Pharm Res (2013) 30:1790–1798
DOI 10.1007/s11095-013-1023-6

RESEARCH PAPER

The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding

Xiang-Wei Zhu • Alexander Sedykh • Hao Zhu • Shu-Shen Liu • Alexander Tropsha

Received: 3 January 2013 / Accepted: 4 March 2013 / Published online: 9 April 2013

ABSTRACT

Purpose To develop accurate *in silico* predictors of Plasma Protein Binding (PPB).

Methods Experimental PPB data were compiled for over 1,200 compounds. Two endpoints have been considered: *(1)* fraction bound (%PPB); and *(2)* the logarithm of a pseudo binding constant (lnKa) derived from %PPB. The latter metric was employed because it reflects the PPB thermodynamics and the distribution of the transformed data is closer to normal. Quantitative Structure-Activity Relationship (QSAR) models were built with Dragon descriptors and three statistical methods.

Results Five-fold external validation procedure resulted in models with the prediction accuracy (R^2) of 0.67 ± 0.04 and 0.66 ± 0.04, respectively, and the mean absolute error (MAE) of 15.3 ± 0.2% and 13.6 ± 0.2%, respectively. Models were validated with two external datasets: 173 compounds from DrugBank, and 236 chemicals from the US EPA ToxCast project. Models built with lnKa were significantly more accurate (MAE of 6.2–10.7 %) than those built with %PPB (MAE of 11.9–17.6 %) for highly bound compounds both for the training and the external sets.

Conclusions The pseudo binding constant (lnKa) is more appropriate for characterizing PPB binding than conventional %PPB. Validated QSAR models developed herein can be applied as reliable tools in early drug development and in chemical risk assessment.

KEY WORDS %PPB · ADMET · drug fraction bound · machine learning · pharmacokinetics

ABBREVIATIONS

%PPB	Percent plasma protein binding
5FCV	5-fold external cross-validation
AD	Applicability domain
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
CCR	Correct classification rate (balanced classification accuracy)
HSA	Human serum albumin
*k*NN	*k* nearest neighbors
lnKa	Natural logarithm of the pseudo binding constant imputed from %PPB
LogP	Octanol-water partition coefficient
LOO-CV	Leave-one-out cross validation
MAE	Mean absolute error
QSAR	Qualitative structure-activity relationship

Electronic supplementary material The online version of this article (doi:10.1007/s11095-013-1023-6) contains supplementary material, which is available to authorized users.

X.-W. Zhu · S.-S. Liu (✉)
Key Laboratory of Yangtze River Water Environment,
Ministry of Education, College of Environmental Science & Engineering
Tongji University, 417 Mingjing Building Shanghai 200092, China
e-mail: ssliuhl@263.net

X.-W. Zhu · A. Sedykh · H. Zhu · A. Tropsha
Division of Chemical Biology and Medicinal Chemistry
Eshelman School of Pharmacy
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27514, USA

H. Zhu
Department of Chemistry, Rutgers University, 315 Penn Street
Camden, New Jersey 08102, USA

H. Zhu
The Rutgers Center for Computational and Integrative Biology
Rutgers University, Camden, New Jersey 08102, USA

A. Tropsha (✉)
100K Beard Hall, University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7568, USA
e-mail: alex_tropsha@unc.edu

R^2	Coefficient of determination
RED	Rapid equilibrium dialysis
RF	Random forest
RMSE	Root mean square error
SVM	Support vector machine
Tc	Tanimoto similarity coefficient

INTRODUCTION

Many small molecules, such as drugs and drug-like compounds in blood circulation form complexes with plasma proteins. The affinity of drugs to plasma proteins varies tremendously directly affecting the free drug concentration and pharmacokinetics (1). Typically, the drug - plasma protein complex serves as a drug reservoir while the drug is eliminated from the body (2). Accurate assessment of small molecules' binding to plasma proteins is necessary for all aspects of ADME-Tox (absorption, distribution, metabolism, excretion, toxicity) (3) in the context of both drug discovery and in chemical risk assessment (4).

Human plasma proteins are made up of albumin and globulin among which the human serum albumin, alpha-1-acid glycoprotein, and lipoproteins are the most abundant (5). Experimentally, rapid equilibrium dialysis (RED) is a conventional method to determine a drug bound to plasma proteins (6) in a simulated *in vivo* environment (*e.g.*, protein composition and concentration, body temperature *etc.*). However, RED and other techniques can still be time-consuming and expensive if applied to every candidate compound in the early drug discovery stage.

