BioImpacts, **2013**, 3(1), 21-27
doi: 10.5681/bi.2013.011
http://bi.tbzmed.ac.ir/

QSAR Models for the Prediction of Plasma Protein Binding

Taravat Ghafourian[*]**, Zeshan Amin**

Medway School of Pharmacy, Universities of Kent and Greenwich, Central Avenue, Chatham Maritime, Kent ME4 4TB, UK

ARTICLE INFO

Article Type:
Research Article

Article History:
Received: 25 Jan. 2013
Revised: 13 Feb. 2013
Accepted: 20 Feb. 2013
ePublished: 21 Feb. 2013

Keywords:
QSAR
ADME
Distribution
Protein Binding
Albumin Binding
Serum Proteins

ABSTRACT

Introduction: The prediction of plasma protein binding (ppb) is of paramount importance in the pharmacokinetics characterization of drugs, as it causes significant changes in volume of distribution, clearance and drug half life. This study utilized Quantitative Structure – Activity Relationships (QSAR) for the prediction of plasma protein binding. ***Methods:*** Protein binding values for 794 compounds were collated from literature. The data was partitioned into a training set of 662 compounds and an external validation set of 132 compounds. Physicochemical and molecular descriptors were calculated for each compound using ACD labs/logD, MOE (Chemical Computing Group) and Symyx QSAR software packages. Several data mining tools were employed for the construction of models. These included stepwise regression analysis, Classification and Regression Trees (CART), Boosted trees and Random Forest. ***Results:*** Several predictive models were identified; however, one model in particular produced significantly superior prediction accuracy for the external validation set as measured using mean absolute error and correlation coefficient. The selected model was a boosted regression tree model which had the mean absolute error for training set of 13.25 and for validation set of 14.96. ***Conclusion:*** Plasma protein binding can be modeled using simple regression trees or multiple linear regressions with reasonable model accuracies. These interpretable models were able to identify the governing molecular factors for a high ppb that included hydrophobicity, van der Waals surface area parameters, and aromaticity. On the other hand, the more complicated ensemble method of boosted regression trees produced the most accurate ppb estimations for the external validation set.

Introduction

Many drugs bind with varying degrees of association to human plasma protein. Plasma protein binding is the reversible association of a drug with the proteins of the plasma due to hydrophobic and electrostatic interactions such as van der Waals and hydrogen bonding. The bound drug exists in equilibrium with the free drug.[1] This reversible interaction can greatly influence the pharmacokinetic properties such as volume of distribution, clearance and elimination, as well as the pharmacological effect of the drug. Only a fraction of unbound (f_u) drug is able to pass across cell membranes.[2] Thus, it can be expected that drugs with high protein binding tend to have a greater half–life compared to those with lower values. The greater the drug is bound to plasma protein, the less fraction of free drug is there for therapeutic effect. The consequences of protein binding are most extensive with drugs that are highly protein bound and have a narrow therapeutic index. This is therefore a vital attribute for the assessment of human risk. Significance of protein binding in pharmacokinetics and pharmacodynamics has been reviewed recently.[3] The plasma protein binding (ppb) is therefore of paramount importance in the pharmacokinetics characterization of drugs. Prediction of the free fraction in tissues and plasma is of interest in drug discovery and development.

Plasma accounts for 55% of the human blood's composition. It is an aqueous solution mainly composed of water (92%), proteins (7%) and others solutes (1%) such as inorganic ions.[1] Plasma proteins include albumin, globulins, clotting factors and regulatory proteins. The most important proteins in terms of drug binding are albumin and α1-acid glycoprotein, followed by lipoproteins.[4] The serum albumin is the primary constituent in human plasma proteins with the concentration of 600 µM accounting for 60% of total plasma protein. There are multiple hydrophobic binding

***Corresponding author:** Taravat Ghafourian, Email: t.ghafourian {at} kent.ac.uk

sites on albumin (a total of eight for fatty acid) that especially bind not to neutral and negatively charged hydrophobic compounds such as NSAIDS, but also to some basic drugs such as tricyclic antidepressants.[5] Binding of acidic drugs to albumin is often considered restrictive as far as the distribution of the drug is concerned.[6] Basic lipophilic drugs such as antidepressants bind to albumin, AGP and lipoproteins. Very lipophilic, water-insoluble compounds bind to lipoproteins.

