European Journal of Pharmaceutical Sciences 204 (2025) 106946

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

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

The state-of-the-art machine learning model for plasma protein binding prediction: Computational modeling with OCHEM and experimental validation

Zunsheng Han [a], Zhonghua Xia [b], Jie Xia [a,*], Igor V. Tetko [b,c,**], Song Wu [a,*]

[a] State Key Laboratory of Bioactive Substance and Function of Natural Medicines, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China
[b] Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[c] BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany

ARTICLE INFO

Keywords:
Plasma protein binding
OCHEM
Machine learning
Prospective study
Retrospective study

ABSTRACT

Plasma protein binding (PPB) is closely related to pharmacokinetics, pharmacodynamics and drug toxicity. Existing models for predicting PPB often suffer from low prediction accuracy and poor interpretability, especially for high PPB compounds, and are most often not experimentally validated. Here, we carried out a strict data curation protocol, and applied consensus modeling to obtain a model with a coefficient of determination of 0.90 and 0.91 on the training set and the test set, respectively. This model (available on the OCHEM platform https://ochem.eu/article/29) was further retrospectively validated for a set of 63 poly-fluorinated molecules and prospectively validated for a set of 25 highly diverse compounds, and its performance for both these sets was superior to that of the other previously reported models. Furthermore, we identified the physicochemical and structural characteristics of high and low PPB molecules for further structural optimization. Finally, we provide practical and detailed recommendations for structural optimization to decrease PPB binding of lead compounds.

1. Introduction

Binding of drugs to plasma proteins is one of the important parameters in pharmacokinetics, as it can affect many key properties of drugs, e.g., distribution volume (Vss), drug-drug interaction (DDI), clearance rate (CL) and therapeutic index (TI) (Di, 2021; Smith et al., 2010; Lambrinidis et al., 2015). Drugs based on compounds with a high affinity to plasma proteins have an increased half-life, and in some cases higher doses of the drug may be required to achieve an effective concentration for treatment. In addition, drugs bind to plasma proteins competitively, and drugs with higher binding rates will occupy most of the plasma protein binding sites. This can give rise to DDI (Di, 2021). Drugs with high PPB values may influence the binding of other drugs to the same plasma proteins, resulting in an increase or decrease in the (free) plasma concentration of the other drugs, rendering them either toxic or ineffective. Situations like this mainly arise for drugs with narrow therapeutic windows, e.g., warfarin (Lambrinidis et al., 2015; Di et al., 2017). Thus, assessment of PPB is very important for the development of new drugs and the safe clinical use of drugs.

The level of binding of a drug to plasma proteins is usually evaluated as the PPB rate (PPB%) or free fraction (fu) (Seyfinejad et al., 2021; Vuignier et al., 2010). There are three commonly used methods (Vuignier et al., 2010; Dimitrijevic et al., 2023) for determining PPB, i.e., equilibrium dialysis (ED), ultrafiltration (UF) and ultracentrifugation (UC). The ED method is the gold standard and is often used as a reference for UF and UC. However, each of the three methods has its own advantages and disadvantages depending on the application (Dimitrijevic et al., 2023). For non-specific compounds with good adsorption, UC is the first choice, but the measurement costs are high. For compounds that are not stable in plasma, UF is preferred, although non-specific adsorption can interfere with measurements quite drastically. ED can be generally applied to most compounds, but its drawbacks, such as

* Corresponding authors at: Institute of Materia Medica, Chinese Academy of Medical Sciences, No. 2 Nanwei Road, Beijing 100050, China.
** Corresponding author at: Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich - German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
E-mail addresses: jie.william.xia@hotmail.com (J. Xia), itetkoai@gmail.com (I.V. Tetko), ws@imm.ac.cn (S. Wu).

https://doi.org/10.1016/j.ejps.2024.106946
Received 30 July 2024; Received in revised form 18 October 2024; Accepted 23 October 2024
Available online 28 October 2024

changes in initial equilibrium state, non-specific binding, volume transformation, Donnan effect and protein leakage, can influence the accuracy of measurements. In any case, experimental assays are complex, time-consuming, and expensive, while *in silico* prediction has the advantages of being economical, simple, and fast, and able to facilitate rapid screening of a large number of compounds, including those that have not yet been synthesized. Therefore, PPB prediction models based on machine learning can play an important role in accelerating drug development process and drug safety.

Over the past decade, significant advances have been made in ML-based drug discovery (Lambrinidis et al., 2015; Vallianatou et al., 2013). In recent years, a number of regression models using ML have been built for PPB prediction. Supporting Materials Table S1 summarizes some of the published models and their performance metrics (Votano et al., 2006; Zhu et al., 2013; Wang et al., 2017; Watanabe et al., 2018; Sun et al., 2018; Yuan et al., 2020; Jimenez-Luna et al., 2021; Lou et al., 2022; Khaouane et al., 2023; Pore and Roy, 2024). These models have made great progress in predicting PPB, but the performance of these models on test sets may still require improvement. Moreover, most of these models were not validated in prospective studies. It is also worth mentioning that none of the previous studies (with the exception of the study by Lou et al. (Lou et al., 2022) using IDL-PPBopt) discussed the substructure and physicochemical properties that are related to PPB activity.

In addition, the successful use of these models usually requires a significant amount of programming expertise, and many of the associated tools are not user-friendly. To address this issue, several web servers offer free PPB prediction services, e.g., ADMETlab3.0 (Li et al., 2024), admetSAR3.0 (H.B. Yang et al., 2019), DruMAP (Kawashima et al., 2023), Pangu Drug (Lin et al., 2022), pkCSM(Deep-PK) (Pires et al., 2015), PreADMET (Lee et al., 2003) have been developed. Although these platforms are easy to use, it appears that the model accuracies within these platforms have not been validated in prospective studies.

The On-line CHEmical database and Modelling environment (OCHEM) is an algorithm-rich, automatized, and simple model training and sharing platform (Sushko et al., 2011). In this study, we used OCHEM to train models for PPB prediction with a variety of ML algorithms and descriptors. Then, we selected several of the best-performing models to develop a consensus model (Zhu et al., 2008). In addition, we performed retrospective and prospective validation of the newly developed and previously published models on external datasets. Finally, we analyzed the molecular features that are associated with high and low PPB values. The workflow adopted in this study is shown in Fig. 1.

2. Data

2.1. Training and test data set

Human PPB data points were collected from several sources: (1) published data by Lou (Lou et al., 2022), Basant (Basant et al., 2016), Watanabe (Watanabe et al., 2018), Ingle (Ingle et al., 2016), etc.; and (2) from DrugBank for drugs on market and in clinical trials (https://go.drugbank.com/, accessed in May 2022). A total of 9500 data points were collected, including SMILES and multiple forms of PPB values (e.g., protein binding rate as 90 % or 0.9 or unbound ratio as 0.1 or 10 %). The compound data were carefully prepared according to the following steps: (1) de-duplication, and removing mixtures, inorganic and organometallic compounds; (2) stripping salts and water; (3) removing compounds with molecular weights >800 Da and rotatable bonds >20.

According to the histogram of %PPB (cf. Figure S1), the distribution of original %PPB values was heavily skewed toward the high-PPB region. To minimize the effects of this skew, %PPB data points were transformed into pseudo-equilibrium constant parameter values (LogIt) in OCHEM automatically (see Eq. (1)) as suggested elsewhere (Gao et al., 2008). After this conversion, the distribution became Gaussian-like (cf. Fig. 2A).

$$LogIt = Log\left(\frac{f_b}{1-f_b}\right) \quad (1)$$

where f_b is fraction bound. This transformation is better for addressing/investigating differences in compounds with very small (i.e., < 1 %) and very large (i.e., > 99 %) PPB values.

