Development of Novel Methods for QSAR Modeling by Machine Learning Repeatedly: A Case Study on Drug Distribution to Each Tissue

*Koichi Handa[1, *], Saki Yoshimura[1], Michiharu Kageyama[1], and Takeshi Iijima[1]*

[1]Toxicology&DMPK Research Department, Teijin Institute for Bio-medical Research, Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan

KEYWORDS

AI, QSAR, machine learning, tissue-to-plasma partition coefficient, Kp, random forest, message passing neural network, missing values, NA, classification of tissues.

Abstract

AI is expected to help identify excellent candidates in drug discovery. However, we face a lack of data as it is time consuming and expensive to acquire raw data perfectly for many compounds. Hence, we tried to develop a novel QSAR method to predict a parameter more precisely from an incomplete dataset via optimizing data handling by making use of predicted explanatory variables. As a case study we focused on the tissue-to-plasma partition coefficient (Kp), which is an important parameter for understanding drug distribution in tissues and building the physiologically based pharmacokinetic (PBPK) model, is a representative of small and sparse datasets. In this study, we predicted the Kp values of 119 compounds in nine tissues (adipose, brain, gut, heart, kidney, liver, lung, muscle, and skin), while some of these were not available. To fill the missing values in Kp for each tissue, firstly we predicted those Kp values by the non-missing dataset using a random forest (RF) model with *in vitro* parameters (log P, fu, Drug Class, and fi) like a classical prediction by a QSAR model. Next, to predict the tissue-specific Kp values in a test dataset, we constructed a second RF model with not only *in vitro* parameters but also the Kp values of other tissues (*i.e.* other than target tissues) predicted by the first RF model as explanatory variables. Furthermore, we tested all possible combinations of explanatory variables and selected the model with the highest predictability from the test dataset as the final model. The evaluation of Kp prediction accuracy based on the root-mean-square error and R^2-value revealed that the proposed models outperformed other machine learning methods, such as the conventional RF and message-passing neural networks. Significant improvements were observed in the Kp values of adipose tissue, brain, kidney, liver, and skin. These improvements indicated that the Kp information of other tissues can be used to predict the same for a specific tissue. Additionally, we found a novel relationship between each tissue by evaluating all combinations of explanatory variables. In conclusion, we developed a novel RF model to predict Kp values. We hope that this method will be applied to various problems in the field of experimental biology which often contains missing values in the near future.

Introduction

Pharmacokinetics (PK) research is crucial in developing new drugs, from searching for seed compounds to conducting clinical trials[1]. In general, the parameters significantly affecting the blood concentration profile of a drug are volume of distribution (Vd), which quantifies the distribution of the drug inside the human body, and total body clearance (CLtot), which shows the drug processing capacity of the entire body[2]. However, these parameters only reveal the drug's exposure to the blood (or plasma). From the viewpoint of the PK/PD concept, more direct information is often obtained from the unbound concentrations in target tissues[3]. Physiologically based pharmacokinetics (PBPK) modeling has been used to predict drug concentration in target tissues. The U.S. Food and Drug Administration (FDA), along with the European Medicines Agency (EMA), announced and released the availability of their guidance documents for the PBPK model[4,5]. Consequently, when submitting a new drug application (NDA), researchers can use the PBPK model to explain the various PK concerns. Based on survey, over 20 PBPK models of drugs have been submitted to the FDA and/or EMA, 80% of which were accepted by 2018[6]. The PBPK model has various utilities, including predicting drug concentration-time profile and PK parameters, DDI magnitude, effect on special populations (e.g., pediatrics, elderly, and pregnancy), bioequivalence, and food effect[7]. The PBPK model consists of organ compartments with specific volumes and blood flow rates connected by the circulatory system.

Tissue-to-plasma partition coefficient (Kp) is one of the most important parameters in PBPK modeling. It is the ratio of the drug concentration in the tissue to that in plasma, representing the amount of drug transferred to or retained in the tissue. Kp is used in a perfusion-limited model, the limiting process of which is the blood flow to the tissue. Here, Kp divides the drug concentration in a tissue to produce the plasma concentration (also blood concentration using Rb), which returns to the blood system, and differential equations can be used to account for it[8]. The guidelines of drug regulatory authorities suggest distribution studies in animals, such as rats and monkeys (i.e., Kp in animals), as it is difficult to determine the concentration in human tissues[9].

To obtain Kp values for various tissues, the drug concentrations in the plasma and tissues must be determined. However, drug concentration measurement in tissues is much more difficult than that in plasma[10]. In addition to being expensive and time-consuming, this procedure is not feasible during the drug discovery stage. Therefore, various methods have been proposed to predict Kp. The three main methods are as follows: the tissue composition-based algorithm (TCB) (also called as mechanistic model of tissue binding), which is based on drug binding to tissue components[11,12,13,14], the correlation-based algorithm (CBA), which uses a specific tissue to find the Kp values for other tissues based on a regression equation using the Kp value of the muscle[15], and a nonlinear regression model using machine learning algorithm for Kp value prediction. The study by Yun et al. was the first to employ machine learning (ML) to predict Kp values[16]. Recently, our group developed a more accurate multimodal model using the minimum required experimental values and physicochemical descriptors [root-mean-square error (RMSE) and % of two-fold error: 0.39 and 64.5% in 10-fold cross-validation][17]. Despite obtaining an accurate Kp prediction model, we noticed several missing Kp values. This was because none of the tissues handled in the study had Kp values for all 119 compounds (the number of compounds with Kp values in adipose:69; bone:42; brain:89; gut:66; heart:94; kidney:92; liver:85; lung:93; muscle:104; skin:64; and spleen:34). Therefore, we believed that missing data may worsen the accuracy of the QSAR model. Several solutions are known to address the problem of missing values. In the area of drug discovery, accurate prediction of drug repositioning has been made possible by the incorporation of missing-value predictions based on the similarity of compound structures[18]. Missing data related to the activity values of different compound targets were predicted using the random forest (RF) method, and a QSAR model was constructed using the data from these predicted values as explanatory variables[19]. In addition, to predict the drug clearance and volume of distribution in humans, the predicted values in other animals were used as explanatory variables[20].

Based on previous studies, we aimed to build a more accurate model by filling the missing values with the predicted values. More specifically, we investigated the multitask message passing neural network (MPNN) model (Chemprop)[21] and repeated RF model to predict the missing values and then built the RF model again using the predicted values as explanatory variables. Furthermore, to obtain a novel aspect of drug distribution through the analysis of the built models, we investigated the features of Kp values for each tissue as well as the relationship between tissues.

