Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Accurate prediction of $K_{p,uu,brain}$ based on experimental measurement of $K_{p,brain}$ and computed physicochemical properties of candidate compounds in CNS drug discovery

Yongfen Ma [a,b], Mengrong Jiang [b], Huma Javeria [a], Dingwei Tian [a], Zhenxia Du [a,*]

[a] *College of Chemistry, Beijing Key Laboratory of Environmentally Harmful Chemical Analysis, Beijing University of Chemical Technology, Beijing, 100029, China*

[b] *DMPK Department, Sironax (Beijing) Co., Ltd, Beijing, 102206, China*

ARTICLE INFO

Keywords:
Central nervous system (CNS)
Physicochemical parameters
Brain-to-plasma unbound fraction ratio ($f_{u,b}/f_{u,p}$)
Unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$)
Quantitative structure-activity relationship (QSAR)

ABSTRACT

A mathematical equation model was developed by building the relationship between the $f_{u,b}/f_{u,p}$ ratio and the computed physicochemical properties of candidate compounds, thereby predicting $K_{p,uu,brain}$ based on a single experimentally measured $K_{p,brain}$ value. A total of 256 compounds and 36 marketed published drugs including acidic, basic, neutral, zwitterionic, CNS-penetrant, and non-CNS penetrant compounds with diverse structures and physicochemical properties were involved in this study. A strong correlation was demonstrated between the $f_{u,b}/f_{u,p}$ ratio and physicochemical parameters (CLogP and ionized fraction). The model showed good performance in both internal and external validations. The percentages of compounds with $K_{p,uu,brain}$ predictions within 2-fold variability were 80.0 %–83.3 %, and more than 90 % were within a 3-fold variability. Meanwhile, “black box” QSAR models constructed by machine learning approaches for predicting $f_{u,b}/f_{u,p}$ ratio based on the chemical descriptors are also presented, and the ANN model displayed the highest accuracy with an RMSE value of 0.27 and 86.7 % of the test set drugs fell within a 2-fold window of linear regression. These models demonstrated strong predictive power and could be helpful tools for evaluating the $K_{p,uu,brain}$ by a single measurement parameter of $K_{p,brain}$ during lead optimization for CNS penetration evaluation and ranking CNS drug candidate molecules in the early stages of CNS drug discovery.

1. Introduction

In the drug discovery process, the brain-to-plasma concentration ratio ($K_{p,brain}$) is commonly used to evaluate drug brain penetration and this parameter has been used as the primary parameter to optimize brain drug delivery in Central Nervous System (CNS) drug discovery for many years. However, with the progress in brain penetration evaluation, its relevance has been questioned. Several researchers proposed that it is difficult to assess brain penetration based upon $K_{p,brain}$ alone [1–4] because it is the unbound concentration of the pharmacologically active entity that matters. The total concentration could largely be due to molecules bound to the brain parenchyma and unable to reach the intended target. It follows that optimizing such parameters may be counterproductive and lead to molecules with very high lipophilicity and ultimately detrimental to efficacy. The unbound brain-to-plasma concentration ratio

* Corresponding author. College of Chemistry, Beijing University of Chemical Technology, Beijing, 100029, China.
E-mail address: duzx@mail.buct.edu.cn (Z. Du).

https://doi.org/10.1016/j.heliyon.2024.e24304
Received 22 October 2023; Received in revised form 4 January 2024; Accepted 5 January 2024
Available online 9 January 2024

Table 1

Measured $K_{p,brain}$, $f_{u,p}$, $f_{u,b}$, $K_{p,uu,brain}$, predicted $K_{p,uu,brain}$ values and physicochemical properties of 36 marketed published drugs in external validation group.