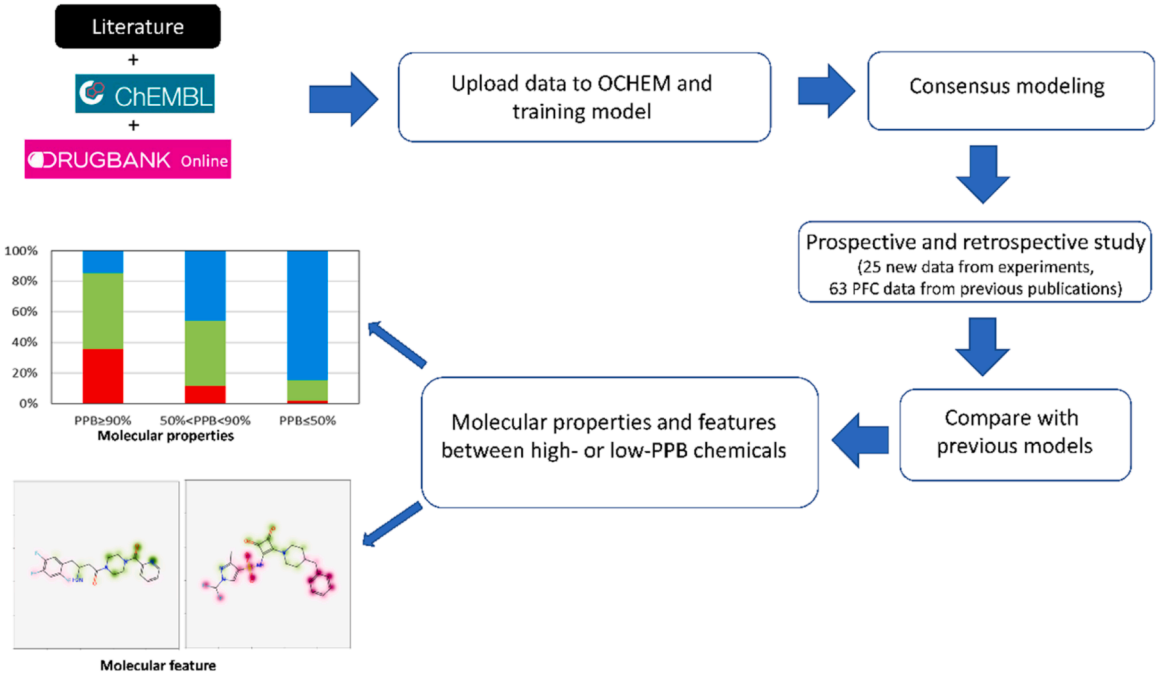

Fig. 1. The workflow to build the PPB prediction model.

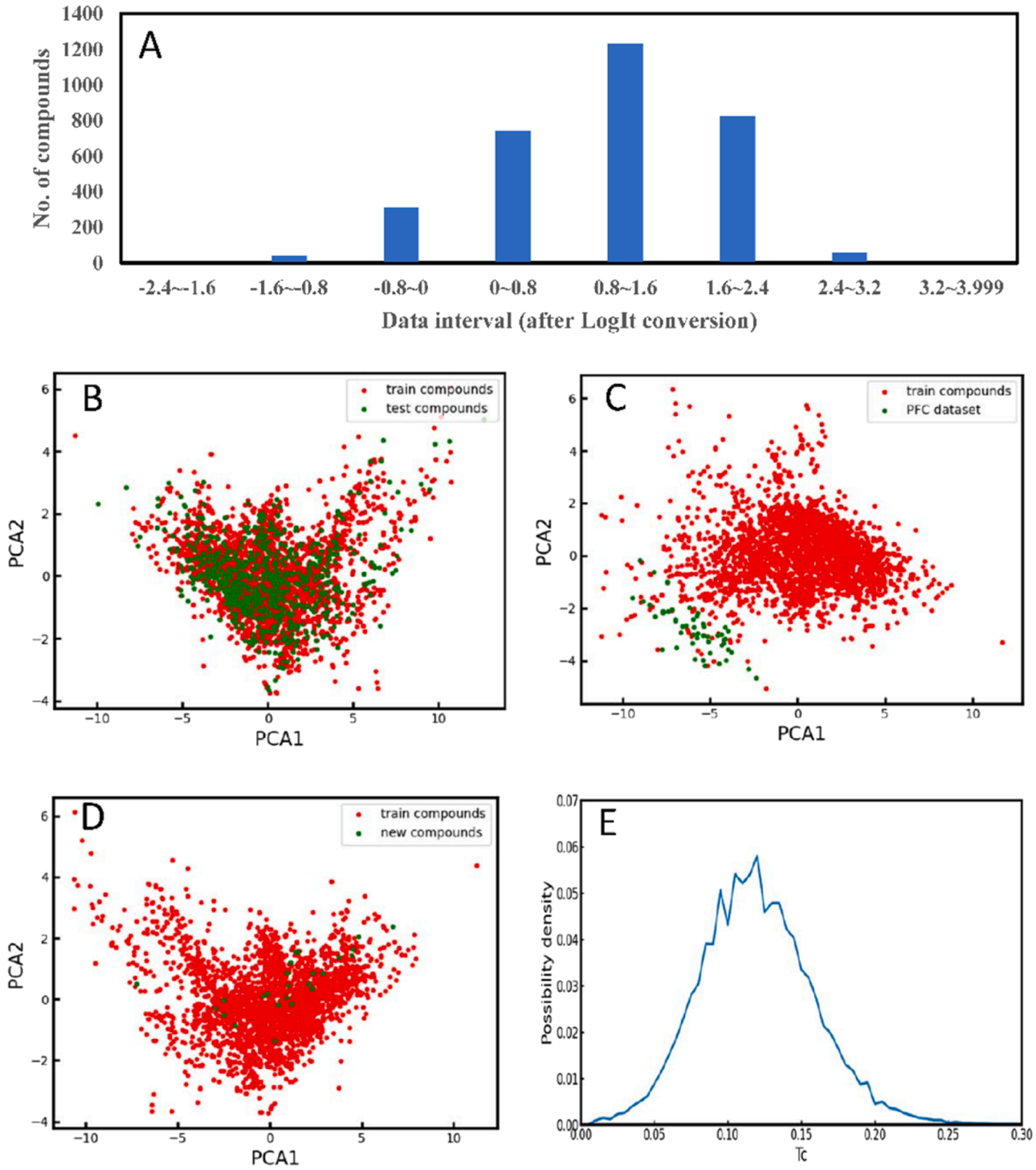

Fig. 2. (A) Distribution of LogIt transformation dataset, and PCA distribution plots: (B) Training and test dataset, (C) Training and PFC dataset, (D) Training and new dataset, (E) Pairwise Tanimoto coefficients distribution of 3214 compounds.

2.2. Retrospective and prospective validation data set

The dataset containing PPB values for Per- and Polyfluoroalkyl Substances was used for retrospective model validation (Smeltz et al., 2023). As the corresponding article was published in 2023, this dataset was not yet encountered by models published by other groups, nor by our own model, which collected data from earlier publications. We found that 5 out of 68 compounds from this dataset were also part of our training dataset. These molecules were excluded and the remaining 63 molecules were used as a validation set, which we called the "PFC" dataset. Furthermore, 36 compounds from the PFC dataset with PPB >99 % formed the "PFC99" set, which was used to compare methods for compounds that bind very strongly to plasma proteins.

A prospective study was performed to estimate model performance for a set of compounds selected using an experimental design method. The 10 K compounds of the ChemDiv diverse subset purchased by Institute of Materia Medica, Chinese Academy of Medical Sciences (structure information can be retrieved at Zenodo: https://zenodo.org/records/12641856), were clustered using the "Cluster Ligands" module of Discovery Studio (version 2016). Specifically, the fixed

"Number of Clusters" was set to 25 and FCFP_6 fingerprints (Rogers and Hahn, 2010) were used to represent compounds. The cluster centers (molecules) were selected to form the external prospective dataset, which we called the "new dataset". These were all new compounds and there was no overlap between these molecules and the modeling set (training, test and retrospective set). The new dataset (25 compounds) was determined by ED combined with HPLC or LC-MS/MS (see Supporting Materials) for detailed experimental methods.

The PPB values for retrospective and prospective dataset molecules were predicted using different computational platforms, including ADMETlab3.0 (Li et al., 2024) (https://admetlab3.scbdd.com), admetSAR3.0 (H.B. Yang et al., 2019) (http://lmmd.ecust.edu.cn/admetsar3), DruMAP (Kawashima et al., 2023) (https://drumap.nibiohn.go.jp), Pangu Drug (Lin et al., 2022) (http://pangu-drug.com), PreADMET (Lee et al., 2003) (https://preadmet.qsarhub.com/), pkCSM(Deep-PK) (Pires et al., 2015) (https://biosig.lab.uq.edu.au/deeppk/data) and our OCHEM model. The prediction accuracies of these platforms were compared with those of the PPB model developed in this study.

