux by various transporters. Structure–brain exposure relationships have for many years been derived from data of the total brain-to-plasma concentration ratio, $K_{p,brain}$ [2], often expressed in its logarithmic form (also known as log BB) [3]:

$$K_{p,brain} = \frac{A_{brain}}{C_p} \quad (1)$$

Here, A_{brain} is the total compound concentration in the brain and C_p is the total compound concentration in the plasma. The pioneering study by Young et al. [4] showed a good correlation between log BB and $\Delta \log P$ (here defined as the difference between the log P for octanol/water and the log P for cyclohexane/water, the latter is a descriptor of the hydrogen bonding capacity) for 20 anti-histamine compounds. Waterbeemd et al. showed that the calculated polar surface area (PSA) is an almost as good descriptor for log BB as the experimentally measured $\Delta \log P$. However, PSA is much easier to obtain, since it is an *in silico* descriptor [5]. Clark et al. [6] built a QSAR model that related log BB to physicochemical descriptors including PSA and C log P for a set of 55 compounds. Ooms et al. correlated log BB data with descriptors based on the 3-dimensional molecular field of 83 compounds [7]. Very recently, Fan et al. [8] reported a log BB multiple linear regression model which uses a genetic algorithm for descriptor selection. Engkvist et al. developed a classification model by using CNS and non-CNS active drugs as training set instead of BBB distribution data [9].

In recent years, it has been observed [10,11] that $K_{p,brain}$ or log BB may not reflect the relevant drug exposure in the brain since it is

* Corresponding author. Tel.: +46 31 7065285.
E-mail address: Hongming.chen@astrazeneca.com (H. Chen).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.04.004

highly influenced by the relative binding affinity of compounds to plasma proteins and brain tissue [3]. It was highlighted that the pharmacological efficacy of a drug in the CNS is not dependent on the total but on the free drug concentration in the brain [3,12]. One suggestion has been to replace log BB by the BBB permeability-surface area (PS) product [10,13], an estimate of the net influx clearance. Abraham [14] reported a good relationship between log PS and solvation parameters for 30 neutral compounds, and, around the same time, Liu et al. [15] developed predictive log PS models based on the TPSA (topological polar surface area) and log *D*. However, it has been argued that the PS product cannot predict the unbound drug concentration in CNS by itself, since the drug concentration is equally influenced by the BBB efflux clearance [3].

The most relevant parameter for predicting free drug exposure is the steady-state unbound brain-to-plasma concentration ratio $K_{p,uu,brain}$. However, until recently there has been a lack of experimental techniques to measure this parameter in an efficient manner [16–18]. $K_{p,uu,brain}$ is defined by the following equation [18]:

$$K_{p,uu,brain} = \frac{C_{u,brainISF}}{C_{u,p}} \quad (2)$$

$C_{u,brainISF}$ is the concentration of unbound compound in the brain interstitial fluid and $C_{u,p}$ is the concentration of unbound compound in the blood plasma. Experimentally, the value can be determined from three different experiments, measuring the total brain and plasma concentrations in an *in vivo* animal experiment and combining these values with determinations of plasma protein binding and binding to brain tissue *in vitro*. Mechanistically, $K_{p,uu,brain}$ is determined by the relative efficiency of BBB influx and efflux and is independent of plasma protein binding in blood and binding to the tissue component of brain. To the best of our knowledge there is currently only one literature report of *in silico* prediction for $K_{p,uu,brain}$; Fridén et al. [18] measured $K_{p,uu,brain}$ values in rat for 41 marketed drugs and reported a QSAR $K_{p,uu,brain}$ model with moderate predictive power.

In the present study, we extended the original $K_{p,uu,brain}$ data set [18] by adding a set of in-house compounds with experimental $K_{p,uu,brain}$ data. This extended dataset was used to build regression models with the random forest (RF) and support vector machine (SVM) algorithms, respectively. Direct and indirect regression models for $K_{p,uu,brain}$ were built, combined into consensus models and evaluated. The models give reasonable continuous predictions and good classification predictions on the test set. Good classification results were also obtained for an additional, external test set of 111 marketed CNS active drugs [19].

2. Experimental

2.1. $K_{p,uu,brain}$ determination

$K_{p,uu,brain}$ is defined as the ratio of the unbound drug concentration in the brain interstitial fluid and the unbound drug concentration in the plasma. The details of the experimental procedure have been described previously [18]. Briefly, $K_{p,uu,brain}$ is calculated by combining the total brain-to-plasma ratio $K_{p,brain}$ determined *in vivo* (Eq. (1)) with estimates of $V_{u,brain}$ and $f_{u,p}$ determined *in vitro* in brain slices [20] and by equilibrium dialysis [21], respectively (Eq. (3)):

$$K_{p,uu,brain} = \frac{K_{p,brain}}{V_{u,brain} \times f_{u,p}} \quad (3)$$

Here, $V_{u,brain}$ represents the unbound volume of distribution in the brain and $f_{u,p}$ is the unbound fraction of drug in plasma. Cassettes of up to three drugs were administered to Sprague-Dawley rats by 4 h constant rate intravenous infusions. Terminal sampling of blood and brain tissue was done under isoflurane anaesthesia. The total drug concentration in the brain (A_{brain}) was corrected for drug in the residual blood by subtracting 0.8% of the drug plasma concentration, an approximation for a more elaborate correction model [22].

2.2. Model building workflow

In the current study, two types of $K_{p,uu,brain}$ models were built: direct models based on the available $K_{p,uu,brain}$ data and indirect models obtained by combining individual models of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ data according to Eq. (3) (Fig. 1). The rationale for developing indirect models was that more experimental data is available for each of $K_{p,brain}$, $V_{u,brain}$ and $f_{u,p}$ than for $K_{p,uu,brain}$. Thus, the indirect models were built on larger training sets. Consensus models were built by averaging the predictions for two or more individual $K_{p,uu,brain}$ models. All models were built on the logarithmic value, i.e., $\log K_{p,uu,brain}$, $\log K_{p,brain}$, $\log V_{u,brain}$ and $\log f_{u,p}$. Accordingly, the obtained RMSE (root mean squared error) values also refer to logarithmic units.

2.3. Data sets

The $K_{p,uu,brain}$ data set consists of 246 compounds for which $f_{u,p}$, $K_{p,brain}$ and $V_{u,brain}$ values were available. Additionally, in-house compounds with either $f_{u,p}$, $K_{p,brain}$ or $V_{u,brain}$ values measured were collected. The $f_{u,p}$, $K_{p,brain}$, and $V_{u,brain}$ data sets contain 3234, 505 and 472 compounds, respectively. The 41 marketed drugs from Fridén's data set [18] are included in the present $K_{p,uu,brain}$ data set. The remaining compounds originate from various in-house drug discovery projects. 30% (73 compounds) were randomly selected as the test set. The remaining 173 compounds form the training set and were used to build the direct $K_{p,uu,brain}$ models. For the indirect models, the same test set of 73 compounds was used and therefore excluded from the respective training set. Thus, the training sets for the $f_{u,p}$, $K_{p,brain}$ and $V_{u,brain}$ models consist of 3161, 432 and 399 compounds, respectively. Additionally, 111 marketed CNS active drugs were collected from the literature [19] and used as an external test set.

A set of 196 2D and 3D descriptors, including descriptors for molecular size, lipophilicity, hydrogen bonding properties, electrostatics and topology, were calculated with an in-house program. The calculated descriptors are described elsewhere [23–25]. For the 41 marketed drugs used in the current study, chemical structures, experimental data and calculated descriptors are available as Supplementary Data. The structures and descriptors for the 111 CNS active drugs used as an independent test set is also provided as Supplementary Data.

2.4. Modelling methods

Two non-linear machine learning methods, support vector machine (SVM) and random forest (RF), were used to build the models. They are both implemented in the in-house machine learning package AZOrange [26], an extension of the open source package Orange [27]. For comparison, linear models based on the partial least squares (PLS) algorithm as implemented in AZOrange were also investigated. Principal component analysis (PCA) was carried out in SIMCA [28]. Further statistical analyses were done in JMP [29].