Drug name	MW	Class	$K_{p,brain}$	$K_{p,uu,brain}$	$f_{u,p}$	$f_{u,b}$	CLogP ADMET Predictor	Acidic_pKa	Basic_pKa	Ionized fraction		TPSA	HBD	HBA	Predicted $K_{p,uu,brain}$	$K_{p,uu,brain}$ Predicted/Observed
			Ref. [1,7]					Ref. [1]		fanion	fcation	ADMET Predictor				
Buspirone	386	Basic	1.60	1.29	0.273	0.220	1.78		7.5		0.5573	69.6	0	7	0.80	0.62
Carisoprodol	260	Neutral	0.66	0.34	0.389	0.202	2.26	11.0		0.0003		90.7	2	6	0.34	1.00
Carbamazepine	236	Neutral	0.76	0.27	0.324	0.116	2.41	10.9		0.0003		46.3	1	3	0.36	1.32
Chlorpromazine	319	Basic	23.0	0.65	0.035	0.001	5.31		9.4		0.9901	31.8	0	2	1.07	1.65
Citalopram	324	Basic	5.10	0.68	0.231	0.031	3.86		9.6		0.9937	36.3	0	3	0.57	0.84
Clozapine	327	Basic	4.10	1.01	0.038	0.009	3.67		7.1		0.3339	30.9	1	4	0.75	0.74
Cyclobenzaprine	275	Basic	12.0	1.62	0.054	0.007	4.79		9.2		0.9844	3.2	0	1	0.77	0.47
Diazepam	285	Neutral	2.00	1.02	0.098	0.050	2.80					32.7	0	3	0.74	0.73
Fluvoxamine	318	Basic	6.10	1.32	0.039	0.008	3.20		9.4		0.9901	56.8	1	4	1.01	0.76
Fluoxetine	309	Basic	12.0	0.89	0.031	0.002	4.39		10.1		0.9980	21.3	1	2	0.97	1.09
Haloperidol	376	Basic	13.0	1.06	0.087	0.007	3.90		8.3		0.8882	40.5	1	3	1.51	1.42
Hydrocodone	297	Basic	2.10	1.96	0.590	0.550	0.69		8.5		0.9264	38.8	0	4	1.63	0.83
Hydroxyzine	375	Basic	7.70	1.51	0.052	0.010	2.99		7.1		0.3339	35.9	1	4	2.11	1.40
Lamotrigine	256	Neutral	1.10	0.64	0.380	0.220	1.98					90.7	2	5	0.67	1.05
Meprobamate	218	Neutral	0.42	0.42	0.760	0.760	0.93	11.3		0.0001		104.6	2	6	0.48	1.14
Metoclopramide	300	Basic	1.20	0.52	0.710	0.310	2.32	11.4	9.6	0.0001	0.9937	67.6	2	5	0.34	0.64
Methylphenidate	233	Basic	12.0	3.43	0.770	0.220	2.25		10.6		0.9994	38.3	1	3	3.50	1.02
Midazolam	326	Neutral	0.23	0.14	0.046	0.027	2.70					30.2	0	3	0.09	0.67
Morphine	285	Basic	0.46	0.72	0.320	0.500	1.06	9.82	8.3	0.0038	0.8882	52.9	2	4	0.29	0.41
Nortriptyline	263	Basic	11.0	1.63	0.031	0.005	3.90		10.1		0.9980	12.0	1	1	1.19	0.73
9-OH-Risperidone	426	Basic	0.06	0.02	0.330	0.086	2.29		7.9		0.7597	84.4	1	7	0.02	1.25
Paroxetine	329	Basic	3.30	0.86	0.015	0.004	3.46		10.3		0.9987	39.7	1	4	0.47	0.54
Phenacetin	179	Neutral	0.87	0.55	0.701	0.442	1.64	11.6		0.0001		38.3	1	3	0.65	1.19
Phenytoin	252	Acidic	0.63	0.28	0.183	0.081	2.09	8.30		0.1118		58.2	2	4	0.39	1.41
Propranolol	259	Basic	19.6	3.08	0.140	0.022	3.48		9.48		0.9918	41.5	2	3	2.74	0.89
Propoxyphene	339	Basic	2.90	0.85	0.111	0.033	4.18		9.2		0.9844	29.5	0	3	0.27	0.32
Quinidine	324	Basic	0.28	0.05	0.286	0.050	2.65		7.95		0.7801	45.6	1	4	0.07	1.49
Risperidone	410	Basic	0.78	0.26	0.204	0.067	3.23		8.4		0.9091	64.2	0	6	0.13	0.52
Selegiline	187	Basic	3.70	1.30	0.160	0.056	2.52		7.5		0.5573	3.2	0	1	1.19	0.92
Sertraline	306	Basic	24.0	1.44	0.011	0.001	4.96		9.5		0.9921	12.0	1	1	1.39	0.96
Sulpiride	341	Basic	0.08	0.06	0.760	0.630	0.86	10.2	8.9	0.0017	0.9693	110.1	2	7	0.05	0.83
Thiopental	242	Acidic	0.36	0.17	0.304	0.146	2.73	7.80		0.2847		90.3	2	4	0.18	1.02
Trazodone	372	Basic	0.61	0.56	0.051	0.047	3.32		7.2		0.3869	45.8	0	6	0.13	0.24
Venlafaxine	277	Basic	4.20	0.98	0.900	0.210	3.12		9.3		0.9876	32.7	1	3	0.73	0.75
Warfarin	308	Acidic	0.07	0.19	0.097	0.281	3.29	5.06		0.9954		67.5	1	4	0.04	0.21
Zolpidem	307	Basic	0.29	0.24	0.240	0.200	2.79		6.9		0.2403	152.7	0	4	0.09	0.39