Materials and Methods

Dataset

The utilized dataset was investigated for 119 compounds in a previous report[16,17]. In this dataset in addition to the compound's name, Kp values in each tissue and several properties described in "Properties for explanatory variables" below were included. The amount of data differed for each tissue (adipose, 69; bone,42; brain,89; gut,66; heart,94; kidney,92; liver,85; lung,93; muscle,104; skin,64; and spleen,34). Because the amount of bone and spleen data was significantly small, we decided to omit these tissues.

Properties for Explanatory Variables

The *in vitro* experimental data [LogP, free fraction of plasma (fu), DrugClass (acid, base, weak base, neutral, zwitter), and fraction of ionization (fi)] were used as parameters for the following machine learning models. These values were obtained from our previous research[17], and shown in Table S1.

Machine Learning Algorithm

The workflow of this study is shown in Figure 1A to D. In this study, we investigated several QSAR modeling methods from different viewpoints, including if they were single task or multitask, whether missing values filled or left empty, and whether explanatory features selected or not selected. Overall, the models built in this study were validated by 3-fold cross validation, and the details are transcribed below.

-Dataset Separation for 3-fold Cross Validation

A total of 119 compounds with Kp values in nine tissues, LogP, fu, DrugClass, and fi, were collected from a previous report[17] (Figure 1A) and were separated into 42 compounds with complete Kp data and 77 compounds with missing Kp values (Figure 1B). In fold 1, 14 compounds were randomly selected from the 42 compounds identified in the previous step and used as the test dataset. A total of 105 compounds consisting of 28 filled Kp compounds and 77 compounds with missing Kp values, were used as the training dataset. Folds 2 and 3 included repeating the processes in fold 1 by selecting 14 different test compounds other than those utilized in fold 1 (Figure 1C). The fold setting (training and dataset) remained the same as in the previous step, and the missing Kp values of the training dataset in each fold were filled using the ML model (Figure 1D).

-Conventional RF Model

The RF method is one of the most reliable machine learning algorithms[16,17,22,23], and the modeling process involved in obtaining the missing values is shown in Figure 1A to C. The objective variable for each tissue was Kp, and the explanatory variables were the chemical properties described above. Because this model was built in a single task manner, nine models for Kp prediction corresponding to the nine tissues were built independently. The caret package of R2.4.1 was used for modeling. The default values of the package were used as the initial parameter values, where Ntree was set to 500 and Mtry was set to the root of the number of explanatory variables observed. Parameter tuning was performed according to the RMSE calculation minimization during cross-validation using the caret package. We describe this approach as the conventional RF (CRF) model.

-Multitask Chemprop Model

We investigated the multitask model using MPNN. The modeling process is shown in Figure 1A to C. The objective variable was Kp for each tissue, and the explanatory variable was the graph of each chemical structure, which was canonicalized smiles through Schrödinger Suite 2016. The Python (ver. 3.7.10) Chemprop (version 1.3.1) library Chemprop function was used, and the parameters were set to default values[21]. As this model was built in a multitask manner, one independent model for Kp was built called the multitask Chemprop (MTC) model.

-Repeated RF Model

To fill the missing values of Kp and utilize it as one of the explanatory variables for the QSAR model, we investigated another RF approach (Figure 1A to D). In this model, we first followed the CRF method to fill the blanks in the training dataset with predicted values. Next, another RF model called the Repeated RF (RRF) model was built for tissue-specific Kp value prediction using the chemical properties (logP, fu, Drug Class, and fi) and Kp values other than the objective variables.

-Repeated RF Model with Best Parameters

We examined all possible combinations of explanatory variables and the model with the lowest RMSE value was selected as the RRF model with the best parameters (RRF-BP).

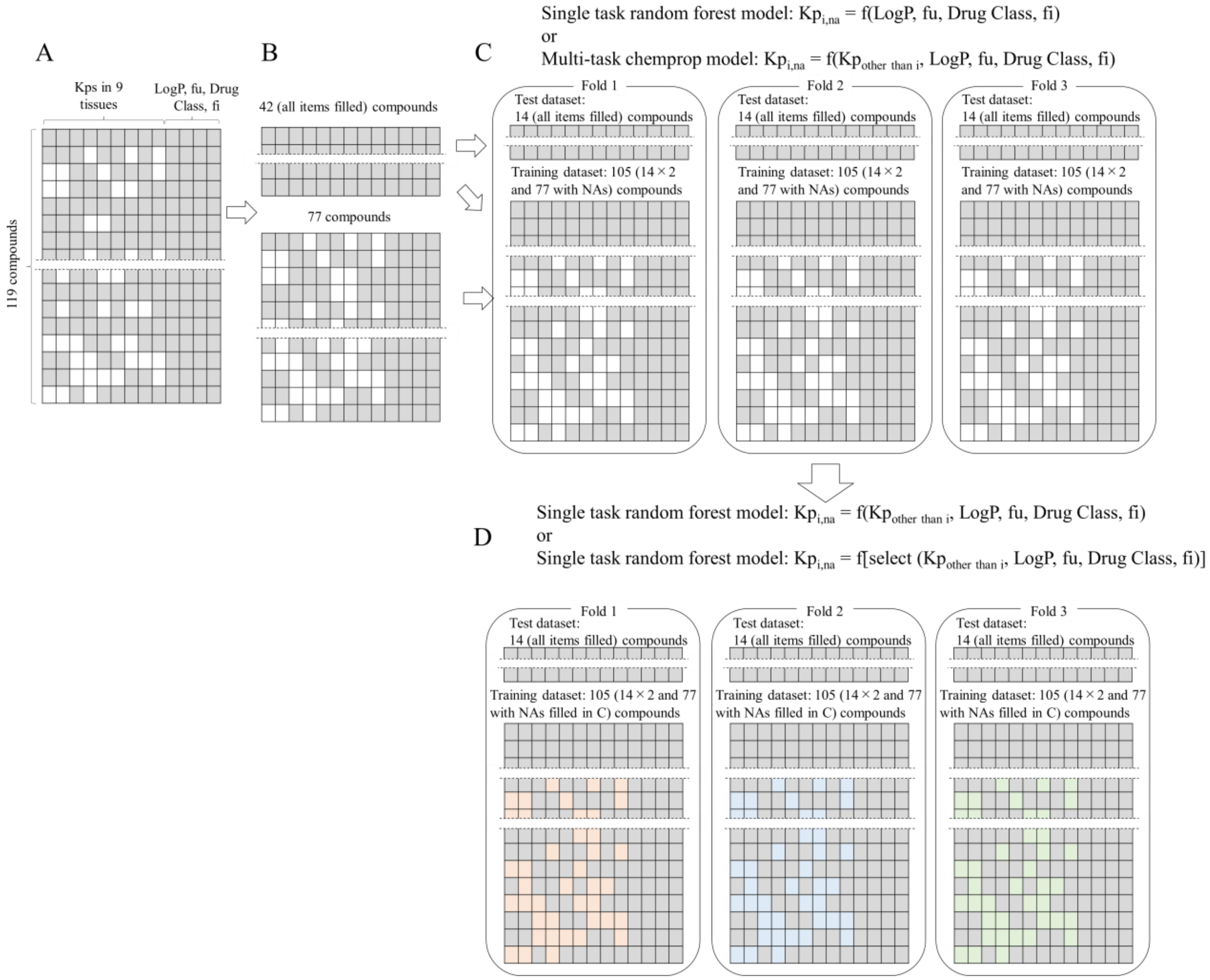
A
Kps in 9 tissues
LogP, fu, Drug Class, fi
119 compounds
B
42 (all items filled) compounds
77 compounds
C
Single task random forest model: $Kp_{i,na}$ = f(LogP, fu, Drug Class, fi)
or
Multi-task chemprop model: $Kp_{i,na}$ = f($Kp_{\text{other than } i}$, LogP, fu, Drug Class, fi)
Fold 1
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs) compounds
Fold 2
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs) compounds
Fold 3
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs) compounds
D
Single task random forest model: $Kp_{i,na}$ = f($Kp_{\text{other than } i}$, LogP, fu, Drug Class, fi)
or
Single task random forest model: $Kp_{i,na}$ = f[select ($Kp_{\text{other than } i}$, LogP, fu, Drug Class, fi)]
Fold 1
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs filled in C) compounds
Fold 2
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs filled in C) compounds
Fold 3
Test dataset: 14 (all items filled) compounds
Training dataset: 105 (14 × 2 and 77 with NAs filled in C) compounds