As one of the most efficient computational tools, Quantitative Structure-Activity Relationships (QSAR) modeling is widely applied to find statistical relations between chemical structural features and a particular biological activity. There have been several attempts to correlate experimental plasma protein binding values with chemical structural features. Hall *et al.* (7) modeled the binding of 115 beta-lactams to human plasma proteins using multiple linear regression resulting in a model with mean absolute error (MAE) in ten-fold cross-validation of 10.9%. Lobel *et al.* (8) reported models with R^2 (coefficient of determination) of 0.68 and 0.51 for training (226 compounds) and test (94 compounds) sets, respectively. Yamazaki et el. (9) developed nonlinear regression models for a set of 300 compounds using pH-dependent octanol-water partition coefficient (LogP) as physicochemical parameters, resulting in R^2 of 0.83 for an external validation set of 20 compounds. Votano *et al.* (10) compiled a diverse dataset of about 1,000 drugs and drug-like compounds with experimental plasma protein binding values. In their study, artificial neural network and support vector machine (SVM) modeling yielded the lowest MAE value of 14.1% and the highest MAE value of 18.3%, respectively, for a validation set of 200 compounds. (For a detailed review of those studies please see the report of Hall *et al.* (11)). Moreover, since the 3D crystal structure of human serum albumin (HSA) is available, structure-based modeling strategies have been employed as well (12). However, mostly due to multiple possible binding sites on HSA, earlier studies were usually limited to small sets of specific chemicals (13) often lacking rigorous external validation. Furthermore, earlier studies lacked special emphasis on accurately predicting highly bound compounds (11,14,15), which is highly important because strong plasma protein binding (90~100%) is often a desirable property in pre-clinical drug screening (14).

In this study, a set of 1,242 compounds with known %PPB was compiled and curated from public sources. To our knowledge, this is the largest human plasma protein binding dataset available publicly. Using this dataset, QSAR models were developed and externally validated. In addition, a set of 173 compounds from DrugBank, and a set of 236 ToxCast chemicals with %PPB values measured using high-throughput screening bioassays were also used to validate our models.

MATERIALS AND METHODS

Modeling Dataset

A set of 1,242 unique compounds with known %PPB (see Supplementary Material Table S4) was compiled and curated from two major sources: the work of Votano *et al.* (10) and the database for pharmacokinetic properties (16). According to the activity histogram (see Supplementary Material Fig. S1), the distribution of original %PPB values is heavily skewed toward highly bound range. We have transformed %PPB into a pseudo-equilibrium constant parameter (lnKa) and observed that the distribution of transformed values became normal-like (see Supplementary Material Fig. S1). The transformation equation (Eq. 1) (for derivation, see Supplementary Material Equations SE. 1–7) is given below:

$$\ln Ka = C \cdot \ln\left(\frac{f_b}{1 - f_b}\right) \tag{1}$$

Where f_b is %PPB × 0.01, and C is a constant set to 0.5. Note that similar transformations have been utilized in previous studies (2,8,11), but the ensuing advantages were not fully explored or discussed.

DrugBank Dataset

A set of drugs or drug-like compounds was curated from DrugBank v3.0 (http://www.drugbank.ca/) that contains plasma protein binding data in a textual form, often as a range of values or a qualitative description. After transforming these into numerical %PPB values, we obtained a set of 173 unique compounds not present in the modeling dataset (see Supplementary Material Table S5).

ToxCast Dataset

A set of 236 unique chemicals with %PPB values measured in a high-throughput screening assay by Wetmore *et al.* (17,18) was obtained from the US EPA ToxCast Phase I project. The ToxCast chemicals are mainly pesticides and were not present in our modeling dataset (see Supplementary Material Table S6).

Chemical Structure Curation

Chemical structures of all employed compounds were curated according to our standard procedure described elsewhere. (19) Briefly, canonical SMILES code for all compounds was generated by ChemAxon Standardizer (v.5.3, ChemAxon, Budapest, Hungary) to normalize chemotypes (neutralization, tautomerization, aromatization, 2D structure cleaning, counter-ions removal) and to remove duplicate structures.

Molecular Descriptors

A set of chemical descriptors annotated as "two-dimensional" was calculated using Dragon (v.5.5, Talete SRL, Milan, Italy). It comprises constitutional and topological descriptors, walk and path counts, connectivity indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, functional group counts, atom-centered fragments, molecular properties, and 2D fingerprints. All descriptors were range-scaled to [0, 1] interval. We then removed descriptors with low variance (standard deviation <0.001) or with high redundancy (if pairwise $R^2>0.90$, one of the pair was randomly removed). A final set of 880 Dragon descriptors was used for QSAR modeling and for the estimation of structural diversity. The modeling dataset, with average pairwise Tanimoto similarity coefficient of 0.58 (see Supplementary Material Fig. S2), is chemically diverse and should have substantial coverage of the chemical space.

k Nearest Neighbors (*k*NN)

This method employs the *k* nearest neighbors' prediction principle with a variable selection procedure (20). In this study, a genetic algorithm was used to drive the variable selection (with a population consisting of 500 solutions, each ranging from 5 to 40 descriptors). The models were evaluated by internal leave-group-out cross-validation (LGO-CV) where a fraction of compounds (~ 20%) is removed from the modeling set and their biological activity was predicted as the weighted average of *k* nearest molecular (*k* was varied from 1 to 6). Individual models were considered acceptable if their LGO-CV R^2 was greater than 0.60.