Note: $K_{p,brain}$, $f_{u,p}$ and $f_{u,b}$ data of Phenacetin, quinidine, warfarin, and propranolol were determined in-house, CLogP, pKa, TPSA, HBD, HBA data were predicted by ADMET Predictor. All data of the other 32 compounds were reference reported, except the CLogP values were predicted by ADMET Predictor.

($K_{p,uu,brain}$) provides a more meaningful value for the extent of blood-brain barrier (BBB) transport, where brain exposure is normalized to systemic exposure [4–6]. In this case, a straightforward method to reliably estimate $K_{p,uu,brain}$ is essential. The common method to estimate $K_{p,uu,brain}$ in preclinical species is to determine the in vivo $K_{p,brain}$, the unbound fraction in plasma ($f_{u,p}$), and the unbound fraction in the brain ($f_{u,b}$), respectively [7]. However, these experimental approaches are labor-intensive, time-consuming, and offer a limited sample throughput capacity, thus limiting their application to support compound optimization in early drug discovery.

Multiple in silico models for the prediction of $K_{p,brain}$ or $K_{p,uu,brain}$ that were built on different datasets have been reported in recent years [8–18], but the results were not satisfactory. The development of $K_{p,uu,brain}$ models is inherently challenging because they must account for the roles of all membrane transporters at the BBB [19]. As of today, P-glycoprotein (P-gp) is the only transporter whose role in brain penetration is relatively well characterized. Other xenobiotic ATP-binding case (ABC) and solute carrier (SLC) transporters such as breast cancer resistance protein (BCRP), multidrug resistance-associated proteins (MRPs), organic anion transporting polypeptides (OATPs), and organic cation transporters (OCTs) are also expressed at the human BBB [20–23], but very little is known about their properties and roles in brain penetration. Morena [14] proposed a statistical model by using experimental $K_{p,brain}$, in silico predicted $f_{u,p}$ and $f_{u,b}$. Experimentally measured $K_{p,brain}$ can directly eliminate the concern about the incomplete investigation of the transporters. However, the model performed poorly as the prediction variability was as high as 10-fold. It is speculated that the prediction bias may be superimposed since the $f_{u,p}$ and $f_{u,b}$ were predicted separately.

The $K_{p,uu,brain}$ of a drug versus $K_{p,brain}$ is described by the following expressions:

$$K_{p,uu,brain} = K_{p,brain} * \text{factor},$$

where $\text{factor} = f_{u,b}/f_{u,p}$

The connection between $K_{p,uu,brain}$ and $K_{p,brain}$ is the brain-to-plasma unbound fraction ratio ($f_{u,b}/f_{u,p}$). It is important to note that due to a very poor correlation, $f_{u,p}$ is not a suitable surrogate for $f_{u,b}$ [24]. In the plasma, albumin and α-1-acid glycoprotein were thought to account for the binding of most drugs. While in the brain tissue, phospholipids drive the non-specific binding [25]. Plasma has twice as much protein as the brain and the brain has 20-fold more lipids than plasma [26]. The very different lipid and protein contents of the two compartments lead to the poor correlation between $f_{u,b}$ and $f_{u,p}$. There has been much research demonstrating that distribution is a function of relative tissue and plasma protein binding, and the protein binding is related to the physicochemical properties of the drugs [27–30]. Based on the information above, it is necessary to conduct detailed studies to explore the relationship between the physicochemical properties and the difference in the binding to plasma and the brain.