2.4.1. Support vector machine (SVM)

In SVM learning [30,31] a hyperplane is constructed, which discriminates between the data points of distinct classes (binary SVM) in such a way that the margin between the classes is maximized. The final position and orientation of the hyperplane is defined by a

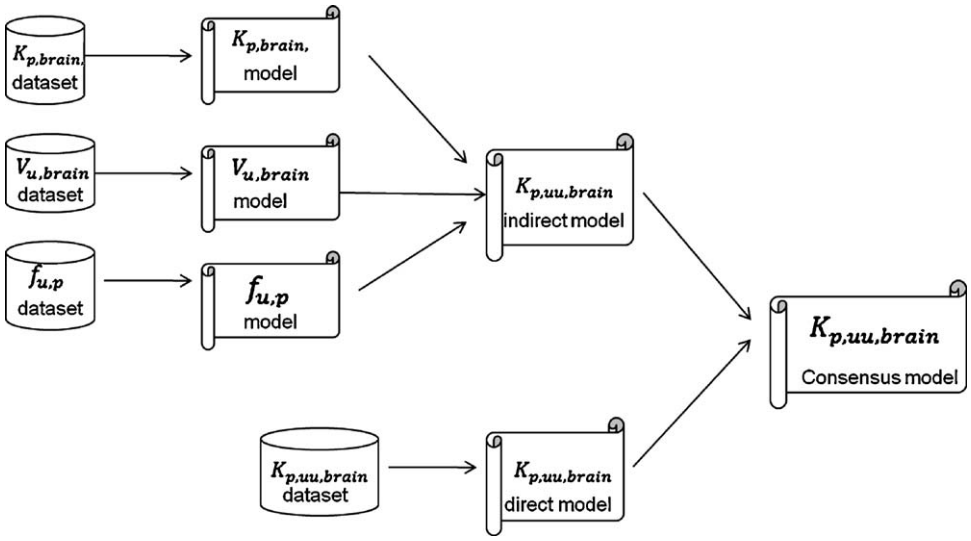

Fig. 1. The workflow applied for the model building.

Table 1
The classification performance measures used.

Measure	Formulae[a]	Description
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Probability to correctly classify compounds
Sensitivity (recall)	$\frac{TP}{TP+FN}$	Probability to predict positive when true class is positive
Specificity	$\frac{TN}{FP+TN}$	Probability to predict negative when true class is negative
Positive precision	$\frac{TP}{TP+FP}$	Probability to correctly classify compounds predicted to be positive
Negative precision	$\frac{TN}{FN+TN}$	Probability to correctly classify compounds predicted to be negative
F-score	$\frac{2\times Sensitivity\times Pos.precision}{Sensitivity+Pos.precision}$	Harmonic average of positive precision and sensitivity
Kappa	$\frac{Accuracy-Cp}{1-Cp}$, where $Cp=\frac{(TP+FP)\times(TP+FN)+(TN+FN)\times(TN+FP)}{(TP+FP+FN+TN)^2}$	True accuracy which is corrected by probability to obtain agreement by chance
Matthews correlation coefficient	$\frac{TP\times TN-FN\times FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TP+FP)}}$	Correlation coefficient between the observed and predicted binary classifications

[a] TP: true positive, FP: false positive, TN: true negative, FN: false negative.

subset of training vectors, the so-called support vectors. The commonly used Gaussian radial basis function kernel was also used in the current study. The optimal Gamma and *C* parameters obtained from a grid search for the best cross validation accuracy were used in the final SVM models. For the final SVM direct $K_{p,uu,brain}$ model, the Gamma value and *C* parameter were set to 0.03 and 32.0, respectively.

2.4.2. Random Forrest (RF) method

This method [32,33] combines a number of tree predictors each of which is built independently from the others. To classify a new object each tree gives a classification or vote for a class. The classification with the most votes is the one proposed. The main advantage of this approach is the possibility of handling thousands of input variables with a low risk for over-training and a new classification is very fast. In this study, the number of trees is set to 100. We have also investigated the influence of the number of trees on the prediction accuracy. The number of trees was increased up to 1000; however, the prediction accuracy did not improve significantly.

2.5. Model validation

The correlation coefficient R^2, RMSE and the cross-validated correlation coefficient Q^2 obtained by a 10 fold cross validation process were assessed for all the models. Q^2 is used as an internal measure of model predictivity. *y* permutation tests [34] were performed to show the difference between the model built on true data and the model built on random data. The *y* ($K_{p,uu,brain}$ or $K_{p,brain}$) value was randomly assigned to a compound in the training set before model building. The scrambling process was repeated 5 times and Q^2 values for the randomized models were calculated. External predictivity was assessed by predicting the test set compounds. Both R^2 and RMSE values were calculated.

The non-parametric Spearman rank correlation coefficient ρ [35] was calculated to evaluate the ranking power of the models: this coefficient is a measure of the extent of the statistical dependence between pairs of observations and can thereby estimate the models' capability of ranking compounds. The sign of the Spearman correlation indicates the direction of the association between *X* (the independent variable) and *Y* (the dependent variable). If *Y* increases when *X* increases, the Spearman correlation coefficient ρ is positive and accordingly, if *Y* decreases when *X* increases, the Spearman correlation coefficient ρ is negative. The Spearman correlation coefficient is zero, if there is no tendency for *Y* to increase or decrease when *X* increases. When *X* and *Y* are perfectly monotonically correlated, the Spearman correlation coefficient ρ is 1.0.

Additionally, several classification performance measurements based on the confusion matrix were generated to assess the quality of the classification (Table 1).

3. Results and discussion

3.1. Analysis of data

The data distributions for all four data sets are shown in Fig. 2. For the $K_{p,uu,brain}$ data set, the mean log $K_{p,uu,brain}$ is −1.1 with a stan-

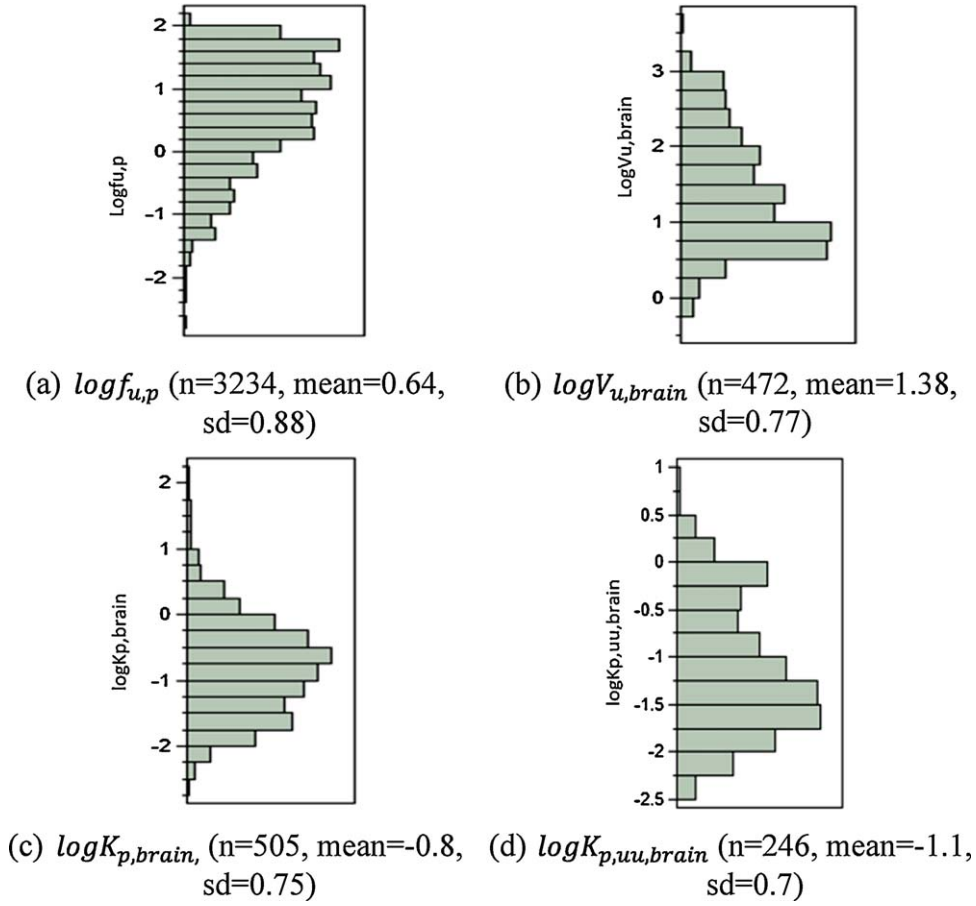

Fig. 2. Distribution of experimental values: (a) $f_{u,p}$, (b) $V_{u,brain}$, (c) $K_{p,brain}$ and (d) $K_{p,uu,brain}$.