Random Forest (RF)

Random Forest is an ensemble of unpruned classification or regression trees created on bootstrap samples of the training data and random subsets of descriptors (m_{try}) for tree induction (21). In this study, we used regression RF with the default parameters (number of trees$=$500 and one-third of the total number of descriptors for m_{try}) (22).

Support Vector Machine (SVM)

The SVM regression approach finds in the descriptor-activity space the narrowest band containing most of the data points. In this study, we used LIBSVM with RBF kernel and the grid-search to optimize its cost and gamma parameters (23).

Applicability Domain (AD)

To avoid over-extrapolation of activity prediction, a global AD is introduced (Eq. 2) to control the distance between a predicted compound and its closest neighbor in the training set (should be less than D_T).

$$D_T = \bar{y} + Z\sigma \tag{2}$$

Here, $\bar{y}$ and σ characterize the training set and are, respectively, the average and standard deviation of the Euclidean distances between each compound and its nearest neighbor. Z is a user-controlled threshold, which in this study was varied from to 0.5 to 3.0 (24,25).

Model Robustness

Y-randomization is carried out to establish model's robustness (26). The process consists of randomly shuffling activity values of the modeling set multiple times and rederiving models with these shuffled values; the "random" model performance is then evaluated using both training and test sets. This procedure was repeated five times and the one-tailed *t*-test p-value was calculated, which is the probability to obtain random models with R^2_{ext} as high as models built with real activities. If the "p-value<0.05" condition is not satisfied, models built with the real data are deemed unreliable.

Modeling and Validation

QSAR models were developed following our standard modeling workflow (27) using %PPB and lnKa as target endpoints (Fig. 1). Briefly, according to the five-fold external cross-validation (5FCV) procedure, the modeling dataset was randomly split into five subsets of nearly equal size.

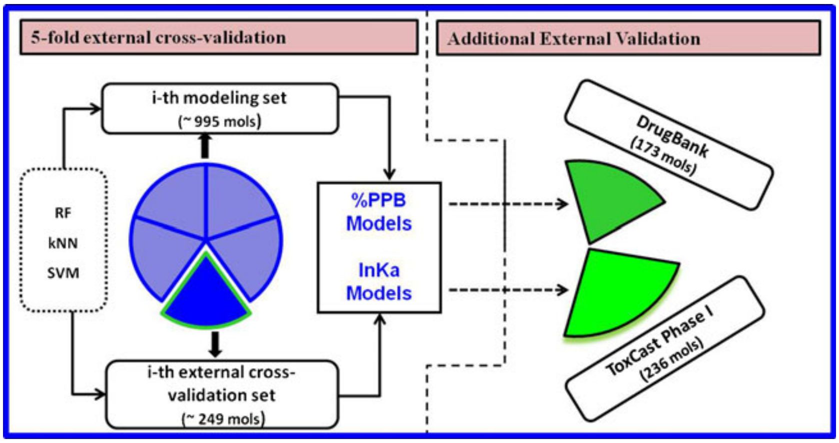


Fig. 1 Plasma protein binding modeling workflow based on five-fold external cross-validation and additional validation on DrugBank and ToxCast chemicals; %PPB and lnKa models are QSAR models of %PPB and lnKa endpoints, respectively.

Four of these were used for developing models and the remaining one - for the external validation of the models. This procedure was repeated five times till each of the five subsets served as a validation set once. The above procedure results in five sets of models (one for each fold) that are then used collectively for screening additional external sets; both the DrugBank and ToxCast chemicals were used to validate the external predictive power of our models. The prediction results of the lnKa models were converted back to fraction bound to facilitate the comparison between two endpoint representations.

RESULTS AND DISCUSSION

QSAR models based on both modeling target endpoints (*i.e.*, %PPB and lnKa) were developed using three modeling approaches (*k*NN, RF, and SVM). All the prediction results were converted into fraction bound (%PPB) for proper comparison. We investigated the effect of varying the threshold value of the AD (Z-parameter from 0.5 to 3, Eq. 2) on the chemical space coverage and prediction accuracy of our models. We found that different AD thresholds yield result with nearly the same prediction accuracy, while the coverage drops significantly (see Supplementary Material Table S1). Thus, all the statistical results listed were based on the largest AD ($\mathcal{Z}=3$). Five-fold external cross-validation prediction performance for the modeling dataset of 1,242 curated compounds is summarized in Table I. Our five-fold external cross-validation yielded MAEs of 15.8−16.3% for %PPB models (*i.e.*, those developed using %PPB as the modeling endpoint) and MAEs of 14.2−14.5 % for lnKa models (Table I). All models developed with real data were found to have significantly better statistical characteristics (*cf.* Table I) than Y-randomized models ($\text{MAE}=35.1-36.9$, $R^2=0.001-0.002$, $n=5$). Consensus prediction was calculated as the average prediction of the models built using RF, *k*NN, and SVM methods. It was slightly better (MAE for lnKa models is 13.6±0.8%) than individual models. These statistical results are comparable or better than those reported previously (11), while based on a larger database size.