For drugs and drug-like compounds, there are two main binding sites on albumin. Both sites are elongated hydrophobic pockets possessing charged lysine's and arginine's residues near the surface, which serve as attachment points for polar ligand features.[5] Sudlow siteI especially binds bulky heterocyclic anions (e.g. warfarin), whilst siteII preferentially recognizes small aromatic carboxylic acids (i.e. ibuprofen).[7] Albumin also has a number of minor binding sites, which allow different drug molecules to bind simultaneously, leading to higher binding capacity.[8] Alpha-1-acid glycoprotein is abundant in serum with a concentration of about 0.01-0.02 mmolar. It binds to a number of endogenous compounds such as steroids, retinoic acid and heparin, as well as a variety of drugs (not mainly basic or neutral, but also some acidic ones e.g. Phenobarbital).[1] The effect of AAG binding on drug disposition is more significant in diseases associated with elevated AAG levels, such as cancer. As a result, interaction of AAG with antineoplastic agents needs to be studied taking into consideration the different levels of the protein in the serum of patients suffering from cancer. Finally, apart from the elevated AAG levels, depressed HSA levels (negative acute-phase protein) should be taken into account, as freer drug might be present as a result for binding with AAG. Lipoproteins are macromolecular complexes containing protein components (apoproteins) and polar lipids (phospholipids) in a surface film surrounding a neutral core. Their concentrations may vary 4-5 folds.[9] There are four types of lipoproteins which differ in density and size. These are chylomicrons, high-density lipoproteins (HDL), low-density lipoproteins (LDL) and very low-density lipoproteins (VLDL). Their physiological role is the transport of cholesterol and triglycerides.[1] Although lipoproteins have been reported to contribute to the binding of extremely hydrophobic drugs, binding occurs when a drug is present at very high concentrations.

There are varieties of *in vitro* assays that can be utilized to determine the extent of plasma protein binding. Such techniques include equilibrium dialysis which is considered to be the 'gold standard', ultrafiltration, ultracentrifugation, chromatographic methods, fluorescence spectroscopy, ultraviolet spectroscopy, circular dichroism, nuclear magnetic resonance spectroscopy, and capillary electrophoresis.[1,10]

However, there is a need for reliable *in silico* techniques which will be able to cope with the enormous amounts of data available for screening and will also be able to predict the plasma protein binding of virtual compounds in order to avoid the synthesis of chemicals which do not have the potentiality of being approved drugs. One such technique is the use of Quantitative Structure-Activity Relationship (QSAR) techniques which aim at estimation of plasma protein binding levels based on the molecular and physicochemical properties of compounds. There have been a number of attempts to understand the molecular factors that influence binding to human plasma proteins. Previous studies of protein binding in homologous series have suggested that plasma protein binding is only related to ligand lipophilicity.[11] Recent studies have considered various chemical structures for the development of QSAR models. Even for the diverse datasets, lipophilicity seems to be the most important determinant of binding affinity to plasma proteins.[12] Among the methods used for the estimation of plasma protein binding are: QSAR based on pharmacophoric similarity concept and partial least square analysis,[5] multiple linear regression, artificial neural networks, k-nearest neighbors and support vector machines,[13] and partial least squares regression on data limited to binding measurements using equilibrium dialysis.[14]

The aim of this study was to use several QSAR modeling methods in order to improve the accuracy of ppb estimation. To this end, a large and diverse dataset of drug ppb values were employed and several linear and nonlinear data mining methods were used. The validity of the models was explored using internal and external validation studies.