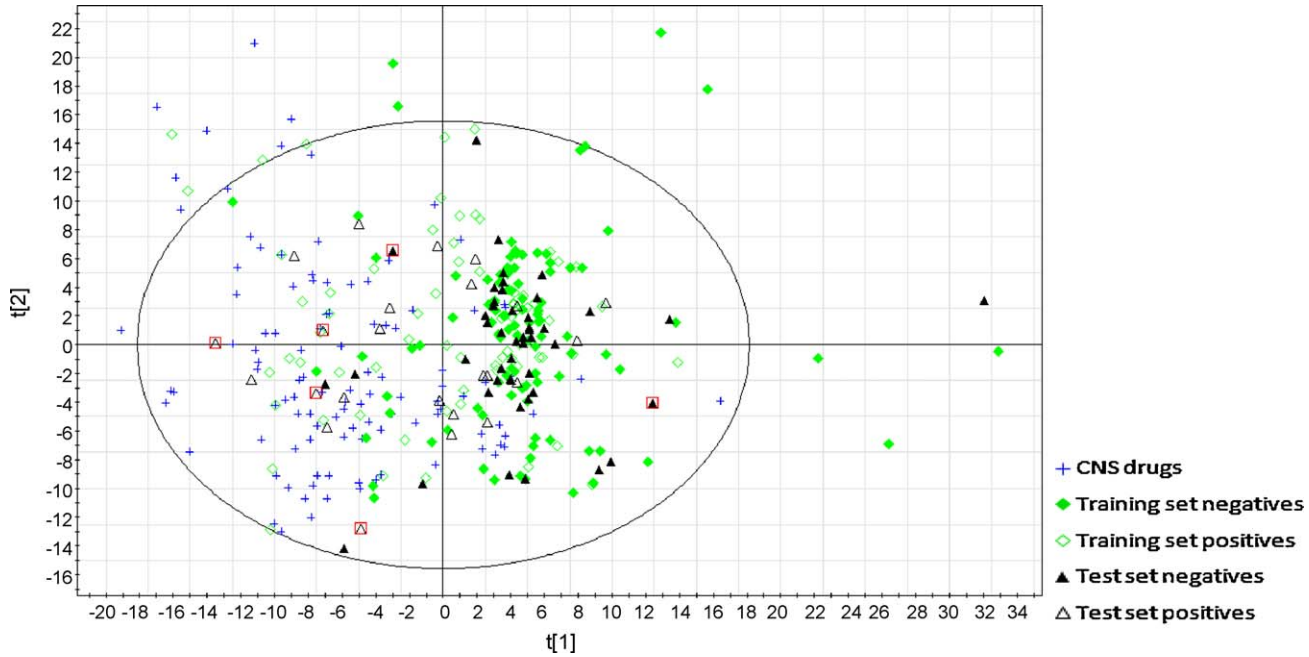

Fig. 3. Score plot from the principal component analysis (PCA) of the $K_{p,uu,brain}$ and CNS drug sets. *X* and *Y* axes correspond to the two most important principal components from PCA analysis. The BBB positive compounds (empty symbols) have a $K_{p,uu,brain}$ larger than or equal to 0.1 and, accordingly, the BBB negative compounds (filled symbols) have a $K_{p,uu,brain}$ less than 0.1. The compounds labelled with red squares are the examples listed in Fig. 8. The CNS-active drugs are labelled with blue crosses.

dard deviation of 0.7. The predominance of negative log $K_{p,uu,brain}$ values is consistent with our previous study which was based on a representative selection from the chemical drug space [18].

The $K_{p,uu,brain}$ data set (n = 246) was randomly split into a training and test set with the ratio of 7:3. A PCA analysis was done for the $K_{p,uu,brain}$ training and test sets and the CNS active drug set based on the 196 descriptor set. The first two principal components for these three sets are plotted in Fig. 3. Test and training sets are well distributed in this plot. Only a minority of compounds is outside the 95% confidence interval. The CNS active drugs are mostly located on the left hand side of the plot as are the compounds that were classified as BBB positive ($K_{p,uu,brain}$ > 0.1).

The data sets contain acids, bases and neutral compounds. The influence of ionization state on $K_{p,uu,brain}$ and $K_{p,brain}$ is shown in Fig. 4. It can be seen that basic compounds tend to have higher $K_{p,brain}$ than acidic compounds. An opposite, however, much smaller trend is seen for $K_{p,uu,brain}$. These findings are in line with our understanding of the determinants of $K_{p,brain}$ and $K_{p,uu,brain}$, respectively. $K_{p,brain}$ is mechanistically dependent of $K_{p,uu,brain}$, $f_{u,p}$ and $V_{u,brain}$. Thus, the higher $K_{p,brain}$ values for bases can be expected from the fact that basic compounds have high affinity to brain tissue phospholipids (high $V_{u,brain}$) and less affinity to plasma proteins (high $f_{u,p}$). Acidic compounds on the other hand are known to bind to albumin in the plasma (low $f_{u,p}$) and the negative charge at physiological pH reduces the partitioning into tissue phospholipids (low $V_{u,brain}$). As shown in Eq. (3), $K_{p,uu,brain}$ is calculated from the three directly measurable components $K_{p,brain}$, $f_{u,p}$ and $V_{u,brain}$. However, mechanistically $K_{p,uu,brain}$ is not dependent on any of these components. As discussed above, $K_{p,uu,brain}$ is solely determined by events at the BBB, i.e., active and passive influx and efflux. The reason for the small difference in $K_{p,uu,brain}$ for basic and acidic compounds is therefore less obvious. One possible explanation is that P-glycoprotein, the most important efflux transporter for the rodent BBB, preferentially transports basic and neutral compounds.

3.2. Model building

Direct and indirect $K_{p,uu,brain}$ models were built with both the SVM and the RF algorithms. Direct models were built based on the 173 compounds in the training set. For indirect models, individual $K_{p,brain}$, (n = 432), $V_{u,brain}$ (n = 399) and $f_{u,p}$ (n = 3161) models were built. The predictions from these three models were then used as input to Eq. (3) to calculate the $K_{p,uu,brain}$ values. The results for the training set for the indirect and direct $K_{p,uu,brain}$ models are shown in Table 2. The cross-validated correlation coefficient Q^2 is given for each model together with R^2 and RMSE. The Q^2 values for the $f_{u,p}$ and $V_{u,brain}$ models (both SVM and RF) are both above 0.7. Thus the models have good internal predictive power. The Q^2 values for the $K_{p,brain}$ SVM and RF models are somewhat lower, 0.5 and 0.6, respectively. The Q^2 values for the direct $K_{p,uu,brain}$ model built on the 173 training compounds are 0.6 for the SVM and 0.4 for the RF model (Table 2). An additional validation for the RF model was done by performing y permutation tests [33]. The average Q^2 values for the y-randomized $K_{p,uu,brain}$ and $K_{p,brain}$ models were both around −0.1, which implies that the models built on the true data is based on the actual variance of the data and not just an accidental correlation. The PLS method was only used to build a direct model and the resulting Q^2 value is 0.23 which is significantly lower than the values obtained with the nonlinear models.