3. Methods

3.1. Analysis of machine learning approaches

We initially analyzed several machine learning approaches available in OCHEM, including traditional methods such as Partial Least Squares, Multiple Linear Regression Analysis, k-Nearest Neighbors as well as more advanced methods based on decision trees, such as Random Forest, XGboost, CatBoost, shallow and deep neural networks, which develop models based on calculated descriptors. Among all investigated methods, the Associative Neural Network (ASNN) method (Tetko, 2009) generally provided higher performances across analyzed descriptors and as such this method was selected for model development. The ASNN is a combination of an ensemble of single-hidden-layer neural networks and k Nearest Neighbors. It was inspired by thalamo-cortical organization (Villa et al., 1999) of the brain and has been shown to improve performance of the ensemble by correcting its bias using errors of the most similar records from the training set. Of the representation learning methods, we selected Transformer Convolutional Neural Networks (Transformer CNN) (Karpov et al., 2020), along with its variation, Transformer Convolutional Neural Fingerprint (CNF2) (Makarov et al., 2021) as well as two Graph Neural networks methods, i.e., attentive fingerprint algorithm (AttFP) (Xiong et al., 2020) and ChemProp (Yang et al., 2019a,b), both of which are implemented as part of the KGCNN (Reiser et al., 2021) package in OCHEM.

3.2. Assessment of model performance

Three statistical parameters were used to evaluate model performance (cf. Eqs. (2)-(4)). MAE and RMSE were used to evaluate the difference between the predicted values and the observed values. In addition to these parameters, the coefficient of determination R^2, which measures explained variance of the model, was also used. A good model usually has small MAE and RMSE values, and an R^2 value close to 1.

$$MAE = \frac{\sum_1^N |x_i - y_i|}{N} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_1^N (x_i - y_i)^2}{N}} \tag{3}$$

$$R^2 = 1 - \frac{\sum_1^N (x_i - y_i)^2}{\sum_1^N (x_i - \overline{x})^2} \tag{4}$$

where x_i is the observed value, y_i is the predicted value, $\overline{x}$ is the average of the observed values, N is the number of data points.

3.3. Applicability domain

The application domain (AD) is an important concept in quantitative-structure activity relationships (QSAR) (Weaver and Gleeson, 2008). The measurement of AD was based on the distance to model (DM) of the compound, where DM is a numerical measure proportional to the model's prediction uncertainty for a given compound (Tetko et al., 2008). A large DM value corresponds to a low prediction accuracy for the target compound. The Consensus standard deviation (STD) of the predicted outcome was used to evaluate AD. On the OCHEM web site, the DM which covers 95 % of compounds from the training set is used to define AD of the model. In the PPB model, 95 % of compounds have STD values <0.33, so we chose 0.33 as the threshold for evaluating AD, that is, when STD is <0.33 for a prediction of a new compound, we consider the compound to be in the AD range. Each prediction was also given a confidence interval to help users judge the reliability of the prediction.

3.4. Feature analysis related to PPB

3.4.1. Physicochemical descriptors analysis

At first, the physicochemical descriptors of all compounds (3214) were calculated using the Mordred package (1.2.0) (Moriwaki et al., 2018), and then the Person correlation coefficient (r^2) of each descriptor with PPB was calculated. Lastly, top 17 descriptors with $r^2>0.4$ were used to plot the heat map.

In order to gain insight into the model's interpretability, the relationships between some of the representative descriptors with high, medium and low PPB values were analyzed, respectively. All compounds were divided into three categories according to their PPB values, i.e., low PPB class (≤ 50 %), medium PPB class (50–90 %), high PPB class (≥90 %). In each class, the compounds were further divided into 3 groups according to the descriptor being studied. Then, the number of compounds in each descriptor group in the categories of low, medium and high PPB was counted. The number of compounds in a certain category was defined as 100 %, and the percentage of molecules in each descriptor group in all compounds in this category was calculated. The data were converted into graphs for visualization after calculation. Taking the descriptor of SlogP as an example, the high PPB compounds were divided into three subgroups (SLogP≥3, 1<SLogP<3, SLogP≤1), and the proportion of compounds in each subgroup was determined. This facilitated the identification of physicochemical properties associated with high and low-PPB compounds.

3.4.2. Privileged substructures analysis with similarity map and SetCompare in OCHEM

A PPB classification model was built and was used to produce a similarity map (Riniker and Landrum, 2013). Specifically, the compounds from the initial set (n = 3128) were divided into high and low PPB sets with thresholds of PPB >90 % and PPB<50 % respectively. The compounds with PPB values between 50 % and 90 % were removed. Each compound was represented with Morgan2 fingerprints. The classification model was built based on the training set and was evaluated on the test set. In the similarity map (Riniker and Landrum, 2013), the atoms of each compound were marked with different colors according to the contribution value of the atom. The substructures composed of atoms with positive contributions or negative contributions were visualized.

The high/low PPB data were also analyzed using the SetCompare utility (Vorberg and Tetko, 2014) in combination with the Extended Functional Groups (Salmina et al., 2016) descriptor type. SetCompare uses hypergeometric distribution with Bonferroni correction to identify overrepresented descriptors amid compounds with high/low PPB data.

4. Results and discussion

4.1. The data set

After careful pre-processing, 3214 PPB data points were obtained. These data were randomly split into training (2571) and test (643) sets. The LogIt-transformed data exhibited a Gaussian-like distribution, as shown in Fig. 2(A). The chemical space of the training, test, PFC and new datasets are shown in Fig. 2(B-D) based on principal component analysis (PCA) of the 17 Mordred (Moriwaki et al., 2018) descriptors (see Supporting Materials). The scatter plot shows substantial overlap in chemical space between training and test, PFC, new compounds dataset. The Tanimoto coefficient (Tc) (Bajusz et al., 2015) based on Morgan2 fingerprints (implementation of ECFP4 (Rogers and Hahn, 2010) in RDKit package (open-source cheminformatics. https://www.rdkit.org/ (accessed 3 July 2024)) values were <0.2 (cf. Fig. 2(E)), indicating high chemical diversity in the PPB data set.

4.2. Models constructed in OCHEM

OCHEM offers a number of ML algorithms and various molecular representations. Almost all ML algorithms implemented in the platform were tested in this study. ASNN (Tetko, 2009) provided on average better performance compared to other descriptor-based machine learning methods and was used for further analysis. Amid ASNN models, the model based on Mordred descriptors achieved the highest accuracy for the training set (cf. Table 1). After comparing all models constructed in OCHEM, we selected those with RMSE equal or <0.33 to build the consensus model. They were five ASNN models developed with ALogPS-OEstate (Tetko et al., 2001), EPA (https://www.epa.gov/-comptox-tools/toxicity-estimation-software-tool-test), Fragmentor (Varnek et al., 2008), MOLD2 (Hong et al., 2008) and Mordred (2D) (Moriwaki et al., 2018) descriptors as well as four models based on representation learning, i.e., Transformer CNN (Karpov et al., 2020), CNF2 (Makarov et al., 2021), ChemProp (Yang et al., 2019a,b) and AttFP (Xiong et al., 2020) (as implemented in KGCNN (Reiser et al., 2021) package). The performance of the consensus model, calculated as a simple average of these selected models, was superior to any individual model for both training and external test sets (cf. Table 1 and Fig. 3). The use of simple average method was based on our previous efforts (using a much larger set with melting point data (Tetko et al., 2016)), which did not demonstrate a significant improvement in the performance of methods when using a weighted average of models. Nevertheless, we also analyzed the weighted average of models in this study, but we did not observe any changes in the model prediction accuracy.

4.3. Prospective and retrospective study: model performance and comparison with other models

For the prospective study, the PPB values of 25 new compounds were determined by ED experiment and predicted by each model. The structures and measured PPB values of these 25 compounds are shown in Fig. 4. The predictive performance parameters of the different platforms are listed in Table 2. The OCHEM consensus model ("Consensus in LogIt") achieved the highest accuracy compared to other published models, with the highest R^2 and the lowest MAE and RMSE.

For the retrospective study, we also compared the predictive performances of the models based on the "63 PFC" dataset (cf. Table 2 and Supporting Materials for more details) and found that the consensus model developed in this study also achieved higher accuracy for this set than the other platforms. In addition, the results for compounds with PPB>99 % ("36 PFC99" set) predicted by the OCHEM model had significantly better RMSE values (0.4 %) compared with the results obtained by other models, which had RMSE values 3 to 100 times higher for the same compounds. Accurately predicting compounds with

Table 1
Performance of individual and the consensus models.