Materials and methods

Dataset

Human plasma protein binding values compiled by Votano *et al.*[13] were used in this study. This value consisted of a set of 794 compounds literature values of percentage compounds bound to plasma proteins (ppb) from a variety of literature sources. In the dataset, 41% had two or more reported values which were averaged. Compounds were not included in the dataset where the reported values differed by 30% or more. Dataset was partitioned into a training set of 662 and validation set of

132 compounds in a way that both sets contained similar ranges of ppb values. To do this, compounds were sorted in ascending order of ppb values and from each group of six, the first 5 were allocated into training and the sixth into the test set.

Molecular descriptors

Molecular descriptors were calculated for the 794 compounds using ACD labs/logD Suite version 12, TSAR 3D version 3.3 (Accelrys Inc), Molecular Operating Environment (MOE) version 2008.10 (Chemical Computing Group) and Symyx QSAR software. Calculation of 3D descriptors was performed on the molecular geometries optimized using AM1 semiempirical method. Descriptors would be removed if more than 98% of the values were identical. One of the two highly inter-correlated descriptors was also excluded. The final descriptor list consisted of around 450 descriptors.

Development and validation of QSARs

Models were developed using Minitab 15.1 statistical software and Statistica Data Mining version 9 (StatSoft, Inc) for the training set comprising 662 compounds. Stepwise regression analysis was used for the development of a linear regression model. Data mining techniques were general regression tree, interactive tree, random forest and boosted trees in Statitistica.

For the development of general regression tree model, 16 trees were generated using several combinations of stopping criteria and the best tree was selected according to the randomly selected internal validation set. In the selected model, the stopping parameters were 10 for the maximum number of levels, 19 for the minimum number of cases in child nodes and 1000 for the maximum number of nodes. The minimum number of cases for partitioning was set at 79. In addition, an interactive tree was generated where the selected general regression tree was pruned in order to reduce the risk of overfitting.

Two random forest tree models were developed in which the number of trees used in each model were 100 trees and 40 trees. The stopping criteria included the minimum number of cases at 19, the maximum number of levels at 10, the minimum number in child node at 5 and the maximum number of nodes at 100. The number of predictors at each tree was set to 9 and the subsample proportion was set at 0.50.

Boosted tree methods in Statistica allow multiple tree models to be generated for the prediction. It computes a sequence of simple trees, where each successive tree is built for the prediction residuals of the proceeding trees. Two boosted tree models were developed comprising 200 trees and 80 trees. The stopping criteria included the minimum number of cases at 19, the maximum number of levels at 10, the minimum number in child node at 1 and the maximum number of nodes at 3. In addition, the learning rate was 0.1 and the subsample proportion was 0.50.

The selected models were used for the calculation of ppb values of the external test set comprising 132 compounds. Mean Absolute Error of prediction and correlation between observed and predicted values were used for assessing the predictive ability of the models.

Results

Linear regression and non-linear data mining methods were used for the development of QSAR for the estimation of ppb. Only training set was used for the development of QSAR. The prediction powers of the models were compared using the test set compounds. The selected models using each statistical technique are presented here.

Stepwise regression model

The first eight most statistically significant ($p < 0.05$) molecular descriptors selected by stepwise regression analysis was used for the development of the regression model below (Equation 1).

$$\mathrm{ppb} = 6.20\ \mathrm{LogP} + 0.0823\ \mathrm{Q\_VSA\_NEG} - 14.6\ \mathrm{FiB7.4} + 79.6\ \mathrm{GCUT\_SLOGP\_3} - 86.5\ \mathrm{GCUT\_PEOE\_3} - 8.34\ \mathrm{FU7.4} - 0.121\ \mathrm{Q\_VSA\_PPOS} - 103\ \mathrm{VAdjEq} + 125$$

$$N = 662\ S = 22.5\ R^2 = 0.558\ F = 103$$

In this model the descriptors are LogP, the logarithm of octanol/water partition coefficient calculated by ACD/ log D Suite; Q_VSA_NEG, total negative van der Waals surface area; FiB7.4, fraction of base ionized at pH 7.4 calculated using Henderson–Hasselbalch equation with most basic pKa from ACD/ log D Suite; GCUT_SLOGP_3, GCUT descriptor using atomic contribution to log P instead of atomic charge; GCUT_PEOE_3, GCUT descriptor using PEOE method for the calculation of atomic charge; FU7.4, fraction of drugs that are unionized at pH 7.4 using Henderson–Hasselbalch equation with most acidic and most basic pKa from ACD/ log D Suite; Q_VSA_PPOS, total positive polar van der Waals surface area; VAdjEq, and vertex adjacency information (equality). Apart from LogP, FiB7.4 and FU7.4, the remaining descriptors of this equation are calculated by MOE software. The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix where the diagonal can take any atomic property. The smallest, 1/3-ile, 2/3-ile and largest eigenvalues are reported.[15]