3.3. Model interpretation

For a better model understanding the variable importance (VIP) was investigated for the RF models. Variable importance estimation is not available for the SVM models. The VIP values of the top 10 descriptors for the $K_{p,brain}$ and $K_{p,uu,brain}$ RF models are listed in Tables 3 and 4. Among the top 10 descriptors in the $K_{p,brain}$ model (see Table 3), there are some descriptors related to hydrogen bonding such as number of hydrogen bond acceptors (HBA_Raevsky [36]), polar surface area (SAS_POL_AREA, SPEC_SAS_POL_AREA) and number of polar atoms (PAT). This is consistent with the reported log BB models [4–6,37], where hydrogen bonding, as described by e.g. PSA, is the main determining factor for distribution across the BBB. Lipophilicity, calculated as log P (ACD log P), has also a big influence on $K_{p,brain}$, again in line with previous studies [6,38–42]. Other descriptors of importance for the model are related to the 3D electrostatic potential and the molecular charge distribution (VDW_EP_N_MEAN, VDW_EP_P_VAR, VDW_EP_P_SUM, etc.) [24]. Fig. 5a–c shows the relationship between log $K_{p,brain}$ and the top three descriptors (SAS_POL_AREA, ACD log P and PAT). The descriptor values are first binned and then plotted against the median log $K_{p,brain}$. Bins with less than three compounds are omitted in the figures. Increasing the molecular polar surface area and number of polar atoms will significantly decrease $K_{p,brain}$, while more lipophilic compounds tend to have higher $K_{p,brain}$ values. Fig. 5d–f shows the influence of the top three descriptors (VDW_EP_P_SUM, Kappa2 and HBD) from the $K_{p,uu,brain}$ model on $K_{p,brain}$. In general, larger VDW_EP_P_SUM values result in smaller $K_{p,brain}$ values. However, this trend does not apply for compounds with the largest VDW_EP_P_SUM (>16). The relationships between $K_{p,brain}$ and Kappa2 and HBD are not monotonic.

The top 10 descriptors for the $K_{p,uu,brain}$ model are different from those of the $K_{p,brain}$ model (see Table 4). Again we see hydrogen bonding descriptors in this list (HBD, HBAsum, and HBAmax). Additionally, descriptors related to the electrostatic potential (VDW_EP_P_SUM, SAS_EP_N_MEAN, VDW_EP_N_MEAN) and topology descriptors (Kappa2, Kappa3 [43,44]) are important for the model. Kappa2 [43] is a molecular topology based shape index, calculated from two-bond fragments, which gives an estimation of the linearity of a molecule. The influence of the top three descriptors (VDW_EP_P_SUM, Kappa2 and HBD) on $K_{p,uu,brain}$ is shown in Fig. 6a–c. Increasing VDW_EP_P_SUM decreases $K_{p,uu,brain}$. Compounds with Kappa2 less than 8.0 have significantly higher $K_{p,uu,brain}$ than compounds with Kappa2 above 8.0. The relationship between HBD and $K_{p,uu,brain}$ is not monotonic. The highest median $K_{p,uu,brain}$ value is found for compounds with two hydrogen bond donors. The relationship between $K_{p,uu,brain}$ and the three most important descriptors for the $K_{p,brain}$ model are shown in Fig. 6d–f. PSA related descriptors (SAS_POL_AREA, PAT) have a similar influence on $K_{p,uu,brain}$ as for $K_{p,brain}$. However, there is no correlation between log P (ACD log P) and $K_{p,uu,brain}$. This observation agrees with earlier findings in a smaller dataset [18] and is likely due to the fact that $K_{p,uu,brain}$ is mainly determined by transporter interactions. In contrast, $K_{p,brain}$ is highly correlated to unspecific tissue binding where lipophilicity is a major factor [46].

3.4. Model validation

The results for the test set (n = 73) are given in Table 2 and the R^2, RMSE and Spearman's ρ are listed for all models developed. For the SVM and RF $f_{u,p}$ models, R^2 is 0.69 and 0.62, respectively. The predictive power of the $V_{u,brain}$ models (SVM and RF) is similar (R^2 = 0.67 and 0.6, respectively). R^2 for the $K_{p,brain}$ models are 0.55 (SVM) and 0.57 (RF). The R^2 for the indirect $K_{p,uu,brain}$ models are 0.42 and 0.51 for the SVM and RF models, respectively. The models predict $K_{p,uu,brain}$ based on Eq. (3) using the individual $K_{p,brain}$ $f_{u,p}$ and $V_{u,brain}$ model predictions. The performance for the $K_{p,uu,brain}$ direct models on the test set is slightly better. The R^2 for the SVM $K_{p,uu,brain}$ direct model (SVMd_$K_{p,uu,brain}$) is 0.53 and 0.52

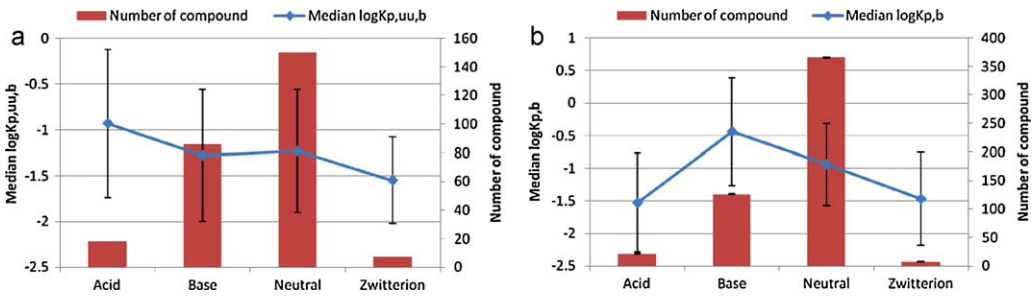

Fig. 4. The influence of the ionization state on (a) $K_{p,uu,brain}$ and (b) $K_{p,brain}$. The error bars here refer to the standard deviation.

Table 2
The model performance for the training and test sets.

	Models[a]	Training set			Test set		
		R^2	RMSE	[b]Q^2	[c]R^2	RMSE	Spearman's ρ
Base models	RF_$V_{u,brain}$	0.97	0.14	0.7	0.6	0.39	0.75
	SVM_$V_{u,brain}$	0.96	0.16	0.8	0.67	0.36	0.73
	RF_$f_{u,p}$	0.93	0.23	0.8	0.62	0.36	0.75
	SVM_$f_{u,p}$	0.88	0.3	0.8	0.69	0.32	0.81
	RF_$K_{p,brain}$	0.94	0.18	0.5	0.57	0.58	0.76
	SVM_$K_{p,brain}$	0.83	0.3	0.6	0.55	0.59	0.73
Indirect models	RFi_$K_{p,uu,brain}$	0.9	0.22	NA	0.51	0.49	0.73
	SVMi_$K_{p,uu,brain}$	0.79	0.32	NA	0.42	0.54	0.68
Direct models	RFd_$K_{p,uu,brain}$	0.94	0.17	0.4	0.52	0.49	0.71
	SVMd_$K_{p,uu,brain}$	0.96	0.15	0.6	0.53	0.48	0.73
	PLSd_$K_{p,uu,brain}$	0.77	0.33	0.23	0.41	0.54	0.63
Consensus models	RFdi_$K_{p,uu,brain}$	NA	NA	NA	0.57	0.46	0.76
	SVMdi_$K_{p,uu,brain}$	NA	NA	NA	0.5	0.5	0.73
	SVMd_RFi_$K_{p,uu,brain}$	NA	NA	NA	0.57	0.47	0.77
	SVMd_RFd_$K_{p,uu,brain}$	NA	NA	NA	0.56	0.47	0.74
	SVMd_RFd_RFi_$K_{p,uu,brain}$	NA	NA	NA	0.58	0.46	0.77
Random permutation models	Permute_RF_$K_{p,uu,brain}$	NA	NA	−0.09	NA	NA	NA
	Permute_RF_$K_{p,brain}$	NA	NA	−0.11	NA	NA	NA

SVMi: SVM indirect model, RFi: random forest indirect model. SVMd: SVM direct model, RFd: random forest direct model, PLSd: PLS direct model; RFdi: consensus model based on RF direct and RF indirect models, SVMdi: consensus model based on SVM direct and SVM indirect models, SVMd_RFi: consensus model based on SVM direct and RF indirect models, SVMd_RFd: consensus model based on SVM direct and RF direct models, SVMd_RFd_RFi: consensus model based on SVM direct, RF direct and RF indirect models.