The main objective of this study is to explore the possibility of building a relationship between $f_{u,b}/f_{u,p}$ ratio and computed physicochemical properties such as the lipophilicity (LogP, octanol-water partition coefficient), ionization (pKa), and possibly others unknown, thereby developing a mathematical equation model for predicting the $K_{p,uu,brain}$ of compounds based on a single experimental measured $K_{p,brain}$ value. To construct the dataset, 256 structurally diverse in-house small molecules were selected, and their K_p,

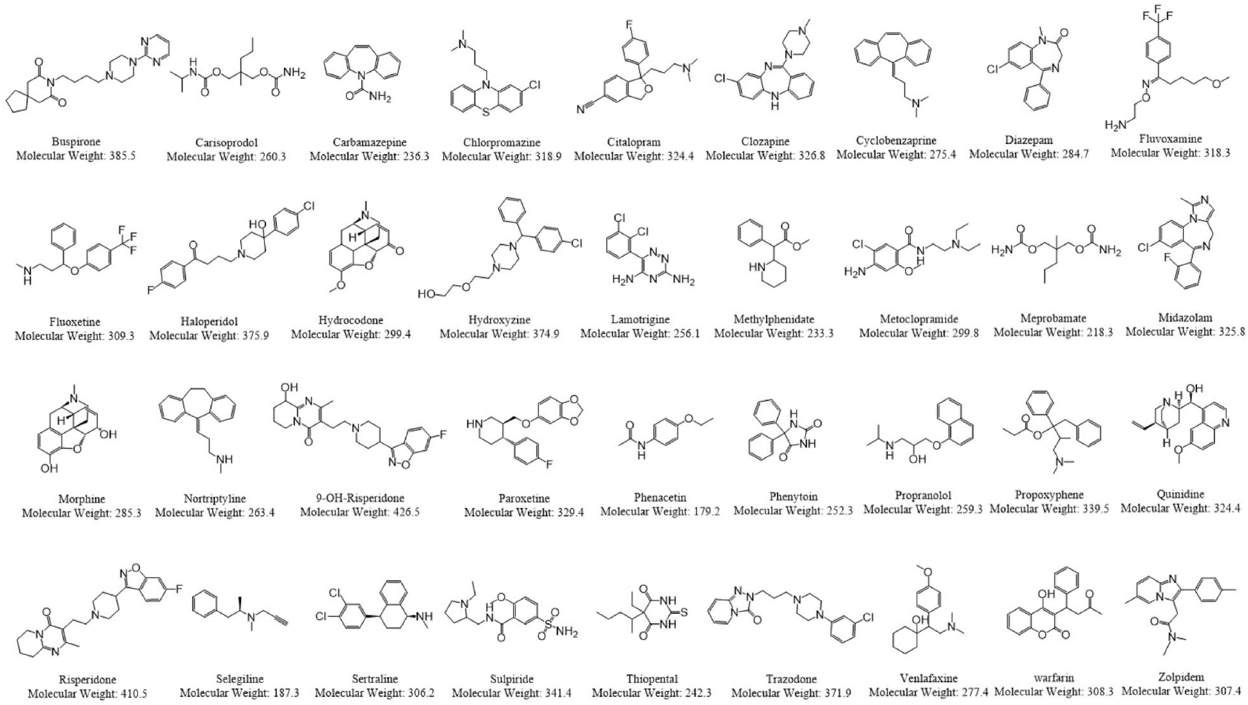

Fig. 1. The structures of 36 marketed drugs in external model validation group. The $K_{p,brain}$ values were from 0.06 to 24.0, and the CLogP ranged from 0.69 to 5.31.

$_{brain}$, $f_{u,b}$ and $f_{u,p}$ values were experimentally determined. Various types of physicochemical properties have been computed. Linear and nonlinear correlation analysis methods were employed to investigate the correlation and sensitivity of each physicochemical parameter with $f_{u,b}/f_{u,p}$ ratio. This work aims at prioritizing the brain penetration potential of discovery compounds at an early stage while reducing resource consumption in the determination of $f_{u,b}$ and $f_{u,p}$.

2. Materials and Methods

2.1. Compounds selection

Three to five compounds were selected from each structural series of different CNS projects in Sironax (Beijing) Co., Ltd, A total of 256 compounds were selected including acidic, basic, neutral, zwitterionic, CNS-penetrant, and non-CNS penetrant compounds with diverse structures and physicochemical properties. This dataset was divided into two parts, 226 compounds for mathematical equation model building (Table S1) and 30 compounds for internal model validation (Table S2). In addition, 32 marketed CNS drugs in the published literature were used for external validation of the model (Table 1). Phenacetin, quinidine, warfarin, and propranolol, used as control drugs in the protein binding assay in-house, comprised in the set of marketed drugs with published structures, were also used for external validation of the model (Table 1). The structures of these 36 marketed drugs are shown in Fig. 1.