The equation indicates the positive effect of lipophilicity measured by LogP and GCUT_SLOGP_3 on plasma protein binding. The negative coefficient of FiB7.4 indicates a lower tendency of basic drugs for binding to plasma proteins and in combination with FU7.4 with negative coefficient, it can be indicative of higher plasma binding of acidic drugs in comparison with basic or neutral compounds. GCUT_PEOE_3 and Q_VSA_PPOS with negative coefficients indicate the negative effect of hydrogen bonding donor groups as these atoms are the main contributors to the positive atomic charge.[16] On the other hand, Q_VSA_NEG has a positive coefficient and indicates the positive contribution of negatively charged surface to ppb.

General regression tree models

Out of 16 trees, one was selected according to the prediction error for the randomly selected internal validation set. Moreover, the pruned version of this selected tree was examined as the interactive tree presented in Fig. 1. The figure shows that similar to the regression model, lipophilicity is the major determinant of ppb with high log P and high SlogP drugs, and those with high hydrophobic interaction field (vsurf_EDmin2) have significantly higher average ppb. Similar to the regression model, compounds with higher negatively charged surface (see the effect of PEOE_VSA_NEG) have higher binding tendency. Moreover, compounds with more than 9.5 aromatic rings (as calculated by MOE) have a significantly higher percentage bound.

Boosted tree models

This is based on Stochastic Gradient Boosting method. The method computed a sequence of very simple trees, where each successive tree is built for the prediction of residuals of the preceding tree. Then the average prediction by many trees is used. The boosted tree analysis was first performed using 200 trees. In this analysis, ensemble of 163 trees was selected as the optimum number of trees by the software (Fig. 2). This figure shows the change in training and test set error with increasing number of trees. It can be seen that after 80 trees, the increase in the number of trees reduces the training set error, but the test set error does not change significantly. Therefore, a second analysis was performed using only 80 trees.