[a] Model annotations are the same as in Table 5.
[b] Q^2 for 10 fold cross-validation of the training set.
[c] Corresponds to test set.

Table 3
VIP of the top 10 descriptors for the $K_{p,brain}$ RF model.

Descriptor name	Weight	Meaning of descriptors
SAS_POL_AREA	0.0135	Solvent accessible surface (SAS) polar area
ACD log P	0.0107	Predicted log P
PAT	0.0107	Number of polar atoms (O, N, S, P)
VDW_EP_N_MEAN	0.0102	Mean of negative electrostatic potential on VDW surface
VDW_EP_P_VAR	0.0093	Variance of positive electrostatic potentials on Van der Waals surface
SPEC_SAS_POL_AREA	0.0086	Percentage of polar SAS area in total SAS area
VDW_EP_P_SUM	0.0085	Sum of positive electrostatic potentials on VDW surface
HBA_Raevsky	0.0084	Number of hydrogen bond acceptors according to Raevsky (HYBOT) [36]
SPEC_VDW_POL_AREA	0.0075	Polar fraction of Van der Waal surface area
PIAT	0.0074	Number of pi atoms (number of atoms linked to double bonds + number of halogen atoms).

Table 4
VIP of the top 10 descriptors for the $K_{p,uu,brain}$ RF model.

Descriptor name	Weight	Meaning of descriptor
Kappa2	0.0113	Topology index [43,44]
VDW_EP_P_SUM	0.0105	Sum of positive electrostatic potentials on VDW surface
HBD	0.0091	Lipinski number of HB donors (OH + NH) [45]
Kappa3	0.0091	Topology index
HBAmax	0.0082	Highest free energy factor for H-bond acceptors [36]
SAS_EP_N_MEAN	0.0080	Mean of negative electrostatic potential on solvent accessible surface
HBAsum	0.0075	Sum of HB acceptor free energy
VDW_EP_N_MEAN	0.0074	Mean of negative electrostatic potential on VDW surface
Chi3c	0.0073	Sum of reciprocal square roots of valences over all 4-count branched atom paths
HDCA	0.0073	Hydrogen bond donor charged surface area (Sum of [(charge on H bond donor atom) × (sqrt area of the atom)]/(sqrt total area) [25]

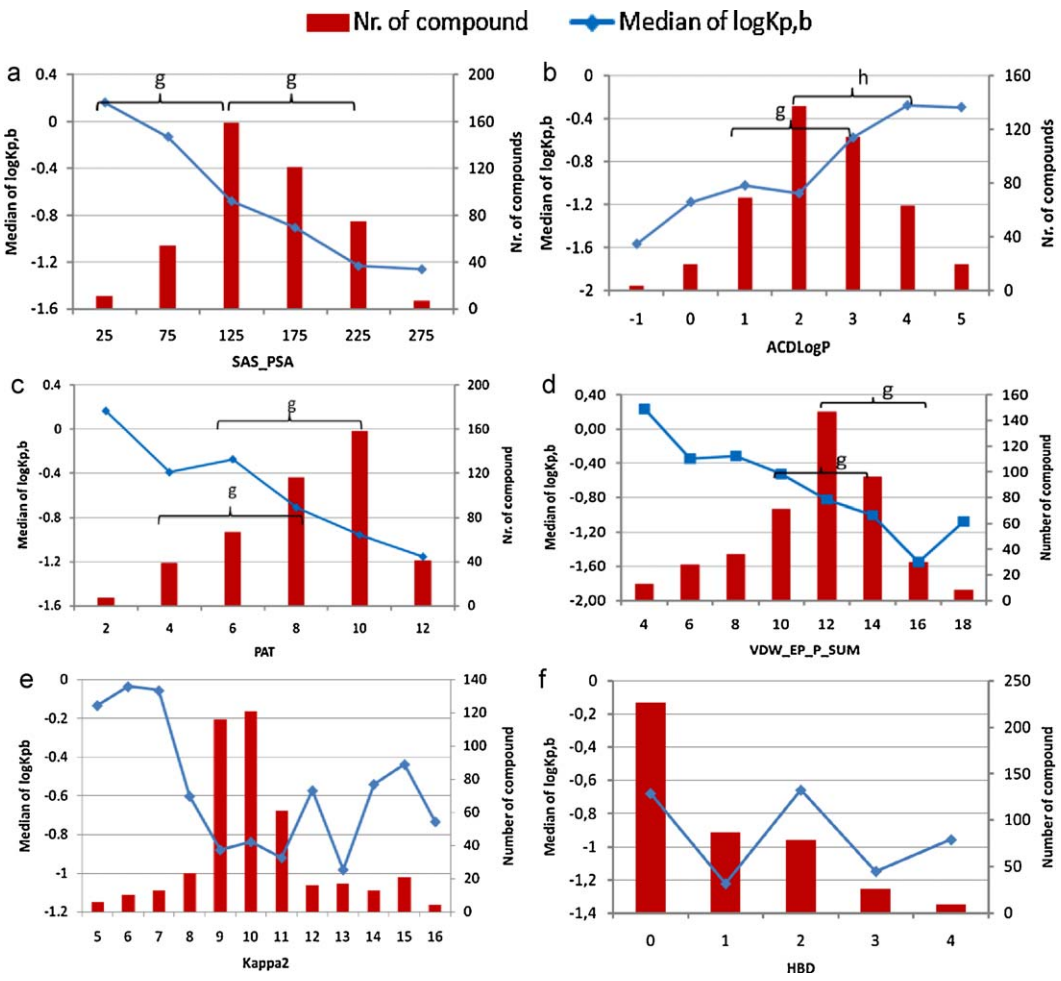

Fig. 5. The relationship between the median log $K_{p,brain}$ and (a) molecular polar surface area (SAS_PSA), (b) lipophilicity (ACD log *P*), (c) number of polar atoms (PAT), (d) VDW_EP_P_SUM, (e) Kappa2, and (f) HBD. The non-parametric Wilcoxon test shows significance at (g) $p < 0.005$ and (h) $p < 0.05$ levels.

for the RF $K_{p,uu,brain}$ direct model (RFd_$K_{p,uu,brain}$). The PLS $K_{p,uu,brain}$ direct model (PLSd_$K_{p,uu,brain}$ model) has significantly less predictive power ($R^2 = 0.41$), however, the performance is comparable to the indirect SVM model.

Consensus models are often used to obtain improved prediction performance [47]. In the current study, the average value of the predictions from the individual components is used as consensus prediction. The best consensus model is built on the SVM and the RF direct models plus the RF indirect model (SVMd_RFd_RFi_$K_{p,uu,brain}$ model). For this consensus model R^2 is 0.58, which is better than for any of the other models. The model has also the lowest RMSE value (0.46) and highest Spearman's ranking coefficient (0.77). The test set predictions for this model are plotted against the experimental values in Fig. 7. Fig. 8 shows the structures of selected example compounds, including the predicted and experimental values for $K_{p,uu,brain}$.

The relationship between the prediction accuracy and the distance to model was explored to understand the applicability domain of the model. The distance to model is defined as the average Mahalanobis distance [48,49] between a test compound and the three nearest neighbours in the training set based on the 196 descriptors used in our study. Fig. 9 shows the variation of the median prediction accuracy (RMSE of test compounds based on the SVMd_RFd_RFi_$K_{p,uu,brain}$ model) at different Mahalanobis distance

Table 5
The classification performance for various $K_{p,uu,brain}$ models.