ID	Method [Descriptors]	Data Set	R^2	RMSE	MAE
1	ASNN [ALogPS, OEstate]	Training	0.835 ± 0.006	0.328 ± 0.005	0.254 ± 0.004
		Test	0.85 ± 0.01	0.304 ± 0.01	0.232 ± 0.007
2	ASNN [Fragmentor]	Training	0.842 ± 0.006	0.321 ± 0.005	0.251 ± 0.004
		Test	0.86 ± 0.01	0.29 ± 0.01	0.228 ± 0.007
3	ASNN [Mordred]	Training	0.85 ± 0.006	0.312 ± 0.005	0.242 ± 0.004
		Test	0.85 ± 0.01	0.305 ± 0.009	0.24 ± 0.007
4	ASNN [MOLD2]	Training	0.832 ± 0.007	0.330 ± 0.006	0.255 ± 0.004
		Test	0.83 ± 0.01	0.33 ± 0.01	0.25 ± 0.008
5	ASNN [EPA]	Training	0.845 ± 0.006	0.318 ± 0.005	0.244 ± 0.004
		Test	0.85 ± 0.01	0.31 ± 0.01	0.24 ± 0.008
6	Transformer-CNN	Training	0.866 ± 0.005	0.292 ± 0.004	0.236 ± 0.004
		Test	0.873 ± 0.009	0.283 ± 0.007	0.227 ± 0.007
7	AttFP	Training	0.841 ± 0.006	0.317 ± 0.005	0.253 ± 0.004
		Test	0.850 ± 0.010	0.310 ± 0.009	0.245 ± 0.008
8	ChemProp	Training	0.862 ± 0.005	0.299 ± 0.005	0.236 ± 0.004
		Test	0.883 ± 0.008	0.277 ± 0.007	0.222 ± 0.006
9	CNF2	Training	0.853 ± 0.006	0.308 ± 0.005	0.247 ± 0.004
		Test	0.860 ± 0.010	0.298 ± 0.008	0.240 ± 0.007
10	Consensus in LogIt	Training	0.901 ± 0.004	0.252 ± 0.004	0.202 ± 0.003
		Test	0.907 ± 0.007	0.240 ± 0.007	0.196 ± 0.005
11	Consensus in % unit	Training	0.902 ± 0.006	7.0 ± 0.2	4.3 ± 0.1
		Test	0.90 ± 0.01	6.6 ± 0.4	4.1 ± 0.2

Fig. 3. The measured PPB vs the PPB predicted by the consensus model for the training and test sets.

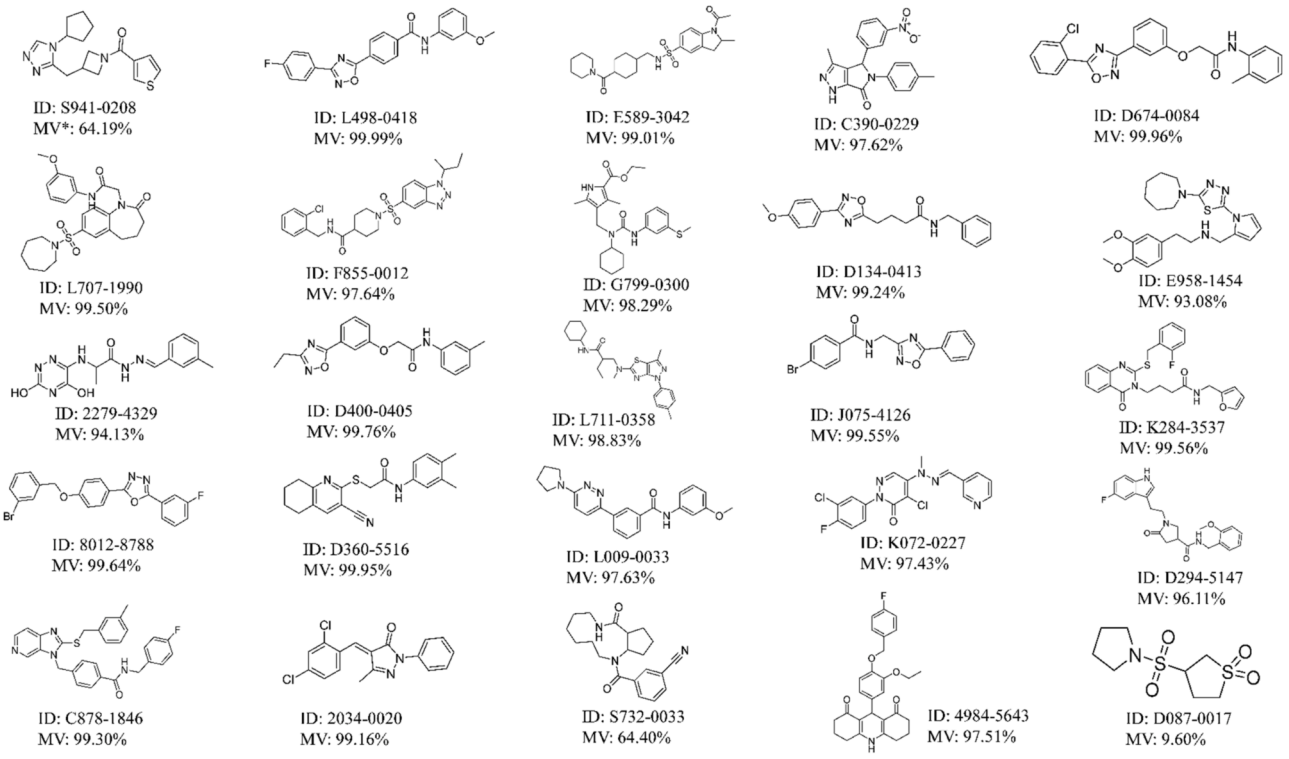

Fig. 4. The structures and measured PPB values of 25 new compounds. (*MV: measured value).

Table 2
The tested platforms' prediction performance for external validation sets using %PPB.

Dataset	Coefficient	ADMETlab v. 3.0	admetSAR v. 3.0	DruMAP	Pangu Drug	Deep-PK	PreADMET	OCHEM
25 (new)	R^2	0.71	0.79	0.77	0.80	0.45	0.57	0.93
	RMSE	10.49	9.01	12.23	9.08	18.33	14.20	5.80
	MAE	5.51	5.77	6.32	6.54	13.76	9.70	3.21
63 PFC	R^2	0.36	0.45	0.27	0.30	0.19	0.53	0.72
	RMSE	11.76	11.66	14.47	17.24	45.77	14.79	9.28
	MAE	4.84	8.10	11.23	9.67	43.55	7.63	3.56
36 PFC99	R^2	0.13	0.22	0.05	0.03	0.01	0.02	0.17
	RMSE	1.50	9.32	12.67	9.82	48.32	5.28	0.40
	MAE	1.09	5.72	10.07	6.21	47.18	3.27	0.27

PPB>99 % is very important and significant for drug discovery projects, because for binding rates of 99 %, 99.9 % or 99.99 %, the free drug concentrations will differ by a factor of 10 or 100. Thus, despite the fact that existing models had a relatively good overall prediction performance when using %PPB as an overall performance measure, most of them were not able to correctly predict compounds with very high PPBs.

4.4. Importance of LogIt function for predicting compounds with high PPB values

We used LogIt units to develop the individual and consensus models, and to estimate their performances. Of course, one can use different units, e.g., percentage or fraction, to estimate performance of the developed models by converting predicted and experimental values to the respective unit. For the consensus model, we first converted LogIt predicted by each individual model to a percentage (%) and then built a consensus model by averaging predictions of individual models given in percentage, i.e., "LogIt-% model". As shown in Table 1, this conversion had a generally negative impact on the models: it decreased R^2 and also led to widened confidence intervals for both RMSE and MAE (cf. Table 1).

OCHEM allows users to select a different target unit when creating a consensus model. We also developed individual models directly using the % unit and created a consensus model with % or LogIt units, respectively (cf. Table 3). The consensus model based on individual models developed with the same unit (LogIt-LogIt model) had a significantly lower RMSE compared to the consensus model based on the individual models developed with the % unit (%-LogIt model) for all but the PFC set. For the latter set, RMSEs of both consensus models were not significantly different due to their large confidence intervals. There were no significant differences for all but the PFC99 subset (see discussion below) when we compared performances of LogIt-% and %-% models using the percentage as the final unit.