Figure 1. Workflow for QSAR Modeling.

The dataset (A), dataset separation into training and test set (B), 3-fold setting and Kp NA filled by ML (C), and ML with filled Kp data (D).

Evaluation of Each Model

To evaluate the QSAR model used in this study, we calculated the RMSE (Eq. 1). We compared our method with previously published methods and cited the results in published literature[16]. The methods used for comparison were tissue composition-based models, including Berezhkovskiy[12], Rodgers[13], and Schmitt[14].

$$\mathrm{RMSE}\ =\sqrt{\frac{\sum_{1}^{n}[\log_{10}(a_i)-\log_{10}(b_i)]^2}{n}} \qquad \text{(Eq. 1)}$$

Additionally, R^2-value were calculated for comparison of the proposed model with other models. The R^2-values were calculated using Eq. 2 and used as log10 transformed values.

$$R^2 = 1-\frac{sum\ of\ squares\ of\ residuals}{total\ sum\ of\ squares} \qquad \text{(Eq. 2)}$$

Analysis of The Explanatory Variables

To understand the relationship between each tissue, we performed a cluster analysis of the nine tissues used in this study. Through the RRF-BP modeling process, we recorded the explanatory variable combinations producing the 10 lowest RMSE values. Based on this data, after normalization clustering was performed using the cluster hierarchy function of the SciPy package (ver. 1.7.3) in Python (ver. 3.7.10). A dendrogram was obtained from hierarchical clustering using Euclidian distance and Ward's method.

Results and Discussion

Comparison of Prediction Accuracies Between ML Models Built in This Study

First, to investigate the applicability of our concept in modeling, we compared the ML models built in this study by calculating the RMSE and R^2-values. The results are shown in Table 1. When comparing CRF and MTC, the RMSE and R^2-values of CRF/MTC in most tissues were 0.57/0.58 and 0.38/0.31 in brain, 0.42/0.51 and 0.69/0.52 in gut, 0.37/0.40 and 0.61/0.48 in heart, 0.40/0.42 and 0.64/0.59 in kidney, 0.46/0.53 and 0.71/0.61 in lung, 0.34/0.42 and 0.67/0.48 in muscle, and 0.33/0.37 and 0.37/0.19 in skin, respectively. This superiority of CRF to MTC indicates that, (1) the properties LogP, fu, Drug Class, and fi worked well, and this inference is consistent with several existing studies[11,12,13,16], and (2) multitasking in Chemprop did not work well as it automatically handled missing entries in the dataset by masking out the respective values in the loss function[21,24] and these results indicate that this method of handling the missing values does not work well in these tissues. However, R^2-value of MTC in adipose (0.39) was better than CRF (0.32), and in the liver both RMSE and R^2-value of MTC (0.48 and 0.47) were better than CRF (0.52 and 0.38). This is in agreement with the specific tissue features, because adipose tissue has different components than the others[14] and liver has different types of transporters[25]; hence, the properties used as explanatory variables might not be effective in these tissues.

When comparing CRF with RRF, no significant difference was found in RMSE and R^2-values, except for the liver. It was the only tissue with RMSE difference over 0.05 [The R^2-value of liver in RRF (0.49) was higher than CRF (0.38)]. Similar results for CRF and RRF in most tissues indicates that using all the predicted Kp values is not always effective as this might lead to the concept of each tissue being similar to a specific tissue and not all tissues.

Next, we compared the CRF and RRF-BP to further investigate the effectiveness of the RRF concept. The RMSE and R^2-values of CRF/RRF-BP were 0.55/0.49 and 0.32/0.47 in adipose, 0.57/0.50 and 0.38/0.52 in brain, 0.42/0.37 and 0.69/0.76 in gut, 0.37/0.34 and 0.61/0.64 in heart, 0.40/0.34 and 0.64/0.75 in kidney, 0.52/0.46 and 0.38/0.52 in liver, 0.46/0.42 and 0.71/0.75 in lung, 0.34/0.34 and 0.67/0.70 in muscle, 0.33/0.29 and 0.37/0.51 in skin, respectively. This proves the superiority of RRF-BP over CRF or other methods. The similar performance of RRF and CRF shows the importance of selecting the best parameters, and it could be derived from the features of each tissue, which are similar to some other specific tissues. The RRF-BP addresses the issue of overfitting by selecting the best parameters to reduce the number of explanatory variables. However, it is important to note that we were unable to create a hold-out test dataset due to the lack of Kp data in all tissues.[26]. Consequently, it should be noted that we are unable to demonstrate the generalizability of the RRF-BP method. To achieve this, we would need to obtain a larger experimental dataset.