Additional Validation of Models

Two datasets (DrugBank and ToxCast chemicals) were used for additional external validation to verify the predictability of our models (see Fig. 1). Since the protein binding of some DrugBank compounds is reported as a range or even as a qualitative description, we represented those compounds in accordance with their original descriptions (*e.g.*, intervals) when comparing with predicted values (Figs. 2 and 3). The prediction error for the set of 236 ToxCast chemicals compounds is higher than that of the DrugBank compounds (Table I). This difference in validation performance might be due to different measuring method employed. The protein binding values for the set of 173 DrugBank drug or drug-like compounds were measured using the same conventional methods as the data in the modeling dataset. In contrast, the fraction bound of ToxCast chemicals was measured using high-throughput screening technique (28), which is different from the conventional methods such as equilibrium dialysis. The experimental variance of conventional methods were estimated to be 0.01~10.3% for various compounds (6), which can be used as the estimation of the lowest prediction error we can ever expect from modeling such data. However, the experimental variance for the ToxCast chemicals is not available.

Comparison Between Two Representations of the Plasma Protein Binding Values

As shown in Table I, lnKa models outperform %PPB models on the modeling data and on two external validation sets. To analyze the distribution of prediction errors further, we plotted MAE values as a function of experimental PPB%

Table I External Prediction Performance for the Modeling Set and for the DrugBank and ToxCast Chemicals[a]

Method	Endpoint	Modeling set (5-fold external cross-validation)			DrugBank			ToxCast	
		R^2 [b]	MAE	RMSE	R^2	MAE	RMSE	MAE	RMSE
kNN	%PPB	0.63 ± 0.04	15.8 ± 0.2	20.6 ± 0.2	0.68 ± 0.03	14.3 ± 0.5	18.6 ± 0.6	18.4 ± 0.5	23.5 ± 0.7
	lnKa	0.62 ± 0.04	14.5 ± 0.2	20.8 ± 0.3	0.61 ± 0.03	14.0 ± 0.6	20.9 ± 0.8	14.9 ± 0.6	22.6 ± 0.9
RF	%PPB	0.62 ± 0.04	16.1 ± 0.2	20.7 ± 0.2	0.64 ± 0.03	14.6 ± 0.5	19.5 ± 0.7	20.6 ± 0.5	25.2 ± 0.6
	lnKa	0.62 ± 0.04	14.4 ± 0.2	21.0 ± 0.3	0.68 ± 0.03	12.7 ± 0.5	18.7 ± 0.7	14.9 ± 0.5	22.5 ± 0.9
SVM	%PPB	0.62 ± 0.04	16.3 ± 0.2	20.7 ± 0.2	0.63 ± 0.03	14.7 ± 0.5	19.8 ± 0.8	20.6 ± 0.6	26.6 ± 0.8
	lnKa	0.63 ± 0.04	14.2 ± 0.2	20.7 ± 0.3	0.70 ± 0.03	11.9 ± 0.5	18.2 ± 0.8	16.5 ± 0.6	25.1 ± 1.0
Consensus	%PPB	0.67 ± 0.04 [c]	15.3 ± 0.2	19.4 ± 0.2	0.67 ± 0.03 [d]	14.3 ± 0.5	18.7 ± 0.6	19.2 ± 0.5	24.4 ± 0.7
	lnKa	0.66 ± 0.04 [e]	13.6 ± 0.2	19.7 ± 0.3	0.68 ± 0.03 [f]	12.5 ± 0.5	18.7 ± 0.7	14.8 ± 0.6	22.6 ± 0.9

[a] Shown values are mean ± standard error based on 20%-off bootstrapping ($n = 1{,}000$). Consensus was calculated as average prediction across three machine learning methods: random forest (RF), *k* nearest neighbors (kNN), and support vector machine (SVM); [b] R^2 - coefficient of determination (not shown for Toxcast as it is heavily skewed toward highly bound range); MAE - mean absolute error; RMSE –root mean square error; [c] with slope = 1.15 and intercept = −9.1%PPB; [d] with slope = 1.10 and intercept = −6.0%PPB; [e] with slope = 0.92 and intercept = 3.3%PPB; [f] with slope = 0.91 and intercept = 4.1%PPB

values (represented by 10 activity bins) for each of the three datasets (Fig. 2). For the modeling dataset (1,242 compounds), the lnKa models outperform the %PPB models for the low (%PPB < 30%) and relatively high (%PPB > 80%) protein binding (~700 compounds in both), while the %PPB models outperform the lnKa models for the medium binding range (%PPB = 30–80%; ~400 compounds) (Fig. 2a). The prediction errors for the DrugBank (Fig. 2b) and ToxCast (Fig. 2c) chemicals show a similar trend.