2.2. Experimental animals

To get the $K_{p,brain}$ values of the studied compounds, CD-1 mice and SD rats of approximately eight weeks of age, weighing about 30g and 220g, respectively, were purchased from Beijing Vital River Laboratory Animal Technology Co., Ltd. The animals were confirmed to be healthy before assignment to the study. Upon arrival, the animals were maintained for at least three days on a 12-h light/dark cycle in a temperature- and humidity-controlled environment with free access to food and water. The animals were housed in clear polycarbonate boxes (three per box) containing sawdust and nesting pads.

2.3. Determination method of $K_{p,brain}$

The modified Cassette-Dosing Approach of up to three compounds was used for the determination of $K_{p,brain}$ [31]. Using this approach, CD-1 mice (n = 3) were administered a single intraperitoneal dose of a mixture of three compounds at 3–10 mg/kg. One plasma and one brain sample were collected at 2 h post-dose. It should be noted that the $K_{p,brain}$ values for model building in this study were determined by a single measurement point of plasma and brain concentration after a single dose, not by the area under the curve (AUC). To compare the difference of $K_{p,uu,brain}$ between the mice and rats, the $K_{p,brain}$ values of 15 compounds in both mice and rats were determined according to the AUC of brain and plasma concentrations at 0.5, 4, 8, and 24hr after a single oral dose.

2.4. Equilibrium dialysis method for $f_{u,b}$ and $f_{u,p}$ measurement

An equilibrium dialysis apparatus (HTDialysis, Cat# 1006) was employed to determine the $f_{u,b}$ and $f_{u,p}$ of mice or rats for each compound. A dialysis membrane with a molecular cutoff of 13K–14K Da was used for dialysis. Plasma and brain homogenate (10 % w/v in 100 mM sodium phosphate buffer, pH = 7.4) were collected from CD-1 mice. These samples were spiked with a test compound at 1 μM and dialyzed against an equal volume of the sodium phosphate buffer. Phenacetin, quinidine, and warfarin were selected as control drugs in $f_{u,p}$ assay. Propranolol [32] was selected as the control drug in $f_{u,b}$ assay. The 96-well equilibrium dialysis apparatus was maintained at 37 °C for 5 h. Post dialysis, the plasma and the 10 % (w/v) brain homogenate samples were mixed with equal volumes of sodium phosphate buffer. The buffer obtained from the apparatus was mixed with an equal volume of either blank plasma or blank brain homogenate. Then these samples were mixed with five volumes of acetonitrile, vortexed, and centrifuged at 3000g for 10 min at 4 °C, the supernatants were analyzed by LC-MS/MS. The $f_{u,p}$ and $f_{u,b}$ were calculated using the equations below:

$f_{u,p}$ = Conc. buffer/Conc. plasma

$f_{u,b}$' = Conc. buffer/Conc. brain homogenate

$$f_{u,b} = 1 \Big/ \left(D * \left(1 \Big/ f_{u,b}'\text{-}1\right) + 1\right)$$

where D and $f_{u,b}$' represent the dilution factor for the brain homogenate and the unbound fraction determined in the 10 % (w/v) brain homogenate, respectively.

2.5. LC-MS/MS detection

Mice or rats' plasma, brain homogenate, and sodium phosphate buffer samples for all compounds were processed by protein precipitation with methanol-acetonitrile (1:1) and analyzed by an AB Sciex API 5500 plus tandem mass spectrometer equipped with a Shimadzu ExionLC AD binary high-pressure gradient pump controlled via PE-Sciex sample control software. Analyst 1.7.1 was used for data acquisition and quantitation.

2.6. *Physicochemical properties calculation of test compounds*

LogP, pKa, molecular weight (MW), Topological Polar Surface Area (TPSA), number of hydrogen bond acceptors (HBA), and number of hydrogen bond donors (HBD) of 256 in-house synthesized small molecule compounds were calculated by ADMET Predictor v10.3.0.0 (Simulations Plus, Inc, Lancaster, CA, USA) based on the molecular structure. Molecular structures were imported in batches and parameters can be calculated in high throughput. The physicochemical properties of 36 marketed published drugs were taken from the literature or calculated by ADMET predictor if data were unavailable from the literature.