Table 1. Evaluation of QSAR Models Using RMSE and R^2-value

Method	Adipose		Brain		Gut		Heart		Kidney	
	RMSE	R^2-value	RMSE	R^2-value	RMSE	R^2-value	RMSE	R^2-value	RMSE	R^2-value
CRF	0.55	0.32	0.57	0.38	0.42	0.69	0.37	0.61	0.40	0.64
MTC	0.55	0.39	0.58	0.31	0.51	0.52	0.40	0.48	0.42	0.59
RRF	0.55	0.34	0.58	0.34	0.41	0.71	0.37	0.58	0.36	0.73
RRF-BP	**0.49**	**0.47**	**0.50**	**0.52**	**0.37**	**0.76**	**0.34**	**0.64**	**0.34**	**0.75**

Method	Liver		Lung		Muscle		Skin	
	RMSE	R^2-value	RMSE	R^2-value	RMSE	R^2-value	RMSE	R^2-value
CRF	0.52	0.38	0.46	0.71	**0.34**	0.67	0.33	0.37
MTC	0.48	0.47	0.53	0.61	0.42	0.48	0.37	0.19
RRF	0.47	0.49	0.48	0.69	**0.34**	0.69	0.35	0.31
RRF-BP	**0.46**	**0.52**	**0.42**	**0.75**	**0.34**	**0.70**	**0.29**	**0.51**

CRF: Conventional Random Forest model, MTC: multitask Chemprop model, RRF: Repeated RF model, RRF-BP: RRF model with the best parameters.

Comparison of Prediction Accuracies of Methods Developed for Each Tissue with The Published Models Using RMSE

Next, we analyzed the RMSE of models developed for each tissue in this study. To compare the accuracy with the already published methods, we cited the RMSE information of the tissue composition-based models[16]. The results of this analysis are shown in Figure 2. It can be seen that when comparing the best methods of tissue-composition-based methods and ML models, the lowest RMSE values were found in ML; RMSE values were 0.61 (Rodger) / 0.50 (RRF-BP) in brain, 0.39 (Rodger) / 0.37 (RRF-BP) in gut, 0.36 (Rodger) / 0.34 (RRF-BP) in heart, 0.54 (Schmitt) / 0.34 (RRF-BP) in kidney, 0.57 (Schmitt) / 0.46 (RRF-BP) in liver, 0.50 (Rodger) / 0.42 (RRF-BP) in lung, 0.37 (Rodger) / 0.34 (RRF-BP) in muscle, and 0.41 (Rodger) / 0.29 (RRF-BP) in skin. Although the difference in RMSEs for a few tissues (gut, heart, and muscle) was less than 0.05, it was clear that the RMSE values of the ML model (RRF-BP) were better than those of the tissue composition-based methods. However, for adipose tissue, Rogers (RMSE:0.47)[13] produced the best RMSE value and the best ML model was RRF-BP with an RMSE of 0.49. This indicates that adipose tissue has different features, which is in agreement with the specific tissue components; namely, this tissue has very little water content and consists mostly of lipid components[14]. Hence we can conclude that RRF-BP outperformed others in predicting Kp values for most tissues. However, considering the algorithm of RRF-BP is ML that is totally different form the tissue composition-based models, we should be care of applicability domain when predicting new compounds[27].

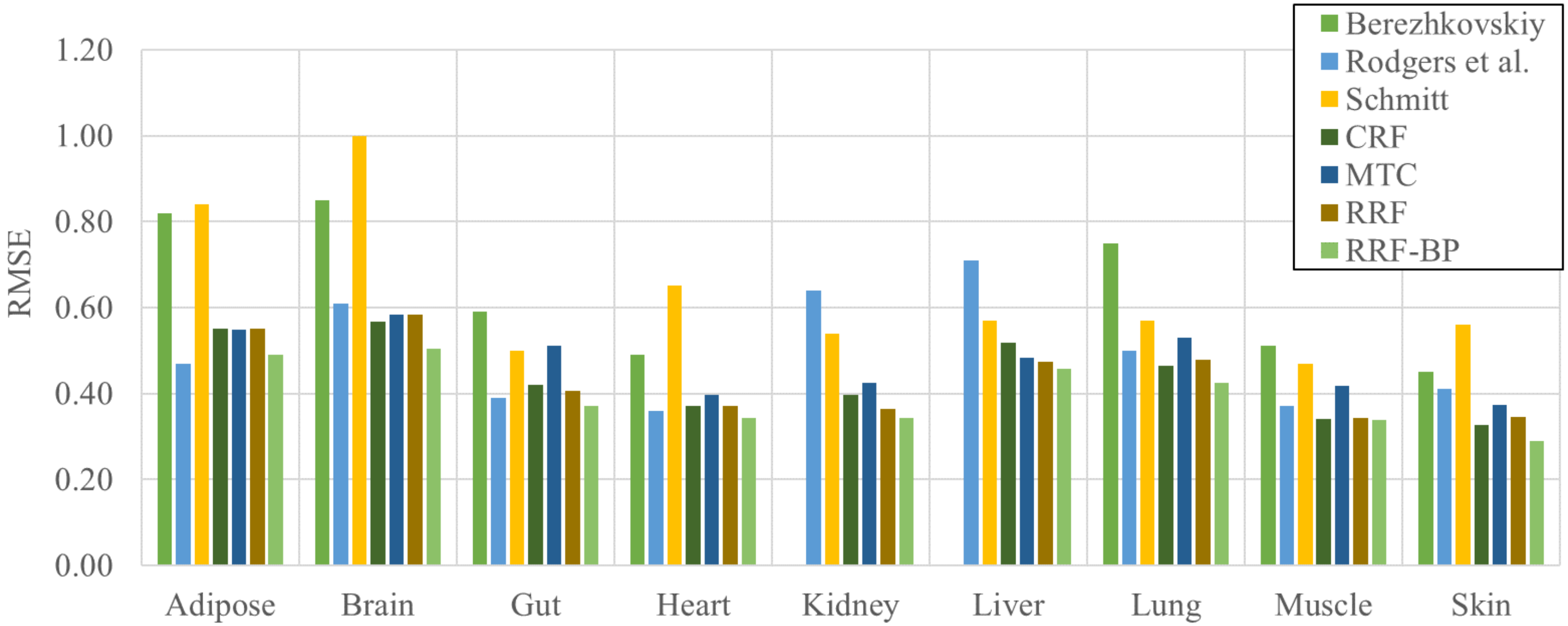

Figure 2. RMSE Analysis of Each Model.