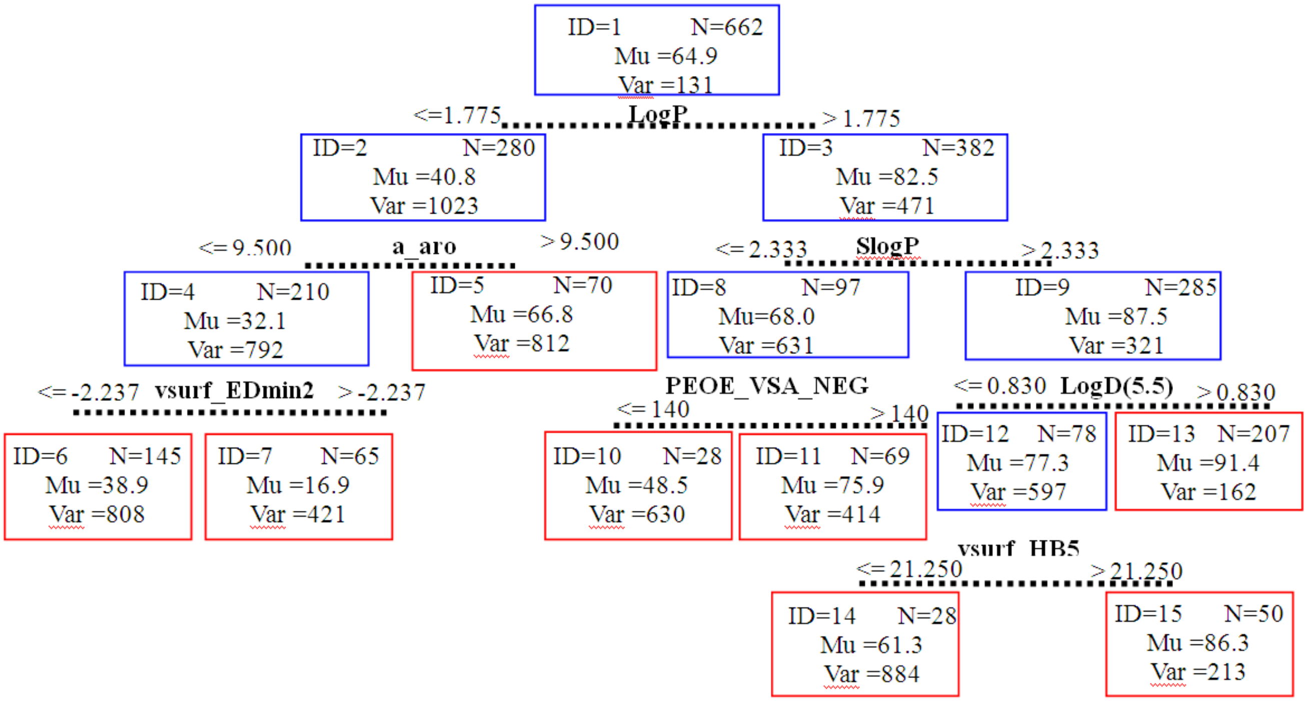

Fig. 1. The interactive tree diagram for protein binding (ppb); each node is identified with an ID number; Mu is the average ppb; Var is the variance; N is the number of compounds in each node. The descriptors used for the classification are Log P, octanol water partition coefficient calculated by ACD/logD; SlogP, the log P calculated using the Wildman and Crippen SlogP method;[17] a_aro, number of aromatic atoms, LogD (5.5), apparent partition coefficient measured by ACD/logD; PEOE_VSA_NEG, Total negative van der Waals surface area where atomic charge is measured by PEOE method; vsurf_EDmin2, Volsurf descriptor[18] indicate ng the second lowest hydrophobic energy; and vsurf_HB5 is a Volsurf descriptor[18] indicating hydrogen bonding interaction. Apart from ACD/logD calculated descriptors identified above, the remaining descriptors are calculated by MOE software.