Models[a]	Prec. Pos.	Prec. Neg.	Sensitivity	Specificity	Accuracy	*F*-score	Kappa	Matthews correlation coefficient
RFd	0.71	0.9	0.85	0.81	0.82	0.77	0.62	0.63
RFi	0.61	0.97	0.96	0.66	0.75	0.77	0.54	0.6
SVMd	0.69	0.9	0.85	0.79	0.81	0.76	0.59	0.61
SVMi	0.63	0.94	0.92	0.7	0.78	0.75	0.56	0.6
RFd_RFi	0.64	0.97	0.96	0.7	0.79	0.77	0.59	0.64
SVMd_SVMi	0.63	0.94	0.92	0.7	0.78	0.75	0.56	0.6
SVMd_RFd	0.72	0.93	0.88	0.81	0.84	0.79	0.65	0.67
SVMd_RFi	0.65	0.94	0.92	0.72	0.79	0.76	0.58	0.62
SVMd_RFd_RFi	0.71	0.97	0.96	0.79	0.85	0.82	0.68	0.72

[a] Model annotations are the same as in Table 2.

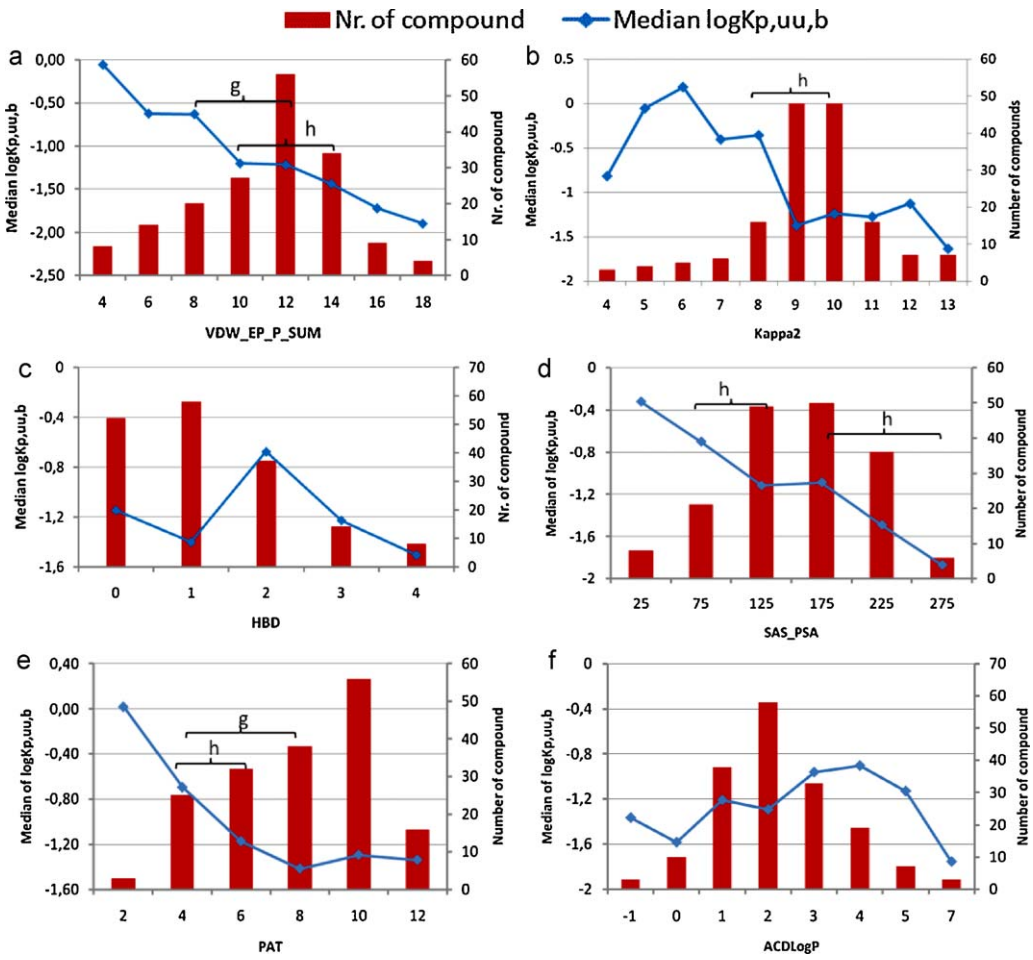

Fig. 6. The relationship between the median log $K_{p,uu,brain}$ and (a) VDW_EP_P_SUM, (b) Kappa2, (c) HBD, (d) SAS_PSA, (e) PAT and (f) ACD log *P*. The non-parametric Wilcoxon test shows significance at the (g) $p < 0.005$ and (h) $p < 0.05$ levels.

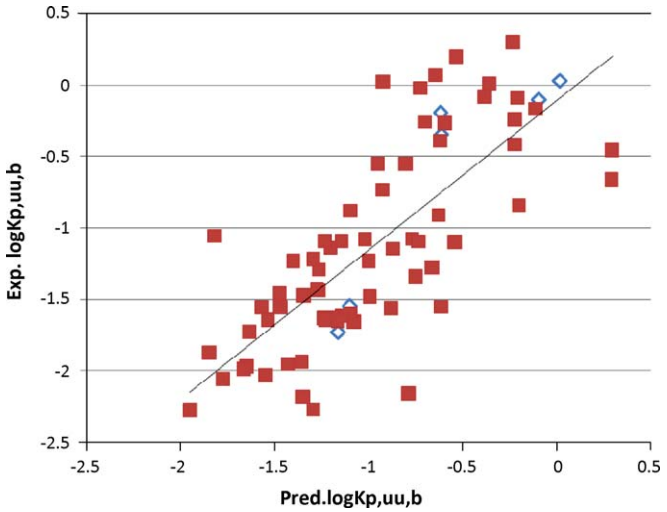

Fig. 7. $K_{puubrain}$ model SVMd_RFd_RFi: predicted vs. experimental values. The empty diamond symbols refer to the example compounds shown in Fig. 8.

bins for the $K_{p,uu,brain}$ test set. As expected, the prediction accuracy decreases for test compounds that have larger Mahalanobis distances to compounds in the training set.

Additionally, we investigated how well the regression models were able to classify the test set compounds. For prioritizing synthesis (or measurements) it is not always necessary to predict an exact value but to understand the likelihood that a compound will have exposure in the brain or not can be very helpful. In the present study, a $K_{p,uu,brain}$ value of 0.1 is used as the cut-off value for the brain exposure classification. If a compound has a $K_{p,uu,brain}$ larger or equal to 0.1, the free compound concentration in the brain is greater than or equal to 10% of the concentration of free compound in the plasma. Such a compound is regarded as having a substantial brain exposure (*Positive* class). Accordingly, if the $K_{p,uu,brain}$ is less than 0.1 the compound will be classified as having a low brain exposure (*Negative* class). This cut-off value was used for both predicted and experimental data and the classification performance of each model is evaluated based on the confusion matrix (Table 1). The results are summarised in Table 5. The consensus model SVMd_RFd_RFi shows the highest values for most of the parameters. Thus, this model, which already showed the best continuous prediction results, is also the best model to classify the test set compounds. The class prediction results for the 73 test compounds are shown in Fig. 10. The model has higher prediction accuracy for the low brain exposure compounds (negative precision = 0.97) than for the high brain exposure compounds (positive precision = 0.71). This is probably due to that the $K_{p,uu,brain}$ data set is more biased towards low brain exposure compounds as is seen from the distribution in Fig. 2d.