Each bar represents RMSE. In the most of tissues, lower RMSEs were observed in machine learning methods investigated in this study than tissue-composition-based methods.

Predictive Feature of Repeated RF model

As the RRF-BP model accurately predicted the Kp values in all tissues, we analyzed predictivity in detail by comparing the observed and predicted values of RRF-BP in all tissues. The results are shown in Figure 3, and it can be seen that the observed and predicted values correlated well in most tissues. However, for adipose and brain tissue, the predicted values were observed in a horizontal manner. Outliers were observed in lower and higher Kp regions of the gut and lungs, respectively. In the liver, outliers were observed in both the low- and high-Kp regions.

Next, for a more quantitative investigation, we counted the compounds with 2- and 3-fold errors. The results are shown in Table 2 [for comparison, the results of the other models are shown in supporting information, CRF in Table S2, MTC in Table S3, and RRF in Table S4]. The % of within 2-fold errors in the muscle and skin was over 75%. With regard to 3-fold errors, those in the skin were over 95%, and those in the adipose, gut, heart, and muscle were over 80%. On the other hand, % of within 2-fold errors in the brain and lungs was below 60%, and the % of within 3-fold errors was still below 70%. Hence, in most tissues, the compounds were within 2- and 3-fold errors, but only for the brain and lungs, which was not the case.

Furthermore, we investigated the tendency of over- and under-prediction of RRF-BP in each tissue type. The results are also shown in Table 2 where it can be seen that values for the adipose, brain, and gut tissues were overpredicted. By contrast, values for the heart, kidney, and liver were under-predicted. Hence, we can conclude that the predictive model for each tissue has good accuracy and prediction tendency, which must be understood before using it in the practical drug discovery process.

Table 2. Fold Error Analysis of Repeated RF Model with Best Parameters

Tissue	% of within 2-fold error	% of within 3-fold error	% of over-estimated more than 2-fold	% of under-estimated less than 2-fold	% of over-estimated more than 3-fold	% of under-estimated less than 3-fold
Adipose	66.7	81.0	21.4	11.9	11.9	7.1
Brain	54.8	66.7	28.6	16.7	26.2	7.1
Gut	61.9	81.0	23.8	14.3	11.9	7.1
Heart	66.7	81.0	11.9	21.4	7.1	11.9
Kidney	66.7	78.6	14.3	19.0	9.5	11.9
Liver	61.9	76.2	16.7	21.4	9.5	14.3
Lung	57.1	69.0	23.8	19.0	14.3	16.7
Muscle	78.6	83.3	9.5	11.9	9.5	7.1
Skin	76.2	95.2	7.1	16.7	4.8	0.0

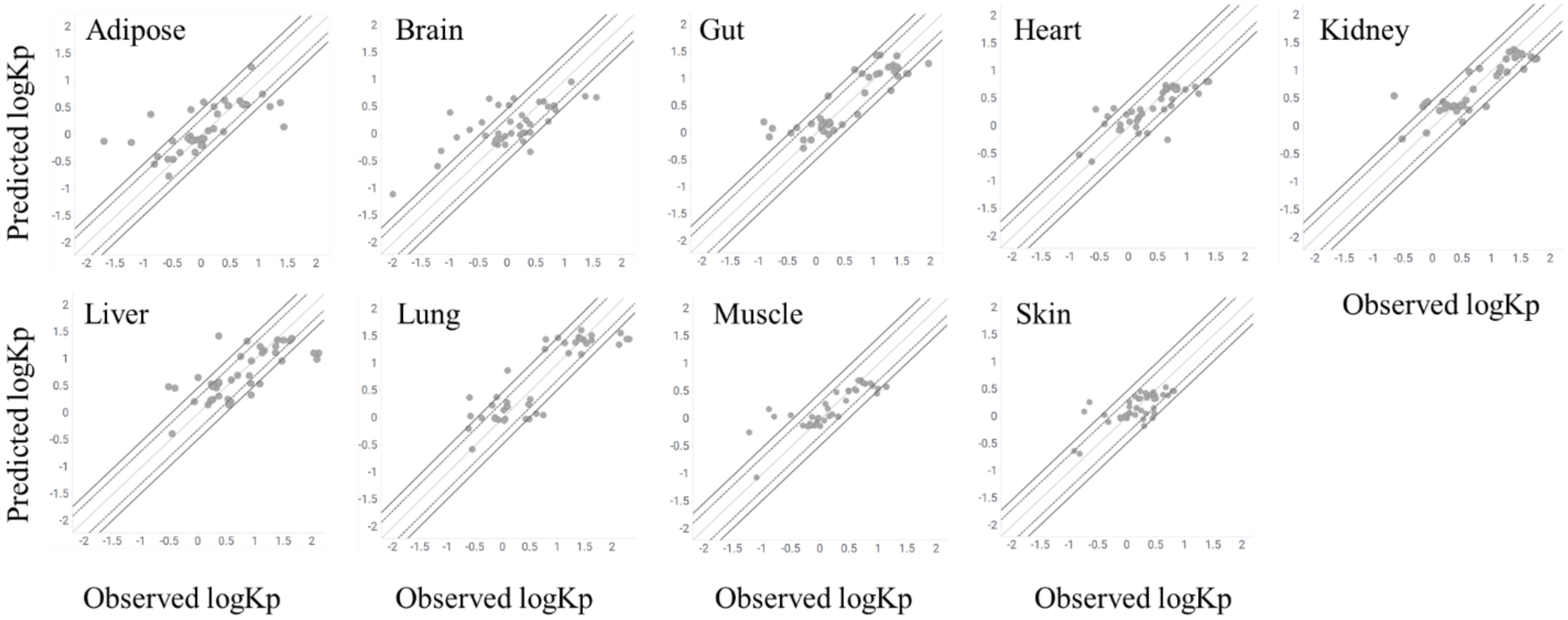

Figure 3. Predictivity of The RRF-BP Model in Each Tissue. Each Grey Circle Represents An Individual Compound.

The x- and y-axes show the observed and predicted log Kp values, respectively. The center diagonal thin line, dotted line, and bold line on each side represent the unity, 2-fold error, and 3-fold error, respectively.