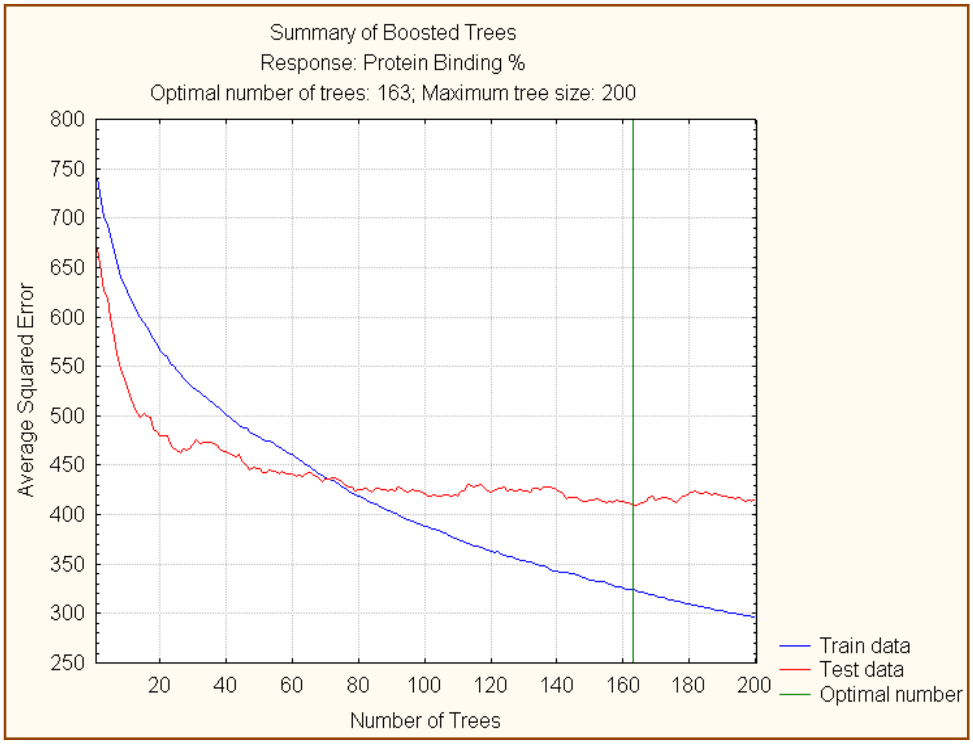

Fig. 2. The effect of number of trees on the average error in boosted tree analysis.

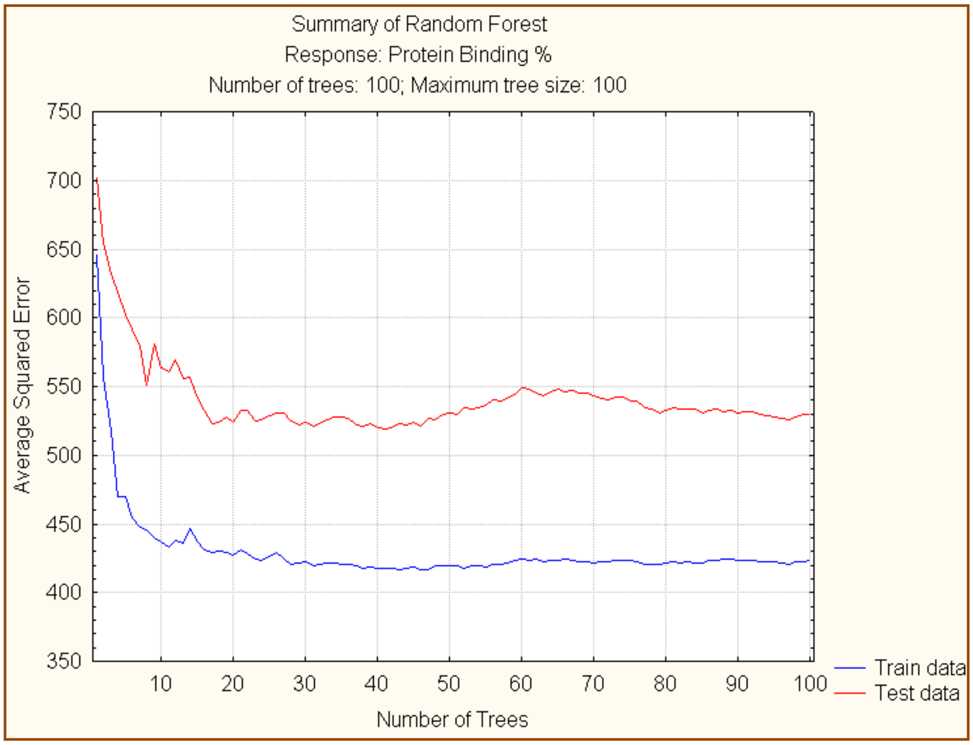

Fig. 3. The effect of number of trees on the average error in random forest analysis.

Random forest models

Random forest consisted of a collection of simple trees generated using a subset of 9 descriptors. Fig. 3 indicates the effect of the number of trees on the training and test errors, indicating insignificant changes of test set error with tree numbers >40. Therefore, a second random forest analysis employed only 40 trees.

Discussion

Interpretable models

Plasma protein binding is an extremely important pharmacokinetic parameter which needs to be investigated in early stages of drug development. Although a number of attempts have been made to investigate the factors that affect plasma protein binding, there is still not a clear picture. The purpose of this study was to produce *in silico* QSAR models which could assist in finding the molecular factors affecting ppb. This was achieved in this investigation by two simple and interpretable models, namely multiple regression equation and regression tree.