2.7. *Statistical analysis and software availability of model development*

t-SNE: t-SNE (t-distributed stochastic neighbor embedding) was used to evaluate structural diversity for 256 in-house compounds and 36 marketed drugs. Molecules are represented by Morgan fingerprint (radius 3, 1024 bits) using RDKit (version 2022.09.5, open source), and then openTSNE (version 0.7.1, open source) was used to perform t-SNE dimensionality reduction. The t-SNE visualization was plotted by Vortex (version 22.1.119751-s, Dotmatics Limited 2007–2022).

Internal Similarity: Internal similarity was measured by all pairwise molecular similarities in each molecule dataset. The molecular similarity was calculated as Tanimoto similarity in RDKit (version 2022.09.5, open source) using Morgan fingerprint (radius 3, 1024 bits). The density of internal similarity was plotted by Matplotlib (version 3.7.0, open source).

Parameter Sensitivity Analysis: To evaluate the relevance and importance of each physicochemical parameter (including CLogP, TPSA, HBA, HBD, and ionized fraction) in the construction of the mathematical equation model, correlation analysis was performed using the techniques of Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) configured in ADMET Predictor software. In addition, a single-parameter linear regression analysis was also conducted with the linear coefficient of correlation (R^2) as an indicator.

"Black Box" QSAR Models: Several machine learning techniques, including ANN, MLR, Support Vector Machine (SVM), and Kernel Partial Least Squares (PLS) methods, were employed to build "black box" Quantitative Structure-Activity Relationship (QSAR) models, to ensure the robustness of the models, each model type was run randomly at least 10 times and finally selected the best-performing model based on the RMSE value and prediction accuracy.

K-fold cross-validation: "Black box" QSAR models were validated with the method of "K-fold cross-validation" to safeguard the models against random bias caused by the selection of only one training and validation set. K-fold cross-validation was performed by Anaconda (version 4.10.3, python version 3.9.7, open source), the molecular descriptor corresponding to the final selected model was taken as the independent variable, and the measured value of f_{ub}/f_{up} was taken as the dependent variable. The ANN/SVM/PLS/MLR model code was customized in Python, and the modeling data was used as input for cross-verification to compare these four modeling methods.

Model performance evaluation: The correlation of the values between the observed and predicted was evaluated using the average fold error (AFE) and residual mean squared error (RMSE), which are indicators of accuracy and precision, respectively. AFE and RMSE were calculated by using equations Eq. 1 and Eq. (2) [33]. The percentage of compounds within 2-fold or 3-fold variability was also used as indicators for prediction accuracy.

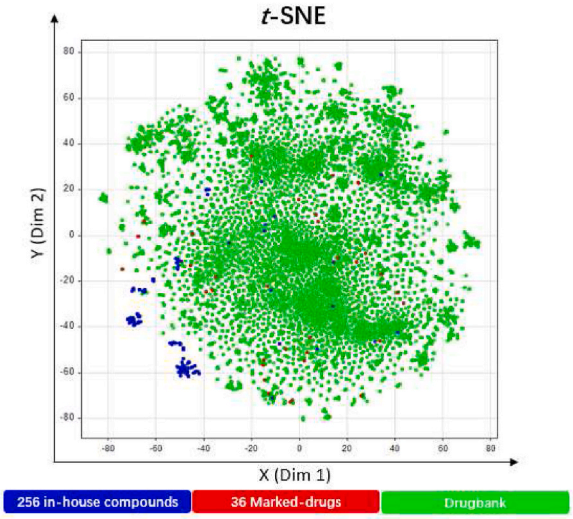

Fig. 2. t-SNE method derived structural diversity for 256 in-house compounds and 36 marketed drugs with Drugbank molecules as background (10485 molecules). X (Dim1) and Y (Dim2) are the projections in 2-dimensions from a multidimensional descriptor space. The blue dot represents 256 in-house compounds, the red dot represents 36 marketed drugs, and the green dot represents the distribution of molecules in Drugbank (10485 molecules).

$$\mathrm{AFE} = 10^{\frac{\sum_{k=1}^{n} \log \frac{\text{predicted Kp,uu}}{\text{observed Kp,uu}}}{n}} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} * \Sigma_{k=1}^{n} (\text{predicted Kp, uu} - \text{observed Kp, uu})^2} \tag{2}$$

where n represents the size of the dataset and k represents kth data.