Best Parameters Selected in Repeated RF Model

Next, we investigated the parameters that are important for Kp prediction in each tissue. The best parameters are marked in Table 3. However, regarding some tissues that highly express transporters (brain[23,24,28], kidney[29], and liver[25]), the discussion is very difficult because Kpuu should be investigated for it[30,31]. Consequently, we have the discussion of adipose, gut, muscle and skin. The key findings from these tissues are as follows.

-Adipose

Only chemical properties (LogP, fi, and DrugClass) were selected for the RRF-BP model. This was the only tissue that did not require the Kp values of other tissues to predict its Kp value (Table 3). This could be in agreement with the fact that adipose tissue is completely different from other tissues in terms of composition as mentioned already[14]. Additionally, the properties selected could include lipophilicity, which is directly linked to the drug's affinity to adipose tissue[32]. Hence, we can conclude that the parameters selected for adipose Kp are interpretable and reliable. However, because this model had a higher RMSE (0.49) than the RRF-BP models of other tissues, the Kp prediction for adipose tissue remained difficult (Table 1).

-Gut

LogP, fi, and DrugClass were selected as chemical properties, and the Kp predicted for the adipose tissue and kidney was used for RRF-BP of the gut. The gut acts as a barrier to drug absorption. In this process, permeation is very important and is strongly linked to lipophilicity and ionization[33]. As lipophilicity is also directly linked to the affinity of the drug to adipose[32], the Kp for adipose tissue is an appropriate parameter for the RRF-BP model in the gut. Therefore, we can conclude that, although it is not easy to interpret the function of kidney Kp in the RRF-BP model of the gut, the chemical properties selected in the gut were understandable.

-Muscle and Skin

The muscle RRF-BP model used eight explanatory variables (Kp-predicted values in adipose, brain, gut, skin, LogP, fi, DrugClass, and fu), which was the highest among the RRF-BP models. This indicates one of the reasons of using muscle Kp as a representative Kp in the other tissues. Additionally, tissues that highly express transporters, such as the kidney and liver, were not selected, which might indicate that the muscle represents the extent of drug diffusion transfer to tissues[15]. The skin RRF-BP model used seven explanatory variables (Kp predicted values in adipose, brain, gut, heart, fi, DrugClass, and fu), which was the second highest among the RRF-BP models. However, the RRF-BP model of the skin did not use LogP as compared to the muscle RRF-BP model. Hence it is truly one of the representative tissues as average Kp of tissues[17]. However, this feature may be different from that of the muscles.

Table 3. Explanatory Variables in RRF-BP Model

Objective tissue	Explanatory Variables in BS-TRF												
	Adipose	Brain	Gut	Heart	Kidney	Liver	Lung	Muscle	Skin	LogP	fi	Drug Class	fu
Adipose	-	□	□	□	□	□	□	□	□	☑	☑	☑	□
Brain	□	-	☑	□	□	□	□	□	☑	□	□	□	□
Gut	☑	□	-	□	☑	□	□	□	□	☑	☑	☑	□
Heart	☑	□	□	-	□	□	□	☑	☑	☑	□	□	☑
Kidney	□	□	☑	☑	-	□	☑	☑	□	□	☑	□	□
Liver	□	☑	☑	☑	□	-	□	□	□	□	☑	☑	□
Lung	□	□	☑	□	□	□	-	□	☑	□	☑	☑	□
Muscle	☑	☑	☑	□	□	□	□	-	☑	☑	☑	☑	☑
Skin	☑	☑	☑	☑	□	□	□	□	-	□	☑	☑	☑

Clustering of The Tissues by The Parameters Needed for Predicting Kp

Next, to understand the relationship between each tissue, we analyzed the parameters used in the top 10 RRF-BP models for each tissue with a lower RMSE. We present the sum of the total instances for each item listed in Table S5. Most of the explanatory variables were selected for different Kp prediction models except the predicted Kp value of the liver. For example, in the liver, in addition to the highly expressed transporters there is another feature that does not follow the drug-free theory[34]. This indicates that its distribution into the liver has a completely different mechanism from that in the other tissues.

Using the results of the number of explanatory variables used in the RRF-BP modeling process, we built a dendrogram for each tissue. The result is shown in Figure 4 where it can be seen that four clusters was obtained, called as group I to IV. The tissues were grouped as follows: kidney, brain, and heart in group I; lung and muscle in group II; adipose tissue in group III; and gut, liver, and skin in group IV. Adipose tissue with completely different components from other tissues was placed in group III. The tissues that highly expressed the transporters were located in groups I (kidney and brain) and IV (gut and liver). Hence, considering the features of each tissue, this grouping reflected the tendency of Kp. By contrast, Yau et. al. clustered tissues into four groups, called A to D for convenience, based on tissue components as follows: adipose in group A; brain and muscle in group B; skin in group C; and kidney, liver, lung, gut, and heart in group D[35]. Although adipose was separated as in our study, the grouping of the other tissues was totally different from ours. Certainly, clustering by tissue components is considered very important, especially for passive diffusion; however, it does not always reflect the tendency of the Kp values of drugs. Through this analysis, we can also see that when only a limited tissue can be obtained, we can suggest a tissue within the same group that could become representative of the Kp value other tissues.

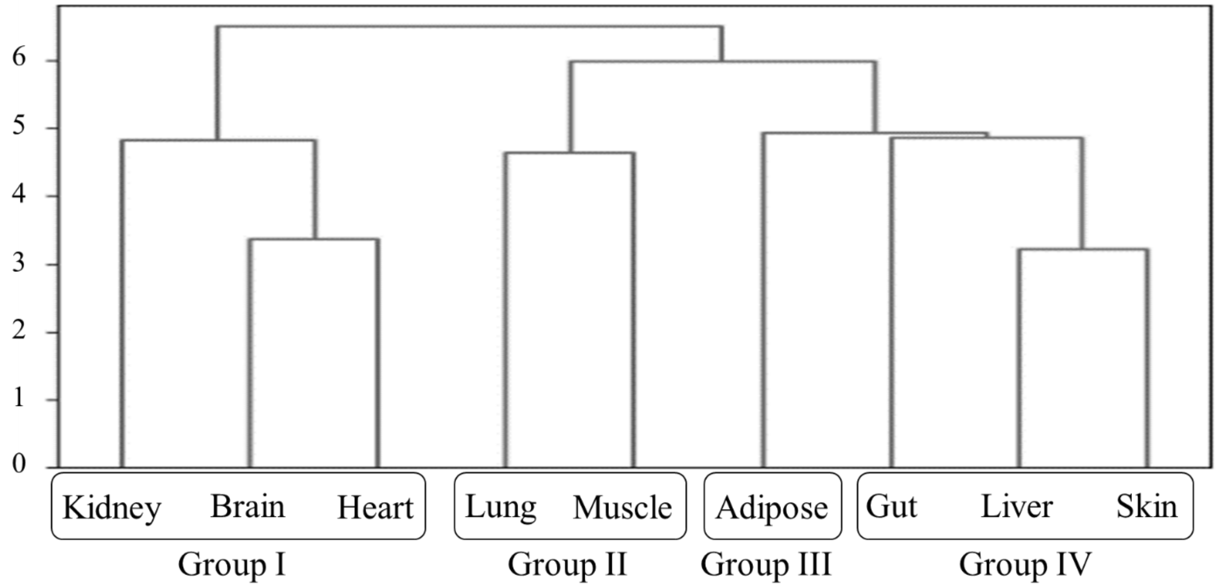

Figure 4. Clustering Analysis of Tissues.