Although there is no unanimous quantitative definition of the highly bound fraction, here we used "%PPB > 90%" as such, which is the same as FDA's definition in its draft guideline on hepatic impairment (29). Consequently, 469 out of 1,242 curated compounds in the modeling dataset were defined as highly bound and their cross validated MAEs were 12.9% and 7.6% for %PPB models and lnKa models, respectively. Likewise, the MAEs of 74 highly bound DrugBank chemicals were 11.9% and 6.2% for %PPB and lnKa models and the MAEs of 156 highly bound ToxCast chemicals were 17.6% and 10.7% for %PPB and lnKa models. That is, for the five-fold external cross-validation and two additional validation cases, the prediction error for highly bound compounds was significantly lower ($p < 0.01$ by permutation test; $n = 10{,}000$) for lnKa models.

This performance difference for the two representations of the target endpoints can be further emphasized by classifying compounds into three categories and then by examining the prediction accuracy for each category. We defined "weakly bound" category as %PPB < 32%, "bound" category as %PPB from 32% to 90%, and compounds with fraction bound ≥ 90% as "highly bound". Our scheme is a modified version of the scheme by Saiakhov *et al.* (30), where %PPB > 32% was defined as "bound", %PPB < 19% as "non-bound", and in-between values as "intermediate". The Classification accuracy for each category and overall correct classification rate (CCR) for both %PPB models and lnKa models are shown in Table II (see also Supplementary Material Table S2). The prediction accuracy of the lnKa models for the highly bound categories exceeds that of %PPB models by 20–40% and for weakly bound – by 10–27%, while for the medium category %PPB models have higher accuracy by about 20%. When lnKa is used instead of %PPB as the modeling endpoint, the overall

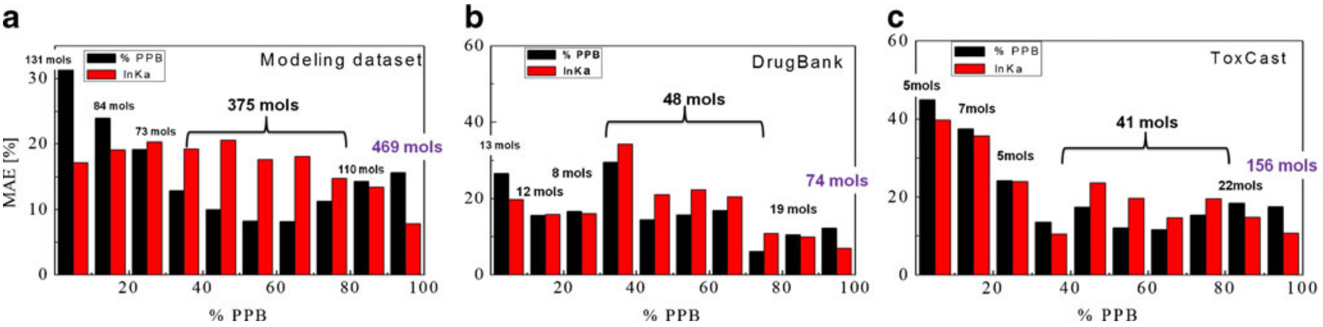


Fig. 2 Distribution of mean absolute errors (MAE) of %PPB models (*black bars*) and lnKa models (*red bars*) predictions for 1,242 modeling set compounds (**a**), 173 DrugBank compounds (**b**), and 236 ToxCast chemicals (**c**). Predictions for each endpoint are based on the consensus of respective kNN, RF, and SVM models.

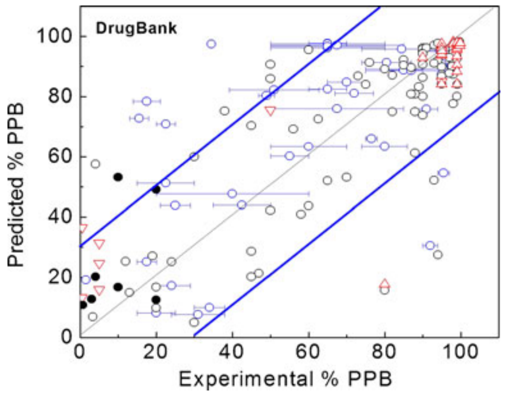


Fig. 3 Prediction results for 173 DrugBank compounds based on the overall consensus of %PPB and lnKa models. Data points with experimental %PPB reported as a range are shown as *blue bars* and *circles*; *unfilled black circles* −%PPB reported as exact values; *filled black circles* – qualitative reports; *down- and upward pointing red triangles* –%PPB reported as "less than *x*" or "greater than *x*" values, respectively. *Gray diagonal* corresponds to $y = x$ correlation line ($R^2 = 0.67$). *Blue diagonals* are 30% off-sets from experimental %PPB values.

CCR increases by 7.5%, 8.8%, and 11.6% for the DrugBank, ToxCast, and modeling sets respectively.