According to both Regression Tree (Fig. 1) and Equation 1, lipophilicity has a positive effect on plasma protein binding meaning in that the more lipophilic drugs have higher ppb values. In these two models, lipophilicity has been represented by a variety of molecular descriptors including logP, logD at pH 5.5, GCUT_SLOGP_3 and vsurf_EDmin2. Effect of lipophilicity has also been indicated by a number of other studies including those of Votano *et al*.,[13] and Colmenarejo *et al*.[12]

In the regression equation, the negative effect of FiB in combination with the negative effect of FU in Equation 1, shows that acidic drugs have higher affinity for plasma proteins and basic drugs are less likely to bind to albumin. This has also been suggested by a volume of distribution models that indicate higher apparent distribution volumes for basic drugs while acidic drugs are retained more in the plasma.[19] These molecular descriptors are not seen in the regression tree, but instead, hydrogen bonding interaction (vsurf_HB5) shows a positive effect on ppb of highly lipophilic drugs in terminal node 15.

A third finding from both regression equation and the nonlinear regression trees is the higher binding tendency of compounds with higher negatively charged surface (see the effect of PEOE_VSA_NEG in Equation 1 and Fig. 1). This also agrees with the traditional idea that albumin binds mainly to acidic compounds.[1]

Prediction accuracy of models

Fitting a model into a dataset does not fully achieve the ultimate goal of a modeling exercise. Any computational model needs to be proven successful by correctly predicting the external compounds (validation set) which have not been used in any stage of the model development.[20] In this investigation, one out of six compounds in the dataset (132 compounds) was kept for external validation of models. Table 1 shows the accuracy of the selected QSAR models in terms of the Mean Absolute error (MAE) for the training and validation sets and the correlation coefficient between the observed and the predicted ppb. The table shows that, for the training set, the order of accuracy according to MAE is boosted tree one, general regression tree and boosted tree two. The accuracy order changes for the validation set with boosted tree model two showing the second most accurate estimation for the external validation set after boosted tree model one, while the general regression tree is in the third place. Moreover, the boosted tree models, along with all the other models presented in Table 1, do not show overfitting to the training data. This is indicated by the small gap between training and validation set MAE values.

Table 1. The summary of prediction accuracy of QSAR models; N, number of cases; S, standard deviation, slope, intercept and R^2 of the correlation between observed and predicted ppb; MAE, the mean absolute error

Model	Name	N	S	Slope	Intercept	R^2	MAE
			Training/internal test set				
1	Stepwise Regression	662	22.38	0.530	30.05	0.589	17.07
2	General Tree	662	19.83	0.653	22.50	0.653	14.21
3	Interactive Tree	662	21.90	0.577	27.44	0.577	16.30
4	Random Forest Tree One	662	19.01	0.486	33.50	0.681	16.90
5	Random Forest Tree Two	662	18.97	0.494	33.29	0.683	16.64
6	Boosted Tree One	662	17.93	0.681	20.62	0.717	13.25
7	Boosted Tree Two	662	20.46	0.615	25.02	0.631	15.16
			Validation set				
1	Stepwise Regression	132	21.30	0.504	34.55	0.626	17.01
2	General Tree	132	23.02	0.579	29.76	0.538	16.86
3	Interactive Tree	132	24.16	0.512	34.79	0.491	18.15
4	Random Forest Tree One	132	21.92	0.413	39.32	0.581	18.91
5	Random Forest Two	132	21.75	0.424	38.94	0.588	18.56
6	Boosted Tree One	132	20.16	0.641	25.13	0.646	14.96
7	Boosted Tree Two	132	20.65	0.622	26.47	0.629	15.34

Boosted tree model one has employed a large number of small trees (163 trees in total) in order to correctly predict ppb values for the compounds. Ensemble methods such as boosted trees and random forest have the advantage of flexibly employing many predictors in a non-linear fashion which can aid prediction accuracy. On the other hand, this improvement in estimation accuracy is accompanied by a loss in the interpretability. Although it must be noted that the most significant descriptors that have an influence on ppb can be identified from the descriptor ranking results of the boosted tree analysis.

Conclusion

The range of simple to more complicated QSAR models developed in this investigation resulted in encouragingly low prediction errors. In particular, the boosted tree models proved to have considerably lower estimation error for the external validation set.