The Y axis represents the height of the dendrogram. Each tissue was classified into four groups from I to IV.

Conclusions

In this study, we investigated a novel QSAR method to predict a parameter more precisely from an incomplete dataset by optimizing the data handling. We focused on the prediction of Kp values, which consisted of the Kp values of 119 compounds in nine tissues (adipose, brain, gut, heart, kidney, liver, lung, muscle and skin), some of which were not available (NA). First, to fill in the missing Kp values for each tissue, we predicted those in the entire dataset using a RF model with *in vitro* parameters (log P, fu, DrugClass, and fi). Second, to predict the Kp value for a certain tissue in a test dataset, we constructed a second RF model with not only *in vitro* parameters but also the Kp values for other tissues predicted by the first RF model as explanatory variables. The prediction accuracies of the Kp values of the final models were higher than those predicted by the other ML methods, and we also observe the usage of the Kp information of the other tissues in predicting the Kp value for a specific tissue. Additionally, through the evaluation of all combinations of explanatory variables for the RRF-BP model, we found that Kp values of no other tissues were needed for the prediction of adipose Kp, whereas liver Kp was not needed for the prediction of Kp for other tissues. Hence, we developed a novel model for predicting of Kp values using a RF model twice, and we hope that this method can be applied to not only the Kp prediction problem but also various other problems.

AUTHOR INFORMATION

Corresponding Author

* Koichi Handa

Present Addresses

†Toxicology & DMPK Research Department, Teijin Institute for Bio-medical Research, Teijin Pharma Limited, 4-3-2 Asahigaoka, Hino-shi, Tokyo 191-8512, Japan

Tel: +81- 42-586-8279

Fax: +81- 42-587-5518

E-mail: koichi.handa@axcelead-twp.com

E-mail: koichi.handa.0722@gmail.com

ORCID: 0000-0003-2748-9742

Author Contributions

Koichi Handa and Saki Yoshimura wrote the manuscript; Michiharu Kageyama made the figures and tables; and Takeshi Iijima reviewed the manuscript.

Funding Sources

The authors (s) report no funding associated with the work featured in this article.

Data and Software Availability statement

The data including chemical structure used in this study are provided in the Supporting Information file. The software used in this study is freely available; R2.4.1 (https://cran.ism.ac.jp/), and Python 3.7.10 (https://www.python.org/downloads/release/python-3710/). The main codes of Python and R used in this study are also provided in the Supporting Information file (Table S6).

ACKNOWLEDGMENT

We wish to thank the researchers at the Teijin Institute for Biomedical Research, such as Seishiro Sakamoto and Shin Umeda, who encouraged us to conduct the study.

REFERENCES

(1) Ballard, P.; Brassil, P.; Bui, K. H.; Dolgos, H.; Petersson, C.; Tunek, A.; Webborn, P. J. The Right Compound in the Right Assay at the Right Time: an Integrated Discovery DMPK Strategy. *Drug Metab. Rev.* **2012**, 44, 224–252.

(2) Sara, E. R. *Basic Pharmacokinetics and Pharamacodynamics*, Wiley: NJ, 2011.

3 Tuntland, T.; Ethell, B.; Kosaka, T.; Blasco, F.; Zang, R. X.; Jain, M., et al. Implementation of Pharmacokinetic and Pharmacodynamic Strategies in Early Research Phases of Drug Discovery and Development at Novartis Institute of Biomedical Research. *Front. Pharmacol.* **2014**, 5, 174.

(4) United States Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). *Physiologically Based Pharmacokinetic Analyses—Format and Content Guidance for Industry*, 2019.

(5) European Medicines Agency. *Reporting of Physiologically Based Pharmacokinetic (PBPK) Modelling and Simulation*, 2018.

(6) Shebley, M.; Sandhu, P.; Emami Riedmaier, A.; Jamei, M.; Narayanan, R.; Patel, A., et al. Physiologically Based Pharmacokinetic Model Qualification and Reporting Procedures for Regulatory Submissions: A Consortium Perspective. *Clin. Pharmacol. Ther.* **2018**, 104, 88–110.

(7) Perry, C.; Davis, G.; Conner, T. M.; Zhang, T. Utilization of Physiologically Based Pharmacokinetic Modeling in Clinical Pharmacology and Therapeutics: an Overview. *Curr. Pharmacol. Rep.* **2020**, 6, 71–84.

(8) Jones, H.; Rowland-Yeo, K. Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development. *CPT Pharmacometrics Syst. Pharmacol.* **2013**, 2, e63.

(9) PMDA. *Guidelines for Nonclinical Pharmacokinetic Studies*. Vol. 469, 1998.

(10) Ho, S. Challenges of Atypical Matrix Effects in Tissue. *Bioanalysis*. **2013**, 5, 2333–2335.

(11) Poulin, P.; Krishnan, K. A Biologically Based Algorithm for Predicting Human Tissue: Blood Partition Coefficients of Organic Chemicals. *Hum. Exp. Toxicol.* **1995**, 14, 273–280.

(12) Berezhkovskiy, L. M. Volume of Distribution at Steady State for a Linear Pharmacokinetic System with Peripheral Elimination. *J. Pharm. Sci.* **2004**, 93, 1628–1640.

(13) Rodgers, T.; Rowland, M. Physiologically Based Pharmacokinetic Modelling 2: Predicting the Tissue Distribution of Acids, Very Weak Bases, Neutrals and Zwitterions. *J. Pharm. Sci.* **2006**, 95, 1238–1257.

(14) Schmitt, W. General Approach for the Calculation of Tissue to Plasma Partition Coefficients. *Toxicol. In Vitro*. **2008**, 22, 457–467.