Superior prediction accuracy of lnKa models for highly-bound compounds has important practical implications, because many of the prescribed drugs fall into that group (31). Small changes in fraction bound for those drugs may cause large changes in their free concentration *in vivo* (15). Furthermore, in an *in vivo* system, a compound with high fraction bound will have a low plasma concentration which leads to its slower clearance and longer half-life time (31). Therefore, prediction errors for highly bound drugs will have larger impact on the subsequent estimation of their *in vivo* pharmacokinetic parameters. Our results indicate that the lnKa models reported in this study are more suitable for practical ADMET calculations than previously reported models (7,10,30).

Interpretation of QSAR Models

Interpretation of QSAR models in terms of the important chemical features can be useful for designing new drug candidates with desired properties. We ranked descriptors by their importance in our lnKa *k*NN and RF models. For ranking, in case of *k*NN, we used descriptor frequency of occurrence in the individual models of the *k*NN ensemble (32,33). In case of RF, we used the mean decrease in accuracy after random permutations of descriptor's values, which can affect multiple decision trees of the forest (22). Ranked top 10 descriptors for each of the *k*NN-lnKa and RF-lnKa models are shown in Supplementary Material Table S3 and six out of them (ALOGP, ALOGP2, MLOGP, BLTA96, Ui and nBM) are actually present in

Table II Classification Accuracies for the Modeling set (5FCV) and for the DrugBank and ToxCast Chemicals [a]

Endpoint	Modeling set (5FCV)				DrugBank				ToxCast			
	WB [b]	B [c]	HB [d]	CCR [e]	WB [b]	B [c]	HB [d]	CCR [e]	WB [b]	B [c]	HB [d]	CCR [e]
%PPB	45.6 ± 1.4	91.2 ± 0.7	31.1 ± 1.1	55.9 ± 0.6	58.7 ± 4.4	86.3 ± 2.4	38.2 ± 2.8	61.1 ± 1.9	22.0 ± 5.4	93.5 ± 1.6	24.8 ± 1.9	46.7 ± 2.0
lnKa	59.7 ± 1.3	77.0 ± 1.0	66.0 ± 1.2	67.5 ± 0.7	70.9 ± 4.1	65.7 ± 3.2	69.1 ± 2.6	68.6 ± 2.0	39.2 ± 6.0	73.7 ± 3.0	53.5 ± 2.1	55.5 ± 2.4

[a] Classification accuracies (in%) are based on the consensus predictions from Table I, values are mean ± standard error based on 20%-off bootstrapping ($n = 1,000$); classification categories are as follows: [b] weakly bound (0≤%PPB <32%), [c] Bound (32≤%PPB <90%), [d] highly bound (90≤%PPB ≤100%); [e] correct classification rate, which is average classification accuracy across all three categories, the expected CCR by random chance is 33.3%

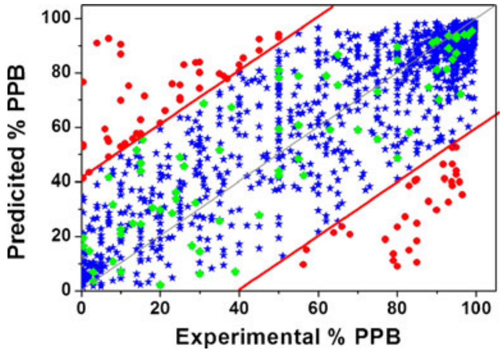


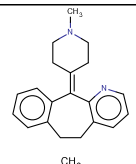
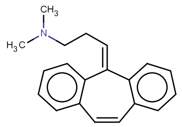
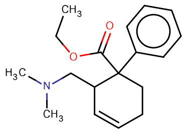
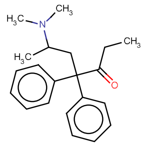
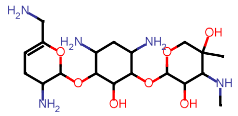
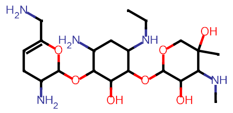
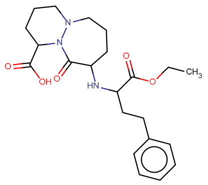
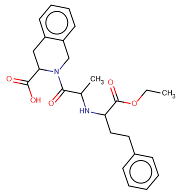
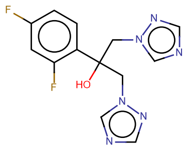
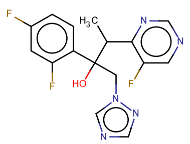
Fig. 4 QSAR modeling results for the 5 fold cross-validation dataset of 1,242 compounds based on consensus prediction of lnKa models (RF, kNN, and SVM). *Gray diagonal* corresponds to $y = x$ correlation line ($R^2 = 0.65$). *Red diagonals* are 40% off-sets (*red lines*) from experimental %PPB values, *Red dots* represent 76 compounds with large prediction errors (MAE > 40%), their corresponding nearest neighbors (*green diamonds*) have average MAE of 14.3%.

both lists. Lipophilicity is considered a major determinant of nonspecific protein binding (10,34), and indeed our results demonstrate that values of descriptors representing octanol-water partition coefficient (*i.e.*, ALOGP, ALOGP2, and MLOGP) are higher for strong binders than for weak binders (see Supplementary Material Fig. S3). Furthermore, the same trend is observed for descriptors representing unsaturated bonds (*i.e.*, Ui and nBM) and, by extension, hydrophobicity (see Supplementary Material Fig. S3).