3. Result

3.1. Data collection

$K_{p,brain}$, $f_{u,b}$, and $f_{u,p}$ values of 256 compounds were determined as described in the Materials and Methods section and are listed in Table S1 (training set) and Table S2 (internal validation set). In the data set for model external validation (Table 1), $K_{p,brain}$, $f_{u,b}$, and $f_{u,p}$ values of 32 marketed drugs were obtained from the literature [1,7], while data of phenacetin, quinidine, warfarin, and propranolol (control drugs in protein binding assay) were determined in house. The range of $K_{p,brain}$ values spanned from 0.06 to 28.5. The range of $f_{u,b}$, and $f_{u,p}$ was from 0.1 %~90 %. The CLogP values ranged from 0.655 to 5.31. The lowest acidic pKa and highest basic pKa were used for acidic and basic compounds, respectively. The selected compounds/drugs have a wide range of physicochemical properties with diverse structures, which meet the requirements of model development.

3.2. Structural diversity evaluation

The t-SNE plot was used to evaluate structural diversity for 256 in-house compounds and 36 marketed drugs. As shown in Fig. 2, these molecules are almost evenly distributed and have sufficient coverage of the chemical space projected by the drugs in the Drugbank (10485 molecules). Furthermore, the distribution of both our in-house compounds and the selected marketed drugs is very similar to that of the drugs in the Drugbank (Fig. 3 (A-D), 10485 molecules). Both results indicate reasonable structural diversity of these molecules.

3.3. Parameter Sensitivity Analysis and mathematical equation model exploration

Two data modeling techniques, MLR and ANN, were employed to assess the importance of each variable. First, 256 compounds were involved in the analysis dataset without using the test set setting. CLogP, f_{cation}, and f_{anion} (ionized fraction of bases and acids at pH 7.4) showed a good correlation with the $f_{u,b}/f_{u,p}$ ratio (relative sensitivity was more than 0.7), as shown in Table 2. Then, a 10 % minimum test set size was defined with the Kohonen self-organizing map selection method [34], taking the RMSE as the evaluation indicator to assess the correlation of the $f_{u,b}/f_{u,p}$ ratio with different combinations of variables. This process was repeated 10 times with ANN and MLR, respectively, 7 of 10 runs in ANN, and all runs in MLR showed the best RMSE results with the combination including S + LogP, f_{cation}, and f_{anion}.

In the single-parameter linear regression analysis, the ratio of $f_{u,b}/f_{u,p}$ was negatively correlated with CLogP in an exponential trend ($R^2 = 0.32$, n = 256), negatively correlated with basic pKa/f_{cation} ($R^2 = 0.36$, n = 182), and negatively/positively correlated with acidic pKa/f_{anion} ($R^2 = 0.33$, n = 49), respectively. While TPSA, HBA, and HBD were less relevant (R^2 less than 0.13, n = 256). The correlation analysis of the $f_{u,b}/f_{u,p}$ ratio with the computed physicochemical properties of the study compounds is shown in Fig. 4.

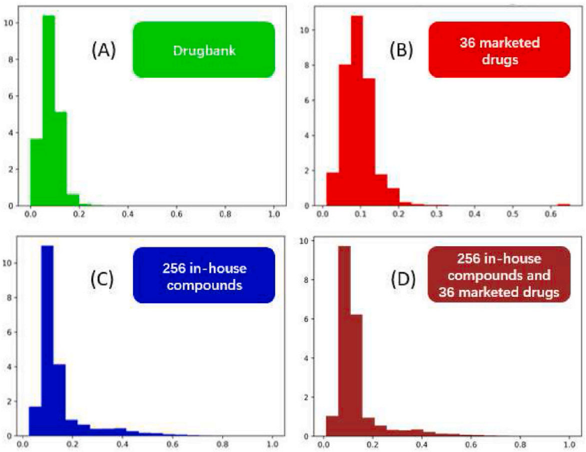

Fig. 3. Similarity distribution for 256 in-house compounds and 36 marketed drugs with Drugbank molecules as background. (A) Drugbank molecules (B) 36 marked-drugs (C) 256 in-house compounds (D) 256 in-house compounds and 36 marked-drugs.

Table 2
Relative sensitivity of physicochemical parameters assessed by MLR and ANN.