Competing interests

The authors declared no competing interests.

References

1. Colmenarejo G. In silico prediction of plasma and tissue protein binding. In: Taylor JB, Triggle DJ. Comprehensive Medicinal Chemistry II, Vol. 5. Oxford: Elsevier;**2007**. pp. 847–66.
2. Jambhekar, SS. Physicochemical and biopharmaceutical properties of drug substances and pharmacokinetics. In: Foye WO, Lemke TL, Williams DA, editors. Foye's Principles of Medicinal Chemistry. New York: Lippincott Williams and Wilkins;**2008**. pp. 247–50.
3. Schmidt S, Gonzalez D, Derendorf H. Significance of protein binding in pharmacokinetics and pharmacodynamics. *J Pharm Sci* **2010**;99:1107–22.
4. Zsila F, Fitos I, Bencze G, Keri G, Orfi L. Determination of human serum α1 acid glycoprotein and albumin binding of various marketed and preclinical kinase inhibitors. *Curr Med Chem* **2009**;16:1964–77.
5. Kratochwil N, Huber W, Müller F, Kansy M, Gerber P. Predicting plasma protein binding of drugs: a new approach. *Biochem Pharmacol* **2002**;64:1355–74.
6. Urien S, Tillement JP, Barré J. The significance of plasma-protein binding in drug research. In: Testa B, van de Waterbeemd H, Folkers G, Guy R, editors. Pharmacokinetic Optimization in Drug Research. Zürich: Wiley-VCH;**2001**. pp. 189–97.
7. Bolli A, Marino M, Rimbach G, Fanoli G, Fasano M, Ascenzi P. Flavonoid binding to human serum albumin. *Biochem Biophys Res Commun* **2010**;398:444–9.
8. Zhu L, Yang F, Chen L, Meehan EJ, Huang M. A new drug binding site on human serum albumin and drug-drug interaction studied by X-ray crystallography. *J Struct Biol* **2008**;102:40–9.
9. Barton P, Austin RP, Fessey RE. In Vitro Models for Plasma Binding and Tissue Storage. In: Taylor JB, Triggle DJ. Comprehensive Medicinal Chemistry II, Vol. 5. Oxford: Elsevier;**2007**. pp. 321–40.
10. Kawai Y, Fujil Y, Akimoto K, Takahashi M. Evaluation of serum protein binding by using *in-vitro* pharmacological activity for the effective pharmacokinetics profiling in drug discovery. *Chem Pharm Bull* **2010**;58:1051–6.
11. Mayer JM, van de Waterbeemd H. Development of quantitative structure-pharmacokinetic relationships. *Environ Health Perspect* **1985**;61:295–306.
12. Colmenarejo G, Alvarez-Pedraglio A, Lavandera JL. Chemoinformatic models to predict binding affinities to human serum albumin. *J Med Chem* **2001**;44:4370–8.
13. Votano RJ, Parham M, Hall LM, Hall HL, Kier BJ, Oloff S, *et al*. QSAR modelling of human serum protein binding with several modelling techniques utilizing structure-information representation. *J Med Chem* **2006**;49:7169–81.
14. Gleeson MP. Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. *J Med Chem* **2007**;50:101–12.
15. Lin A. QuaSAR: The MOE System for QSAR. Chemical Computing Group Inc; **2013**. Available from: URL: http://www.chemcomp.com/journal/qsar.htm.
16. Dearden JC, Ghafourian T. Hydrogen bonding parameters for QSAR: comparison of indicator variables, hydrogen bond counts, molecular orbital and other parameters. *J Chem Inf Comput Sci* **1999**;39:231–5.
17. Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci* **1999**;39:868–73.
18. Cruciani G, Pastor M, Guba W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur J Pharm Sci* **2000**;11:S29.
19. Ghafourian T, Barzegar-Jalali M, Dastmalchi S, Khavari-Khorasani T, Hakimiha N, Nokhodchi A. QSAR models for the prediction of apparent volume of distribution. *Int J Pharm* **2006**;319:82–97.
20. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* **2007**;26:694–701.