Compounds Mispredicted by Individual Models

In effort to understand limitations of the models it is helpful to analyze compounds with large prediction errors. There were 76 compounds with prediction errors large than 40%PPB (Fig. 4 and Supplementary Material Table S7). However, the average MAE of their corresponding nearest neighbors is only 14.3%. Table III shows five examples from that list. The predicted %PPB values for azatadine, tilidine, sisomicin, cilazapril, and fluconazole are different from their corresponding experimental values, but close to the experimental and predicted %PPB of their individual nearest neighbors (cyclobenzaprine, methadone, netilmicin, quinapril, and voriconazole, respectively). The pairs of similar compounds with large difference in activity (so-called "activity cliffs") present challenges for QSAR model

Table III Five Compounds with Large Prediction Errors and Their Corresponding Nearest Neighbors

Compound	Structure	Exp. PPB%	Pred. PPB% [a]	Tc [b]	Nearest Neighbor	Structure	Exp. PPB %	Pred. PPB% [a]
Azatadine		10.0	86.8	0.94	Cyclo-benzaprine		93.0	92.6
Tilidine		25.0	70.0	0.92	Methadone		89.0	82.2
Sisomicin		85.0	10.5	0.99	Netilmicin		10.0	22.6
Cilazapril		25.0	75.5	0.93	Quinapril		97.0	88.5
Fluco-nazole		21.2	66.1	0.85	Voriconazole		58.0	66.2

[a] consensus of RF, kNN, and SVM models for lnKa as a target endpoint (Table I); [b] Tanimoto similarity coefficient between each compound and its corresponding nearest neighbor based on 2D Dragon descriptors

development (35); they can also point to data errors, when structures or activity values are wrong. During the dataset curation, we found compounds with contradictory %PPB values reported in different literature sources. For example, the fraction bound of biotin was reported as 80% in one publication (10), but 20% in another (36). When we checked the original experimental work, we found that "approximately 12% of total biotin in plasma is covalently bound, 7% is reversibly bound, and 81% is free" (36). It is obvious that free biotin fraction was mistaken as bound in the first source. On the other side, there are also some compounds that are likely to be true activity outliers. For example, the experimental %PPB value for fluconazole was reported as 21.2%, 21.8%, and 11% by three different measuring methods (6). However its predicted %PPB is 66.1%, which is close to that of its nearest neighbor of voriconazole (Table III, (37,38)). This discrepancy could be due to specific interactions of fluconazole that are missed by our models; hence, additional data on similar compounds may be needed to update the models.

CONCLUSION

In this study, we curated the largest publicly available plasma protein binding dataset and developed predictive QSAR models that were rigorously validated on diverse external sets containing both, drugs and industrial chemicals. We compared the results of modeling plasma protein binding using two modeling target endpoints: fraction bound (*i.e.*, %PPB) and lnKa ("binding constant"-like parameter). We found that lnKa models achieve higher prediction accuracy for highly bound compounds: The MAEs were 7.6%, 6.2% and 10.7% for the highly bound compounds in the modeling dataset, DrugBank, and ToxCast, respectively. The computational models developed in this study can accurately predict plasma protein binding of new chemicals, especially so for suspected strong binders, which is crucial for practical ADMET applications. Therefore, models developed herein can serve as useful virtual screening tools in human health risk assessment (for example, for toxicokinetic adjustment of estimated levels of exposure) and in early drug development. All models developed in this study are available for open access at our Chembench web-server (http://chembench.mml.unc.edu).

ACKNOWLEDGMENTS AND DISCLOSURES

We thank Drs. Alexander Golbraikh and Denis Fourches for their useful comments on this study. This work was supported, in part, by grants from National Institutes of Health (GM66940 and R21GM076059), The Johns Hopkins Center for Alternatives to Animal Testing (20011-21), and the National Natural Science Foundation of China (20977065).