The results for the PFC set had significantly large errors (RMSE=0.79 and 0.74 in LogIt unit) compared to the training set compounds (RMSE=0.25 and 0.29) for LogIt-LogIt and %-LogIt models, thus indicating that this set was particularly difficult to predict. This was the expected result, since the distribution of the compounds in the PFC set was markedly different to that of the training set, as shown on PCA plot (cf. Fig. 2C). However, there were no significant differences in RMSEs for the training and test sets when predicted values were compared using percentage units. However, for the subset of the PFC set with PPB>99 % (PFC99, $n = 36$), the consensus model LogIt-(LogIt or %) models yielded significantly smaller errors than the %-(LogIt or %) consensus model for

Table 3
RMSE of consensus models developed with LogIt and percentage units.

Unit of individual models	Consensus model unit	Model name	Training set CV, n = 2571	Test set, n = 643	New set n = 25	Test set PFC, n = 63	Test set PFC99, n = 36
LogIt	LogIt	LogIt-LogIt	0.252 ± 0.004*	0.24 ± 0.007*	0.68 ± 0.1*	0.79 ± 0.08	0.49 ± 0.05*
	%	LogIt-%	7 ± 0.2 %	6.6 ± 0.4 %	6 ± 2 %	9 ± 3	0.5 ± 0.1 %*
%	LogIt	%-LogIt	0.293±0.004	0.273±0.008	0.78±0.1	0.74±0.07	0.9 ± 0.1
	%	%-%	7.1 ± 0.2 %	6.4 ± 0.4 %	7.1 ± 2 %	7.5 ± 1	3.6 ± 0.9 %

PFC99 – dataset comprising compounds with PPB>99 %. *indicates significantly lower RMSE ($p < 0.05$) using the respective consensus unit.

both unit types (cf. Table 3). Thus, the LogIt-(LogIt or %) consensus models were much better at predicting the PPB of compounds with very high binding. The ability to differentiate compounds with high PPB values is important for drug discovery, and we have shown that developing models using LogIt units allows for a deeper exploration of the data and more accurate predictions of PPB.

4.5. Physicochemical descriptors highly related to PPB

We calculated the physicochemical descriptors of the molecules, along with their corresponding Person correlation coefficients r^2 with PPB using the Mordred program (Moriwaki et al., 2018). A total of 397 related physicochemical descriptors ($r^2 > 0.05$) were obtained. Among them, 17 descriptors (cf. Supporting Materials Table S3) were highly correlated to PPB, with r^2 values greater than 0.4. The heatmap below (cf. Fig. 5) illustrates the correlations between the selected descriptors and PPB.

As mentioned in the Methods section, we investigated which of the representative descriptors were correlated with high, medium and low PPB values. We observed that the SLogP (predicted lipophilicity) and LogS (predicted solubility) were very important physicochemical indicators for PPB, with R^2 values greater than 0.6. When considering three subgroups of SLogP (SLogP≥3, 1<SLog P < 3, SLog $P \leq 1$) and three subgroups of LogS (LogS<−6, −6≤logS<−4, −4≤logS), we observed that a large proportion (around 80 %) of high-PPB compounds had high SLogP (≥3) and low LogS (<−4), as shown in Fig. 6A and 6B. In medium- or low-PPB compounds, the proportion of compounds with SLogP≥3 and LogS<−4 decreased significantly. Therefore, SLogP was positively correlated with PPB and LogS were negatively correlated with PPB. If The SLogP is greater and the groups are more lipophilic, the hydrophobic effect is stronger. If the LogS is smaller, the groups are less hydrophilic, the bonding between the drug and plasma protein as well as hydrophobic binding is stronger, and thus the plasma protein binding rate is higher.

The number of AromAtom or nAromBond was also correlated with PPB (cf. Fig. 6C.). A reduction in the number of AromAtom or nAromBond led to a significant decrease in the value of PPB – not a surprising observation, as it is generally known that the increase in the number of aromatic rings can improve the lipophilic properties of a drug, leading to greater SLogP value and thus higher PPB. In addition, we found that a higher number of halogen atoms (nX) may be associated with higher PPB values (cf. Fig. 6D), which may be related to an increased propensity for hydrogen bonding, as the introduction of halogens increases electronegativity.

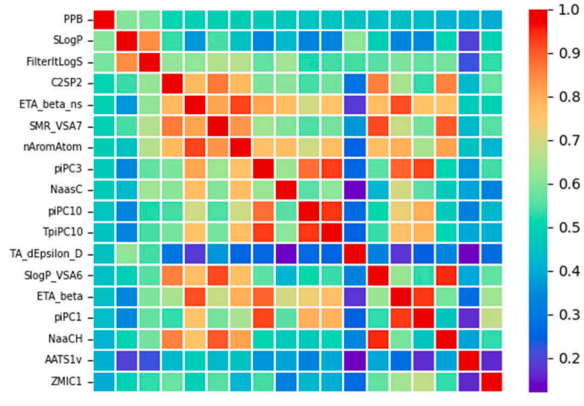

Fig. 5. A heatmap of r^2 between any two descriptors or between any descriptor and PPB. The heatmap was plotted based on the absolute value of r^2.

The number of acid (nAcid) or basic (nBase) groups is closely related to the pKa of compounds. It was found that, for compounds with 0–1 acid groups, nAcid had no effect on PPB. For nAcid ≤ 2, the increase of nAcid may lead to lower PPB values. As for nBase, the effect of basic groups on PPB was more significant than that of acidic groups. In general, more basic groups may lead to lower PPB values (cf. Fig. 6E and F).

Atom-bond Connectivity Index (ABC Index) refers to the strong interaction between two or more adjacent atoms, including covalent, ionic and metallic bonds. The study found that lower ABC Indices were associated with lower PPB values, as shown in Supporting Materials Figure S2A. Hydrogen bonding is known to be a very important intermolecular force, and our results showed that the more hydrogen bond donors a molecule contained (i.e., >2), the more likely it was to have a lower PPB, whereas the number of hydrogen bond acceptors had no significant influence on PPB (cf. Supporting Materials Figure S2B and S2C). In addition to the above mentioned descriptors, we identified other physicochemical descriptors that showed positive correlations with PPB, e.g., KappaShapeIndex, sum of atomic volume parameters (McGowanVolume), atomic polarizability, rotatable bond (nRot), van der Waals volume (VdwVolumeABC), and some topological indicators, such as Wiener index and ZagrebIndex, cf. Supporting Materials Figure S2D-J.

4.6. Substructures that affect PPB

To identify structures that significantly affect PPB, we built a classification model (PPB>90 % and PPB< 50 %) using Morgan2 fingerprints and generated a similarity map for several representative compounds (Riniker and Landrum, 2013). The representative similarity maps of 3 low-PPB and 3 high-PPB compounds are displayed in Fig. 7. For low-PPB compounds (cf. Fig. 7A-C), we can conclude that: (1) amino groups often existed in low-PPB compounds, while secondary and primary amines were dominant. It was exactly consistent with the observation from the descriptor analysis (e.g., more basic groups may lead to lower PPB). Also, the presence of five-membered nitrogen heterocyclic rings, saturated polycyclic rings, carbonyl groups, hydroxyl groups and carboxyl groups appeared to play an important role in the reduction of PPB. (2) For high-PPB compounds (cf. Fig. 7D-F), the presence of aromatic rings, halogen atoms (F, Cl, Br) in a benzene ring or alkyl chain, alkyl chains, sulfonyl groups, thiazoles, oxazoles and oxadiazoles were associated with higher PPB values. More details are shown in Supporting Materials Figure S3 and Figure S4.