A set of 111 marketed CNS active drugs collected from the literature [19] was used as an additional external test set for validating the classification ability of the various models. Since these drugs are CNS active compounds, they are regarded as BBB positive com-

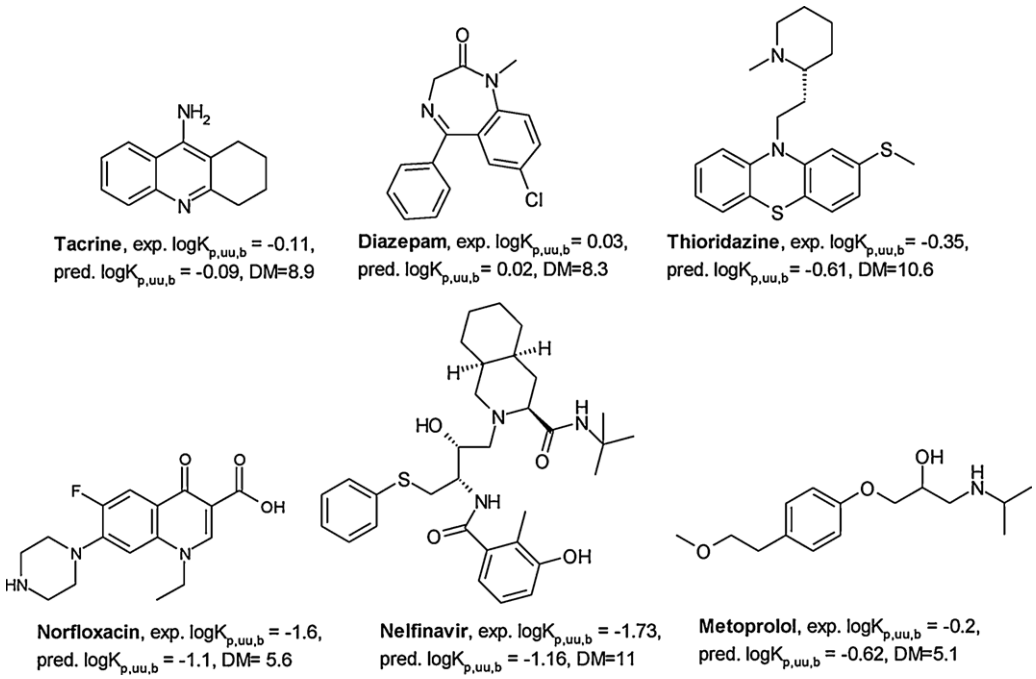

Fig. 8. Structures of selected drugs with experimental and predicted $K_{p,uu,brain}$ values, DM value refers to the distance to model.

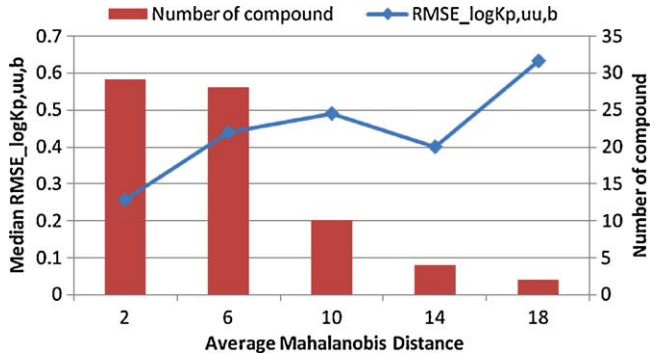

Fig. 9. The relationship between the average Mahanalobis distance between a test compound and its three nearest neighbour in the training set and the RMSE for the predicted log $K_{p,uu,brain}$ value.

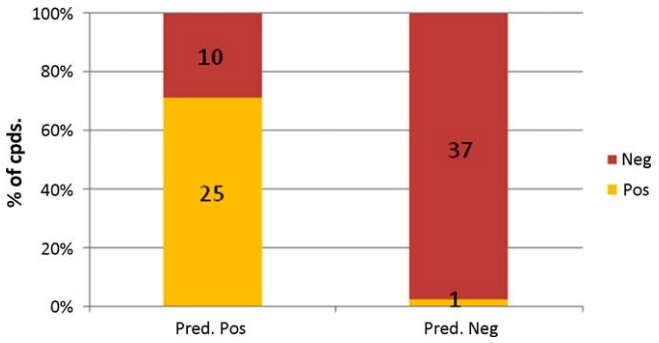

Fig. 10. The class prediction results of the test set for the consensus model SVMd_RFd_RFi. The numbers in each section of the bar refer to the number of compounds belonging to the class.

pounds. The prediction results for all the classification models are shown in Fig. 11: 83–92% of the CNS active drug set was classified as high brain exposure compounds. SVMd_RFd_RFi, previously selected as best model, correctly predicts 89% of the CNS drugs as BBB permeable. Thus, this model is not only capable of classifying the test set with reasonable accuracy, but also has a high success rate in classifying the external CNS active drug set as high brain exposure compounds.

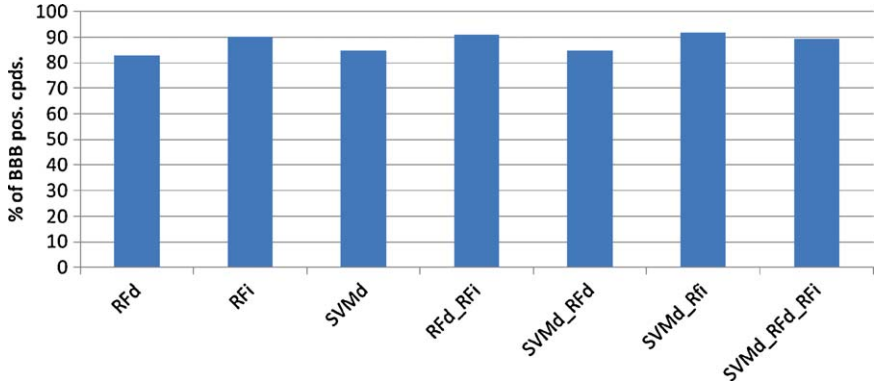

Fig. 11. The performance of the different models for the CNS active drug set.

4. Conclusions

log BB has been the standard measure to describe the level of exposure in the brain. However, in recent years it has been proposed that it is not the total but instead the unbound brain exposure which is relevant. log BB is highly influenced by plasma protein and brain tissue binding, whereas the pharmacologically relevant, unbound, brain concentration is mainly dependent on the actual transport across the BBB. Therefore, the unbound concentration ratio between blood and plasma, $K_{p,uu,brain}$, has been proposed as a measure to estimate the level of free drug exposure in the brain. In the current study, in-house $K_{p,uu,brain}$ data was used to build non-linear predictive models with two machine learning algorithms. Several models were built and validated. Consensus models that combine separate single models were also investigated. Our results show that a three component consensus model was the most accurate and gives a reasonable prediction of the unbound compound exposure in the brain. These models can also be used for classification by defining a cutoff value to distinguish compounds with high and low brain exposure. The classification performance was evaluated and the best continuous model classifies the test set into BBB positive and BBB negative compounds with an overall accuracy of 85%. The model also has an excellent success rate in predicting the brain exposure class of 111 marketed CNS drugs: 89% of the CNS drugs was correctly classified. These validation results demonstrate the applicability of a nonlinear $K_{p,uu,brain}$ model in drug discovery projects.

We also analyzed the descriptors that are important for the $K_{p,uu,brain}$ and $K_{p,brain}$ RF models and found that hydrogen bonding descriptors are important for both properties, whereas lipophilicity is an influential factor only for the $K_{p,brain}$ model; however, not for the $K_{p,uu,brain}$ model. These findings confirm the differences between the unbound and total brain–plasma concentration ratio seen previously [18], and highlight the need to use the relevant parameter ($K_{p,uu,brain}$) for estimating the likelihood of a compound to reach the desired CNS exposure.

Acknowledgements

We thank Dr. Jonna Stålring (Computational Toxicology, Global Safety Assessment, AstraZeneca R&D Mölndal) for help with AZOrange and Dr. Niklas Blomberg for a critical review and comments on the manuscript. We also would like to thank Drs. Britt-Marie Fihn, Gunilla Jerndal, and all other colleagues of DMPK department in Mölndal R&D centre that have been involved in generating the experimental data used in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2011.04.004.