REFERENCES

1. Bow DAJ, Perry JL, Simon JD, Pritchard JB. The impact of plasma protein binding on the renal transport of organic anions. J Pharmacol Exp Ther. 2006;316(1):349–55.
2. Kratochwil NA, Huber W, Müller F, Kansy M, Gerber PR. Predicting plasma protein binding of drugs: a new approach. Biochem Pharmacol. 2002;64(9):1355–74.
3. Mager DE, Xu C. Quantitative structure-pharmacokinetic relationships. Expert Opin Drug Met. 2011;7(1):63–77.
4. Banker MJ, Clark TH. Plasma/serum protein binding determinations. Curr Drug Metab. 2008;9(9):854–9.
5. Kuchinskiene Z, Carlson LA. Composition, concentration, and size of low density lipoproteins and of subfractions of very low density lipoproteins from serum of normal men and women. J Lipid Res. 1982;23(5):762–9.
6. Waters NJ, Jones R, Williams G, Sohal B. Validation of a rapid equilibrium dialysis approach for the measurement of plasma protein binding. J Pharm Sci. 2008;97(10):4586–95.
7. Hall LM, Hall LH, Kier LB. QSAR modeling of beta-lactam binding to human serum proteins. J Comput Aided Mol Des. 2003;17(2):103–18.
8. Lobell M, Sivarajah V. In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pK(a) and AlogP98 values. Mol Divers. 2003;7(1):69–87.
9. Yamazaki K, Kanaoka M. Computational prediction of the plasma protein-binding percent of diverse pharmaceutical compounds. J Pharm Sci. 2004;93(6):1480–94.
10. Votano JR, Parham M, Hall LMH, Kier LB, Oloff S, Tropsha A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. J Med Chem. 2006;49(24):7169–81.
11. Hall LM, Hall LH, Kier LB. Methods for predicting the affinity of drugs and drug-like compounds for human plasma proteins: a review. Curr Comput-Aid Drug. 2009;5(2):90–105.
12. Zsila F, Bikadi Z, Malik D, Hari P, Pechan I, Berces A, *et al.* Evaluation of drug–human serum albumin binding interactions with support vector machine aided online automated docking. Bioinformatics. 2011;27(13):1806–13.
13. Li H, Chen Z, Xu X, Sui X, Guo T, Liu W, *et al.* Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). Biopharm Drug Dispos. 2011;32(6):333–42.
14. Zhang F, Xue J, Shao J, Jia L. Compilation of 222 drugs' plasma protein binding data and guidance for study designs. Drug Discov Today. 2012;17(9–10):475–85.
15. Pellegatti M, Pagliarusco S, Solazzo L, Colato D. Plasma protein binding and blood-free concentrations: which studies are needed to develop a drug? Expert Opin Drug Metab Toxicol. 2011;7(8):1009–20.
16. Moda TL, Torres LG, Carrara AE, Andricopulo AD. PK/DB: database for pharmacokinetic properties and predictive in silico ADME models. Bioinformatics. 2008;24(19):2270–1.
17. Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, Freeman K, *et al.* Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicol Sci. 2012;125(1):157–74.
18. Wetmore BA, Wambaugh JF, Ferguson SS, Li L, Clewell HJ, Judson RS *et al.* The relative impact of incorporating pharmacokinetics on predicting *in vivo* hazard and mode-of-action from high-throughput *in vitro* toxicity assays. Toxicol Sci. 2013. doi:10.1093/toxsci/kft012.
19. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics

and QSAR modeling research. J Chem Inf Model. 2010;50(7):1189–204.

20. Zheng WF, Tropsha A. Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Model. 2000;40(1):185–94.
21. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
22. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Model. 2003;43(6):1947–58.
23. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM TIST. 2011;2(3):1–27.
24. Golbraikh A, Shen M, Xiao ZY, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des. 2003;17(2–4):241–53.
25. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 2003;22(1):69–77.
26. Rucker C, Rucker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model. 2007;47(6):2345–57.
27. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6–7):476–88.
28. Kruhlak NL, Contrera JF, Benz RD, Matthews EJ. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. Adv Drug Deliv Rev. 2007;59(1):43–55.
29. CDER/CBER. Guidance for industry:pharmacokinetics in patients with impaired hepatic function: study design, data analysis, and impact on dosing and labeling. U.S. Department of Health and Human Services Food and Drug Administration. 2003.
30. Saiakhov R, Stefan L, Klopman G. Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. Perspect Drug Discov Des. 2000;19(1):133–55.
31. Smith DA, Di L, Kerns EH. The effect of plasma protein binding on *in vivo* efficacy: misconceptions in drug discovery. Nat Rev Drug Discov. 2010;9(12):929–39.
32. Zhang LY, Zhu H, Oprea TI, Golbraikh A, Tropsha A. QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. Pharm Res. 2008;25(8):1902–14.
33. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, *et al.* Use of *in vitro* HTS-derived concentration–response data as biological descriptors improves the accuracy of QSAR models of *in vivo* toxicity. Environ Health Perspect. 2011;119(3):364–70.
34. Gleeson MP. Plasma protein binding affinity and its pelationship to molecular structure: an in silico analysis. J Med Chem. 2006;50(1):101–12.
35. Maggiora GM. On outliers and activity cliffs why QSAR often disappoints. J Chem Inf Model. 2006;46(4):1535.
36. Mock D, Malik M. Distribution of biotin in human plasma: most of the biotin is not bound to protein. Am J Clin Nutr. 1992;56(2):427–32.
37. Sandherr M, Maschmeyer G. Pharmacology and metabolism of voriconazole and Posaconazole in the treatment of invasive aspergillosis: review of the literature. Eur J Med Res. 2011;16(4):139–44.
38. Kethireddy S, Andes D. CNS pharmacokinetics of antifungal agents. Expert Opin Drug Metab Toxicol. 2007;3(4):573–81.