We further analyzed the frequency of occurrence of substructures of high/low PPB compounds using SetCompare tool in the OCHEM platform. It was found that several functional groups were strongly associated with one of the two analyzed classes (cf. Fig. 8a). Aromatic and halogen derivatives occurred significantly more often in high-PPB compounds. Additionally, arenes and aryl chlorides derivatives were

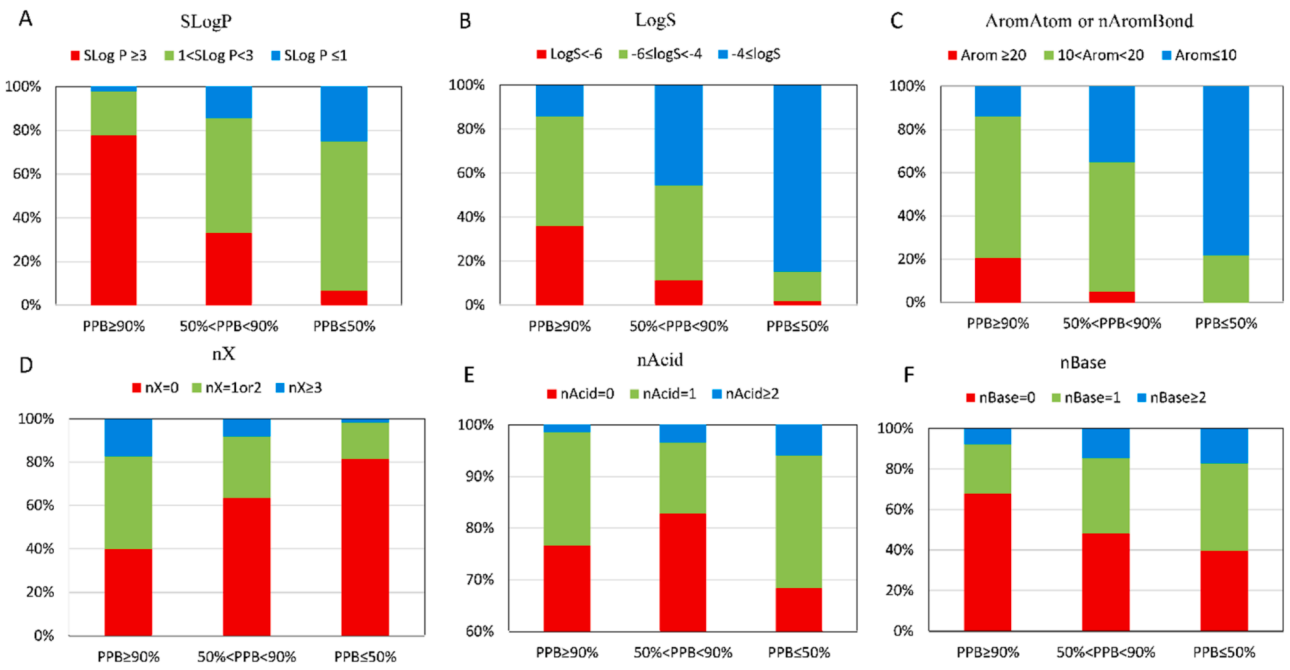

Fig. 6. The differences in physicochemical properties of high-, medium- and low- PPB compounds. (A) SLogP, (B) LogS, (C) AromAtom or nAromBond, (D) nX, (E) nAcid, (F) nBase.

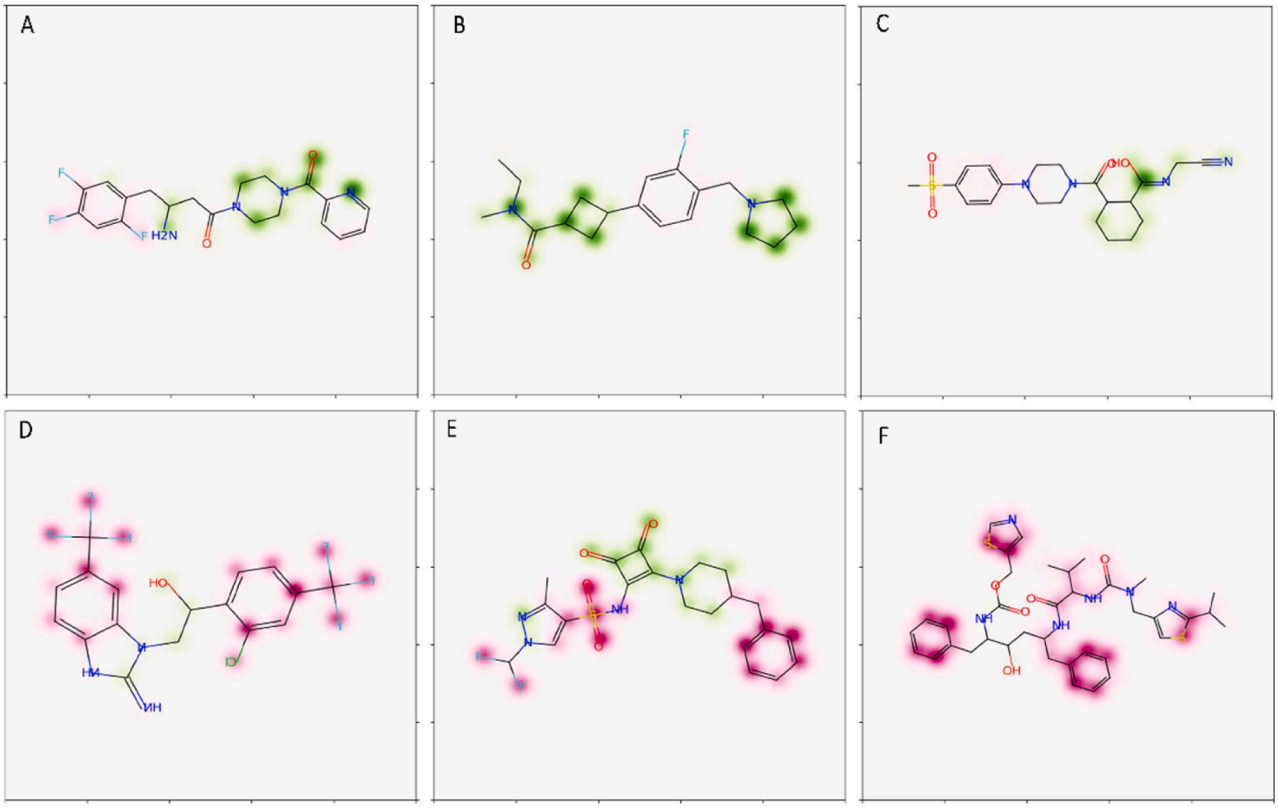

Fig. 7. Similarity maps of representative low-PPB (A-C) and high-PPB compounds (D-F). Atoms colored in red may increase PPB, while atoms colored in green are associated with lower PPB values.

overrepresented in the high PPB dataset. At the same time, primary amines and secondary aliphatic amines occurred more often in the low PPB dataset (cf. Fig. 8b). Other groups associated with low PPB are secondary alcohols, heterocycles, α, β-unsaturated carboxylic acids and tetrahydrofuran. The results from this analysis can provide useful suggestions for chemists seeking to design compounds with low PPB. The full list of calculated groups is shown in Supporting Materials.xls (SetCompare results).

The model was developed and made publicly available at OCHEM platform. OCHEM is a pretty good online service platform. After 15 years

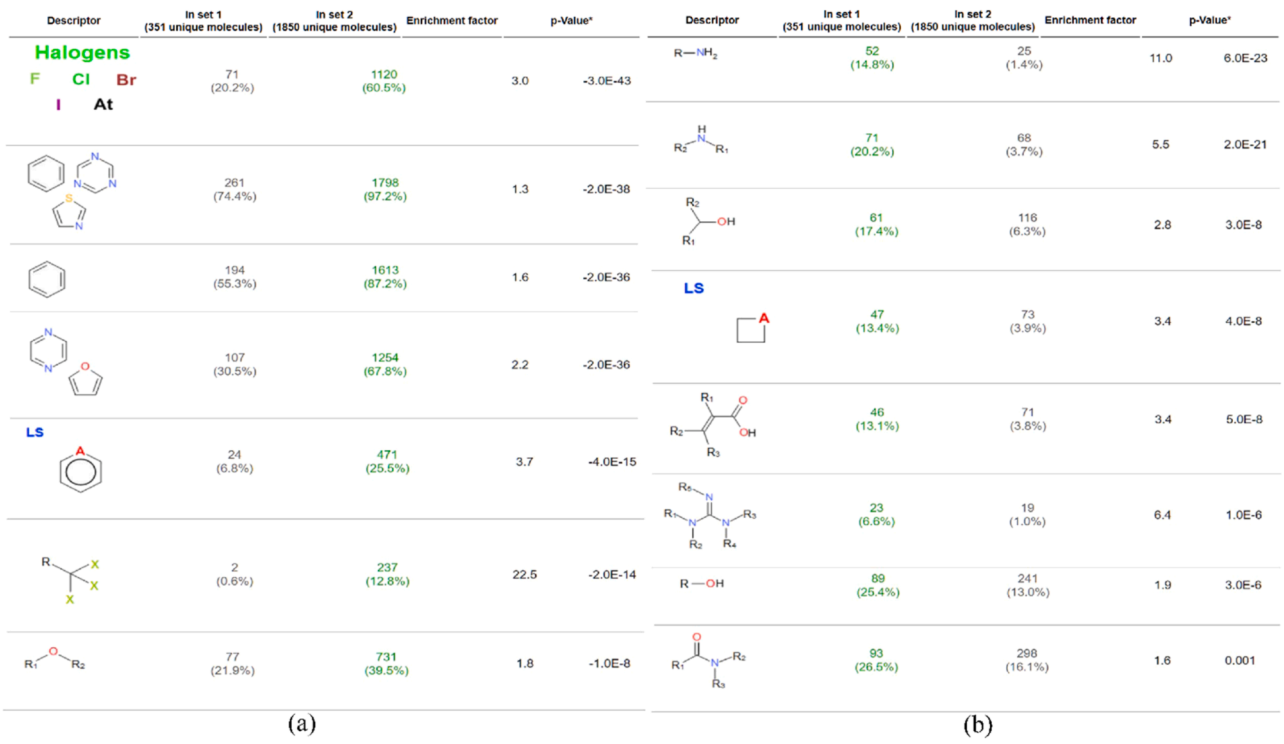

Fig. 8. Functional groups that are overrepresented in high (a) / low (b) PPB compound datasets. Appearance counts are listed, as well as the *p*-value of the respective distribution. Negative and positive p-values indicate groups overrepresented in high/low PPB datasets, respectively.