(15) Björkman, S. Reduction and Lumping of Physiologically Based Pharmacokinetic Models: Prediction of the Disposition of Fentanyl and Pethidine in Humans by Successively Simplified Models. *J. Pharmacokinet. Pharmacodyn.* **2003**, 30, 285–307.

(16) Yun, Y. E.; Cotton, C. A.; Edginton, A. N. Development of a Decision Tree to Classify the Most Accurate Tissue-Specific Tissue to Plasma Partition Coefficient Algorithm for a Given Compound. *J. Pharmacokinet. Pharmacodyn.* **2014**, 41, 1–14.

(17) Handa, K.; Sakamoto, S.; Kageyama, M.; Iijima, T. Development of a 2D-QSAR Model for Tissue-to-Plasma Partition Coefficient Value with High Accuracy Using Machine Learning Method, Minimum Required Experimental Values, and Physicochemical Descriptors. *Eur. J. Drug Metab. Pharmacokinet.* **2023 Jun 2**, 48, 341–352. doi: 10.1007/s13318-023-00832-w. Epub ahead of print. PMID: 37266860.

(18) Sawada, R.; Iwata, H.; Mizutani, S.; Yamanishi, Y.; Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data. *J. Chem. Inf. Model.* **2015**, 55, 2717–2730. DOI: 10.1021/acs.jcim.5b00330.

(19) Martin, E. J.; Polyakov, V. R.; Zhu, X. W.; Tian, L.; Mukherjee, P.; Liu, X. All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, 59, 4450–4459. doi: 10.1021/acs.jcim.9b00375.

(20) Iwata, H.; Matsuo, T.; Mamada, H.; Motomura, T.; Matsushita, M.; Fujiwara, T. et al. Predicting Total Drug Clearance and Volumes of Distribution Using the Machine Learning-Mediated Multimodal Method through the Imputation of Various Nonclinical Data. *J. Chem. Inf. Model.* **2022**, 62, 4057–4065. doi: 10.1021/acs.jcim.2c00318.

(21) https://github.com/chemprop/chemprop

(22) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today*. **2018**, 23, 1538–1546.

(23) Tietz, S.; Engelhardt, B. Brain Barriers: Crosstalk between Complex Tight Junctions and Adherens Junctions. *J. Cell Biol.* **2015**, 209, 493–506.

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H. et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019 Aug 26**, 59, 3370–3388. doi: 10.1021/acs.jcim.9b00237. (Epub 2019 Aug 13). Erratum in: *J. Chem. Inf. Model.* **2019 Dec 23**, 59, 5304–5305. PMID: 31814400; PMCID: PMC6727618.

(25) Mikkaichi, T.; Nakai, D.; Yoshigae, Y.; Imaoka, T.; Okudaira, N.; Izumi, T. Liver-Selective Distribution in Rats Supports the Importance of Active Uptake into the Liver via Organic Anion Transporting Polypeptides (OATPs) in Humans. *Drug Metab. Pharmacokinet.* **2015**, 30, 334–340.

(26) Khan, P. M.; Roy, K. Current Approaches for Choosing Feature Selection and Learning Algorithms in Quantitative Structure-Activity Relationships (QSAR). *Expert Opin. Drug Discov.* **2018**, *13* (12), 1075–1089. https://doi.org/10.1080/17460441.2018.1542428.

(27) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13* (34), 3494–3504. https://doi.org/10.2174/138161207782794257.

(28) Pardridge, W. M. Blood-Brain Barrier Biology and Methodology. *J. Neurovirol.* **1999**, 5, 556–569.

(29) César-Razquin, A.; Snijder, B.; Frappier-Brinton, T.; Isserlin, R.; Gyimesi, G.; Bai, X., et al. A Call for Systematic Research on Solute Carriers. *Cell*. **2015**, 162, 478–487.

(30) Di, L.; Riccardi, K.; Tess, D. Evolving Approaches on Measurements and Applications of Intracellular Free Drug Concentration and Kpuu in Drug Discovery. *Expert Opin. Drug Metab. Toxicol.* **2021**, *17* (7), 733–746. https://doi.org/10.1080/17425255.2021.1935866.

(31) Orozco, C. C.; Atkinson, K.; Ryu, S.; Chang, G.; Keefer, C.; Lin, J.; Riccardi, K.; Mongillo, R. K.; Tess, D.; Filipski, K. J.; Kalgutkar, A. S.; Litchfield, J.; Scott, D.; Di, L. Structural Attributes

Influencing Unbound Tissue Distribution. *Eur. J. Med. Chem.* **2020**, *185*, 111813. https://doi.org/10.1016/j.ejmech.2019.111813.

(32) Bruno, C. D.; Harmatz, J. S.; Duan, S. X.; Zhang, Q.; Chow, C. R.; Greenblatt, D. J. Effect of Lipophilicity on Drug Distribution and Elimination: Influence of Obesity. *Br. J. Clin. Pharmacol.* **2021 Aug**, 87, 3197–3205. doi: 10.1111/bcp.14735. (Epub 2021 Feb 16). PMID: 33450083.

(33) Kokate, A.; Li, X.; Jasti, B. Effect of Drug Lipophilicity and Ionization on Permeability across the Buccal Mucosa: a Technical Note. *AAPS PharmSciTech*. **2008**, 9, 501–504. doi: 10.1208/s12249-008-9071-7. (Epub 2008 Mar 20). PMID: 18431653; PMCID: PMC2976956.

(34) Kim, S. J.; Lee, K. R.; Miyauchi, S.; Sugiyama, Y. Extrapolation of In Vivo Hepatic Clearance from In Vitro Uptake Clearance by Suspended Human Hepatocytes for Anionic Drugs with High Binding to Human Albumin: Improvement of In Vitro-to-In Vivo Extrapolation by Considering the “Albumin-Mediated” Hepatic Uptake Mechanism on the Basis of the “Facilitated-Dissociation Model”. *Drug Metab. Dispos.* **2019 Feb**, 47, 94–103. doi: 10.1124/dmd.118.083733. (Epub 2018 Nov 30). PMID: 30504137.

(35) Yau, E.; Olivares-Morales, A.; Ogungbenro, K.; Aarons, L.; Gertz, M. Investigation of Simplified Physiologically-Based Pharmacokinetic Models in Rat and Human. *CPT Pharmacometrics Syst. Pharmacol.* **2023 Mar**, 12, 333–345. doi: 10.1002/psp4.12911. (Epub 2023 Feb 8). PMID: 36754967; PMCID: PMC10014059.