References

[1] A. Reichel, The role of blood–brain barrier studies in the pharmaceutical industry, Curr. Drug Metab. 7 (2006) 183–203.
[2] U. Norinder, M. Haeberlein, Computational approaches to the prediction of the blood–brain distribution, Adv. Drug Deliv. Rev. 54 (2002) 291–313.
[3] M. Hammarlund-Udenaes, M. Fridén, S. Syvänen, A. Gupta, On the rate and extent of drug delivery to the brain, Pharm. Res. 25 (2008) 1737–1750.
[4] R.C. Young, R.C. Mitchell, T.H. Brown, C.R. Ganellin, R. Griffiths, M. Jones, K.K. Rana, D. Saunders, I.R. Smith, N.E. Sore, Development of a new physiochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists, J. Med. Chem. 31 (1988) 656–671.
[5] H. Van de Waterbeemd, M. Kansy, Hydrogen-bonding capacity and brain penetration, Chimia 46 (1992) 299–303.
[6] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena 2. Prediction of blood–brain barrier penetration, J. Pharm. Sci. 88 (1999) 815–821.
[7] F. Ooms, P. Weber, P.A. Carrupt, B. Testa, A simple model to predict blood–brain barrier permeation from 3D molecular fields, Biochim. Biophys. Acta 1587 (2002) 118–125.
[8] Y. Fan, R. Unwalla, R.A. Denny, D. Li, E.H. Kerns, D.J. Diller, C. Humblet, Insights for predicting blood–brain barrier penetration of CNS targeted molecules using QSPR approaches, J. Chem. Inf. Model. 50 (2010) 1123–1133.
[9] O. Engkvist, P. Wrede, U. Rester, Prediction of CNS activity of compound libraries using substructure analysis, J. Chem. Inf. Comput. Sci. 43 (2003) 155–160.
[10] I. Martin, Prediction of blood–brain barrier penetration: are we missing the point? Drug Discov. Today 9 (2004) 161–162.
[11] H. Van de Waterbeemd, D.A. Smith, B.C. Jones, Lipophilicity in PK design: methyl, ethyl, futile, J. Comput. Aided Mol. Des. 15 (2001) 273–286.
[12] H. Wan, M. Rehngren, F. Giordanetto, F. Bergstrom, A. Tunek, High-throughput screening of drug-brain tissue binding and in silico prediction for assessment of central nervous system drug delivery, J. Med. Chem. 50 (2007) 4606–4615.
[13] W.M. Pardridge, log(BB), PS products and in silico models for drug brain penetration, Drug Discov. Today 9 (2004) 392–393.
[14] M.H. Abraham, The factors that influence permeation across the blood–brain-barrier, Eur. J. Med. Chem. 39 (2004) 235–240.
[15] X. Liu, M. Tu, R.S. Kelly, C. Chen, B.J. Smith, Development of a computational approach to predict blood brain barrier permeability, Drug Metab. Dispos. 32 (2004) 132–139.
[16] M. Fridén, A. Gupta, M. Antonsson, U. Bredberg, M. Hammarlund-Udenaes, In vitro methods for estimating unbound drug concentration in the brain interstitial and intracellular fluids, Drug Metab. Dispos. 35 (2007) 1711–1719.
[17] M. Hammarlund-Udenaes, U. Bredberg, M. Fridén, Methodologies to assess brain drug delivery in lead optimization, Curr. Topic Med. Chem. 9 (2009) 148–162.
[18] M. Fridén, S. Winiwarter, G. Jerndal, O. Bengtsson, H. Wan, U. Bredberg, M. Hammarlund-Udenaes, M. Antonsson, Structure–brain exposure relationship in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids, J. Med. Chem. 52 (2009) 6233–6243.
[19] T.T. Wager, R.Y. Chandrasekaran, X. Hou, M.D. Troutman, P.R. Verhoest, A. Villalobos, Y. Will, Defining desirable central nervous system drug space through the alignment of molecular properties, in vitro ADME, and safety attributes, ACS Chem. Neurosci. 1 (2010) 420–434.
[20] M. Friden, F. Ducrozet, B. Middleton, M. Antonsson, U. Bredberg, M. Hammarlund-Udenaes, Development of a high-throughput brain slice method for studying drug distribution in the central nervous system, Drug Metab. Dispos. 37 (2009) 1226–1233.
[21] H. Wan, F. Bergstrom, High throughput screening of drug–protein binding in drug discovery, J. Liq. Chromatogr. Related Technol. 30 (2007) 681–700.
[22] M. Friden, H. Ljungqvist, B. Middleton, U. Bredberg, M. Hammarlund-Udenaes, Improved measurement of drug exposure in brain using drug-specific correction for residual blood, J. Cereb. Blood Flow Metab. 30 (2010) 150–161.
[23] S.W. Paine, P. Barton, J. Bird, R. Denton, K. Menochet, A. Smith, N.P. Tomkinson, K.K. Chohan, A rapid computational filter for predicting the rate of human renal clearance, J. Mol. Graphics Modell. 29 (2010) 529–537.
[24] P. Bruneau, Search for predictive generic model of aqueous solubility using Bayesian neural nets, J. Chem. Inf. Comput. Sci. 41 (2001) 1605–1616.
[25] A. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Kalrelson, QSPR studies on vapor pressure, aqueous solubility, and the prediction of water–air partition coefficient, J. Chem. Inf. Comp. Sci. 38 (1998) 720–725.
[26] AZOrange 0.3, http://github.com/AZCompTox/AZOrange (accessed 07.10.10).
[27] Orange official web site, http://www.ailab.si/orange/ (accessed 01.09.10).
[28] SIMCA-P+, version 12.0, Umetrics, Umeå, Sweden.
[29] JMP Version 7, SAS Institute Inc., Cary, NC.
[30] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[31] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons, Inc., New York, 1998.
[32] L. Breimann, Random forests, Mach. Learn. 45 (2001) 5–32.
[33] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.
[34] M.E. Stokes, C.S. Davis, G. Koch, Observer agreement, in: M.E. Stokes, C.S. Davis, G. Koch (Eds.), Categorical Data Analysis Using the SAS System, 2nd ed., SAS Institute, Cary, NC, 1995, pp. 98–102.
[35] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.
[36] O.A. Raevsky, K.J. Schaper, Analysis of water solubility data on the basis of HYBOT descriptors. Part1. Partitioning of volatile chemicals in the water–gas phase system, QSAR Comb. Sci. 22 (2004) 926–942.
[37] J. Kelder, P.D.J. Grootenhuis, D.M. Bayada, L.P.C. Delbressine, J.P. Ploemen, Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs, Pharm. Res. 16 (1999) 1514–1519.
[38] S. Vilar, M. Chakrabarti, S. Costanzi, Prediction of passive blood–brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors, J. Mol. Graphics Modell. 28 (2010) 899–903.
[39] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, QSAR modeling of the blood–brain barrier permeability for diverse organic compounds, Pharm. Res. 25 (2008) 1902–1914.

[40] T.J. Hou, X.J. Xu, ADME evaluation in drug discovery. 3. Modelling blood–brain barrier partitioning using simple molecular descriptors, J. Chem. Inf. Comput. Sci. 43 (2003) 2137–2152.
[41] A.R. Katritzky, M. Kuanar, S. Slavov, D.A. Dobchev, D.C. Fara, M. Karelson, W.E. Acree Jr., V.P. Solov'ev, A. Varnek, Correlation of blood–brain penetration using structural descriptors, Bioorg. Med. Chem. 14 (2006) 4888–4917.
[42] M. Iyer, R. Mishra, Y. Han, A.J. Hopfinger, Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis, Pharm. Res. 19 (2002) 1611–1621.
[43] L.B. Kier, A shape index from molecular graphs, Quant. Struct. Act. Relat. 4 (1985) 109–116.
[44] L.B. Kier, Shape indexes of orders one and three from molecular graphs, Quant. Struct. Act. Relat. 5 (1986) 1–7.
[45] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.
[46] K. Lanevskij, J. Dapkunas, L. Juska, P. Japertas, R. Didziapetris, QSAR analysis of blood–brain distribution: the influence of plasma and brain tissue binding, J. Pharm. Sci. 100 (2011) 2147–2160.
[47] P. Willett, Enhancing the effectiveness of ligand-based virtual screening using data fusion, QSAR Comb. Sci. 25 (2006) 1143–1152.
[48] P.C. Mahalanobis, On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India 2 (1936) 49–55.
[49] R. Gnanadesikan, J.R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics 28 (1972) 81–124.