of development, the platform can be very convenient for data management and processing, model training and prediction, model tuning and optimization, data analysis and a whole chain of other services, which are all free of charge. More importantly, OCHEM is easy to use, especially for chemists who have little computational expertise. However, for some chemists, they may want to understand the details of models. In such a situation, the convenience of OCHEM may not be a merit for them. Additionally, some users may concern about the data privacy. In this regard, we recently made OpenOCHEM https://github.com/openochem publicly available. The users can install it on their computers and develop models locally. In the future, we will add a way to both export and import OCHEM models, allowing the same model to be run online or as standalone version.

5. Conclusions

In the present study, we established a highly accurate PPB consensus model using the OCHEM platform with high-quality datasets. In the retrospective and prospective validation, our model demonstrated higher performances compared to the mainstream prediction platforms. Accurate prediction of a compound's PPB value is of high importance for drug development and clinical use, especially when the PPB of a compound is >99 %, and our model showed excellent predictive performance for such compounds. More importantly, we analyzed the chemical features closely related to PPB, including physicochemical properties and substructure features and functional groups, which are of great significance for structural optimization. Nevertheless, there is still room for improvement. For example, the current prediction accuracy for high-PPB compounds is generally not satisfactory. In addition, though our PPB prediction model demonstrated good performance for most compounds, more experimental measurements are required to further improve model performance, in particular for compounds with high PPB values.

Notes

The authors declare no competing financial interest.

CRediT authorship contribution statement

Zunsheng Han: Writing – review & editing, Writing – original draft, Validation, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zhonghua Xia:** Writing – review & editing, Supervision, Methodology. **Jie Xia:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Igor V. Tetko:** Writing – review & editing, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Song Wu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Acknowledgements

We thank Dr. Katya Ahmad for proof-reading the manuscript for English and her comments.

Funding

This work was supported by Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (No. 2021-I2M-1–069), the National Science and Technology Major Projects for Major New Drugs Innovation and Development (No. 2018ZX09711001–012–003) and the Program for Foreign Talent of Ministry of Science and Technology of the People's Republic of China (No. G2021194015L).

Supplementary materials

The performance of the Published Models; The distribution of original %PPB values; HPLC or LC/MS/MS methods for 25 ChemDiv

compounds; Mordred descriptors determined via feature selection and Pearson correlation coefficient (R^2) between the descriptor and PPB, The differences in molecular properties between high- and low- PPB chemicals; substructural features of low-PPB and high-PPB compounds (Supporting Materials.docx); model output from six web servers for 25 new measurements, PFC test and PFC99 test set; substructure feature analysis by SetCompare (Supporting Materials.xls).

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejps.2024.106946.

Data availability

Data will be made available on request.

References

Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminf. 7, 20. https://doi.org/10.1186/s13321-015-0069-3.

Basant, N., Gupta, S., Singh, K.P., 2016. Predicting binding affinities of diverse pharmaceutical chemicals to human serum plasma proteins using QSPR modelling approaches. SAR QSAR Environ. Res. 27, 67–85. https://doi.org/10.1080/1062936X.2015.1133700.

Di, L., 2021. An update on the importance of plasma protein binding in drug discovery and development. Expert Opin. Drug Discov. 16, 1453–1465. https://doi.org/10.1080/17460441.2021.1961741.

Di, L., Breen, C., Chambers, R., Eckley, S.T., Fricke, R., Ghosh, A., Harradine, P., Kalvass, J.C., Ho, S., Lee, C.A., Marathe, P., Perkins, E.J., Qian, M., Tse, S., Yan, Z.Y., Zamek-Gliszczynski, M.J., 2017. Industry perspective on contemporary protein-binding methodologies: considerations for regulatory drug-drug interaction and related guidelines on highly bound drugs. J. Pharm. Sci. 106, 3442–3452. https://doi.org/10.1016/j.xphs.2017.09.005.

Dimitrijevic, D., Fabian, E., Funk-Weyer, D., Landsiedel, R., 2023. Rapid equilibrium dialysis, ultrafiltration or ultracentrifugation? Evaluation of methods to quantify the unbound fraction of substances in plasma. Biochem. Biophys. Res. Commun. 651, 114–120. https://doi.org/10.1016/j.bbrc.2023.02.021.

Gao, H., Yao, L.L., Mathieu, H.W., Zhang, Y., Maurer, T.S., Troutman, M.D., Scott, D.O., Ruggeri, R.B., Lin, J., 2008. Silico modeling of nonspecific binding to human liver Microsomes. Drug Metab. Dispos. 36, 2130–2135. https://doi.org/10.1124/dmd.107.020131.

Hong, H.X., Xie, Q., Ge, W.G., Qian, F., Fang, H., Shi, L.M., Su, Z.Q., Perkins, R., Tong, W. D., 2008. Mold2, molecular descriptors from 2D structures for Chemoinformatics and Toxicoinformatics. J. Chem. Inf. Model. 48, 1337–1344. https://doi.org/10.1021/ci800038f.

Ingle, B.L., Veber, B.C., Nichols, J.W., Tornero-Velez, R., 2016. Informing the human plasma protein binding of environmental chemicals by machine learning in the pharmaceutical space: applicability domain and limits of predictability. J. Chem. Inf. Model. 56, 2243–2252. https://doi.org/10.1021/acs.jcim.6b00291.

Jimenez-Luna, J., Skalic, M., Weskamp, N., Schneider, G., 2021. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. J. Chem. Inf. Model. 61, 1083–1094. https://doi.org/10.1021/acs.jcim.0c01344.

Karpov, P., Godin, G., Tetko, I.V., 2020. Transformer-CNN: swiss knife for QSAR modeling and interpretation. J. Cheminf. 12, 17. https://doi.org/10.1186/s13321-020-00423-w.

Kawashima, H., Watanabe, R., Esaki, T., Kuroda, M., Nagao, C., Natsume-Kitatani, Y., Ohashi, R., Komura, H., Mizuguchi, K., 2023. DruMAP: a novel drug metabolism and pharmacokinetics analysis platform. J. Med. Chem. 66, 9697–9709. https://doi.org/10.1021/acs.jmedchem.3c00481.

Khaouane, A., Ferhat, S., Hanini, S., 2023. A quantitative structure-activity relationship for human plasma protein binding: prediction, validation and applicability domain. Adv. Pharm. Bull. 13, 784–791. https://doi.org/10.34172/apb.2023.078.

Lambrinidis, G., Vallianatou, T., Tsantili-Kakoulidou, A., 2015. In vitro, in silico and integrated strategies for the estimation of plasma protein binding. Adv. Drug Delivery Rev. 86, 27–45. https://doi.org/10.1016/j.addr.2015.03.011.

Lee, S.K., Lee, I.H., Kim, H.J., Chang, G.S., Chung, J.E., No, K.T., 2003. The PreADME approach: web-based program for rapid prediction of physico-chemical, drug absorption and drug-like properties. EuroQSAR 2002 Designing Drugs and Crop Protectants: processes, problems and solutions 418–420.

Li, F., Shi, S.H., Yi, J.C., Wang, N.N., He, Y.H., Wu, Z.X., Peng, J.F., Deng, Y.C., Wang, W. X., Wu, C.K., Lyu, A.P., Zeng, X.X., Zhao, W.T., Hou, T.J., Cao, D.S., 2024. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res 52, W422–W431. https://doi.org/10.1093/nar/gkae236.

Lin, X.Y., Chi, X., Xiong, Z.P., Zhang, X.F., Ni, N.J., Chang, J.L., Pan, R.Q., Wang, Z.D., Yu, F., Tian, Q., Jiang, H.L., Zheng, M.Y., Qiao, N., 2022. PanGu drug model: learn a molecule Like a human. Sci. China: Life Sci. 66, 879–882. https://doi.org/10.1007/s11427-022-2239-y.

Lou, C.F., Yang, H.B., Wang, J.Y., Huang, M.T., Li, W.H., Liu, G.X., Lee, P.W., Tang, Y., 2022. IDL-PPBopt: a strategy for prediction and optimization of human plasma protein binding of compounds via an interpretable deep learning method. J. Chem. Inf. Model. 62, 2788–2799. https://doi.org/10.1021/acs.jcim.2c00297.

Makarov, D.M., Fadeeva, Y.A., Shmukler, L.E., Tetko, I.V., 2021. Beware of proper validation of models for ionic Liquids! J. Mol. Liq. 344, 117722. https://doi.org/10.1016/j.molliq.2021.117722.

Moriwaki, H., Tian, Y.S., Kawashita, N., Takagi, T., 2018. Mordred: a molecular descriptor calculator. J. Cheminf. 10, 4. https://doi.org/10.1186/s13321-018-0258-y.

Pires, D.E.V., Blundell, T.L., Ascher, D.B., 2015. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J. Med. Chem. 58, 4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104.

Pore, S., Roy, K., 2024. Insights into pharmacokinetic properties for exposure chemicals: predictive modelling of human plasma fraction unbound (fu) and hepatocyte intrinsic clearance (Clint) data using machine learning. Digital Discov. 3, 1852–1877. https://doi.org/10.1039/D4DD00082J.

Reiser, P., Eberhard, A., Friederich, P., 2021. Graph neural networks in TensorFlow-Keras with RaggedTensor representation (kgcnn). Softw. Impacts. 9, 100095. https://doi.org/10.1016/j.simpa.2021.100095.

Riniker, S., Landrum, G.A., 2013. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. J. Cheminf. 5, 43. https://doi.org/10.1186/1758-2946-5-43.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742–754. https://doi.org/10.1021/ci100050t.

Salmina, E.S., Haider, N., Tetko, I.V., 2016. Extended Functional Groups (EFG): an efficient set for chemical characterization and structure-activity relationship studies of chemical compounds. Molecules 21, 1. https://doi.org/10.3390/molecules21010001.

Seyfinejad, B., Ozkan, S.A., Jouyban, A., 2021. Recent advances in the determination of unbound concentration and plasma protein binding of drugs: analytical methods. Talanta 225, 122052. https://doi.org/10.1016/j.talanta.2020.122052.

Smeltz, M., Wambaugh, J.F., Wetmore, B.A., 2023. Plasma protein binding evaluations of Per- and Polyfluoroalkyl substances for category-based Toxicokinetic assessment. Chem. Res. Toxicol. 36, 870–881. https://doi.org/10.1021/acs.chemrestox.3c00003.

Smith, D.A., Di, L., Kerns, E.H., 2010. The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery. Nat. Rev. Drug Discov. 9, 929–939. https://doi.org/10.1038/nrd3287.

Sun, L.X., Yang, H.B., Li, J., Wang, T.D.Y., Li, W.H., Liu, G.X., Tang, Y., 2018. Silico prediction of compounds binding to human plasma proteins by QSAR models. ChemMedChem 13, 572–581. https://doi.org/10.1002/cmdc.201700582.

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A.K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q.Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., Tetko, I.V., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput.-Aided Mol. Des. 25, 533–554. https://doi.org/10.1007/s10822-011-9440-2.

Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., Varnek, A., 2008. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J. Chem. Inf. Model. 48, 1733–1746. https://doi.org/10.1021/ci800151m.

Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N., Villa, A.E.P., 2001. Estimation of aqueous solubility of chemical compounds using E-state indices. J. Chem. Inf. Comput. Sci. 41, 1488–1493. https://doi.org/10.1021/ci000392t.

Tetko, I.V., 2009. Associative Neural Network. In: Livingstone, D.J. (Ed.), Artificial Neural Networks: Methods and Applications. Humana Press, Totowa, NJ, pp. 180–197.

Tetko, I.V., Lowe, D.M., Williams, A.J., 2016. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. J. Cheminf. 8, 2. https://doi.org/10.1186/s13321-016-0113-y.

Vallianatou, T., Lambrinidis, G.M., Tsantili-Kakoulidou, A., 2013. In silico prediction of human serum albumin binding for drug leads. Expert Opin. Drug Discovery 8, 583–595. https://doi.org/10.1517/17460441.2013.777424.

Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I.V., Marcou, G., 2008. ISIDA - platform for virtual screening based on fragment and Pharmacophoric descriptors. Curr. Comput.-Aided Drug Des. 4, 191–198. https://doi.org/10.2174/157340908785747465.

Villa, A.E.P., Tetko, I.V., Dutoit, P., De Ribaupierre, Y., De Ribaupierre, F., 1999. Corticofugal modulation of functional connectivity within the auditory thalamus of rat, guinea pig and cat revealed by cooling deactivation. J. Neurosci. Methods 86, 161–178. https://doi.org/10.1016/S0165-0270(98)00164-2.

Vorberg, S., Tetko, I.V., 2014. Modeling the biodegradability of chemical compounds using the online chemical modeling environment (OCHEM). Mol. Inf. 33, 73–85. https://doi.org/10.1002/minf.201300030.

Votano, J.R., Parham, M., Hall, L.M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., 2006. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure−information representation. J. Med. Chem. 49, 7169–7181. https://doi.org/10.1021/jm051245v.

Vuignier, K., Schappler, J., Veuthey, J.L., Carrupt, P.A., Martel, S., 2010. Drug-protein binding: a critical review of analytical tools. Anal. Bioanal. Chem. 398, 53–66. https://doi.org/10.1007/s00216-010-3737-1.

Wang, N.N., Deng, Z.K., Huang, C., Dong, J., Zhu, M.F., Yao, Z.J., Chen, A.F., Lu, A.P., Mi, Q., Cao, D.S., 2017. ADME properties evaluation in drug discovery: prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling.

Chemom. Intell. Lab. Syst. 170, 84–95. https://doi.org/10.1016/j.chemolab.2017.09.005.

Watanabe, R., Esaki, T., Kawashima, H., Natsume-Kitatani, Y., Nagao, C., Ohashi, R., Mizuguchi, K., 2018. Predicting fraction unbound in human plasma from chemical structure: improved accuracy in the low value ranges. Mol. Biopharm. 15, 5302–5311. https://doi.org/10.1021/acs.molpharmaceut.8b00785.

Weaver, S., Gleeson, M.P., 2008. The importance of the domain of applicability in QSAR modeling. J. Mol. Graphics Modell. 26, 1315–1326. https://doi.org/10.1016/j.jmgm.2008.01.002.

Xiong, Z.P., Wang, D.Y., Liu, X.H., Zhong, F.S., Wan, X.Z., Li, X.T., Li, Z.J., Luo, X.M., Chen, K.X., Jiang, H.L., Zheng, M.Y., 2020. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J. Med. Chem. 63, 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959.

Yang, K., Swanson, K., Jin, W.G., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R., 2019a. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59, 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

Yang, H.B., Lou, C.F., Sun, L.X., Li, J., Cai, Y.C., Wang, Z., Li, W.H., Liu, G.X., Tang, Y., 2019b. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. Bioinformatics 35, 1067–1069. https://doi.org/10.1093/bioinformatics/bty707.

Yuan, Y.W., Chang, S., Zhang, Z., Li, Z.G., Li, S.Z., Xie, P., Yau, W.P., Lin, H.S., Cai, W.M., Zhang, Y.C., Xiang, X.Q., 2020. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemom. Intell. Lab. Syst. 199, 103962. https://doi.org/10.1016/j.chemolab.2020.103962.

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Öberg, T., Dao, P., Cherkasov, A., Tetko, I.V., 2008. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. J. Chem. Inf. Model. 48, 766–784. https://doi.org/10.1021/ci700443v.

Zhu, X.W., Sedykh, A., Zhu, H., Liu, S.S., Tropsha, A., 2013. The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. Pharm. Res. 30, 1790–1798. https://doi.org/10.1007/s11095-013-1023-6.