Index	Parameters	Sensitivity	Relative Sensitivity
1	S + LogP	0.478	1
2	f_{anion}	0.349	0.731
3	f_{cation}	0.343	0.718
4	HBD	0.111	0.233
5	HBA	0.109	0.228
6	TPSA	0.016	0.033
7	$f_{zwitterion}$	0.006	0.013

These 256 compounds were randomly divided into two groups. 226 compounds (Table S1) were used for mathematical equation model exploration and development (training set), and the other 30 compounds (Table S2) were used for internal validation of the model (test set). Based on the results of parameter correlation studies, with 226 compounds as the dataset and taking RMSE as the indicator, a correlation between the $f_{u,b}/f_{u,p}$ ratio and the computed physicochemical properties was established by optimizing the model equation by adjusting the weight for one of the parameters while fixing the other two (Eq. (3)). Finally, an equation for prediction of the $K_{p,uu,brain}$ based on experimental $K_{p,brain}$ and physicochemical properties was developed (Eq. (4)).

$$\frac{fu,b}{fu,p}=\frac{2*10^{(0.35*fanion)}}{10^{(0.25*fcation)}*e^{(0.6*CLogP)}} \tag{3}$$

$$Kp,uu,brain=2*Kp,brain\ *\ \frac{10^{(0.35*fanion)}}{10^{(0.25*fcation)}*e^{(0.6*CLogP)}} \tag{4}$$

where f_{cation} and f_{anion} designate the ionized fraction of basic and acidic compounds at pH 7.4, and were calculated as follows (Eq. (5) and Eq. (6)):

$$fcation=\frac{100}{1+10^{(7.4-basic\ pKa)}} \tag{5}$$

$$fanion=\frac{100}{1+10^{(acidic\ pka-7.4)}} \tag{6}$$

3.4. Predictive performance of the mathematical equation model

A total of 226 compounds were in the training set, and relevant information used for modeling is summarized in Table S1. The correlation of observed and predicted values for $K_{p,uu,brain}$ is shown in Fig. 5, and the model performance parameters are summarized in Table 3. The percentage of compounds with $K_{p,uu,brain}$ predicted within 2-fold variability was 82.3 %, and 92.5 % within 3-fold variability. The models' accuracy and reliability were further demonstrated by the AFE value of 1.13 and RMSE value of 0.2. The predicted $K_{p,uu,brain}$ was within 2-fold of the observed $K_{p,uu,brain}$ for 186 of 226 (82.3 %) compounds, which is a significant achievement given the physicochemical property differences and the range of $K_{p,brain}$ that covered two orders of magnitude. This result should support the use of the predicted $K_{p,uu,brain}$ for rank order compounds in the early stage of CNS drug discovery.

3.5. Internal validation of the mathematical equation model with 30 additional compounds

Internal validation was performed to assess the predictive performance of the model with 30 additional compounds (Table S2). The f_{cation} and f_{anion} were calculated using Eq. (5) and Eq. (6), respectively. When the $K_{p,brain}$ values and their corresponding physicochemical parameters of these 30 compounds were put into Eq. (4), it was observed that 80.0 % of the compounds fell within a 2-fold variability, and 93.3 % of the compounds fell within a 3-fold variability (as shown in Table 3 and Fig. 6). The values of RMSE and AFE were 0.28 and 0.99, respectively, indicating that limited prediction bias was observed for $K_{p,uu,brain}$ in the model.

3.6. External validation of the mathematical equation model with 36 marketed published drugs

A total of 36 marketed drugs (32 literature-reported drugs and four control drugs in protein binding assay in-house, Table 1) were selected to assess the predictive performance of the model. The $K_{p,brain}$, $f_{u,p}$, $f_{u,b}$, and pKa values of 32 published drugs were from the literature [1,7], and the CLogP values were calculated by ADMET Predictor. The $K_{p,brain}$ values ranged from 0.06 to 24.0, and the CLogP ranged from 0.69 to 5.31. The f_{cation} and f_{anion} were calculated using Eq. (5) and Eq. (6), respectively. Put the $K_{p,brain}$ values and the corresponding physicochemical parameters of these 36 compounds into Eq. (4), it was observed that 83.3 % of the drugs fell within a 2-fold window of linear regression, and 91.7 % within a 3-fold variability (as shown in Fig. 7 and Table 3). The values of RMSE and AFE were 0.30 and 0.80, respectively.

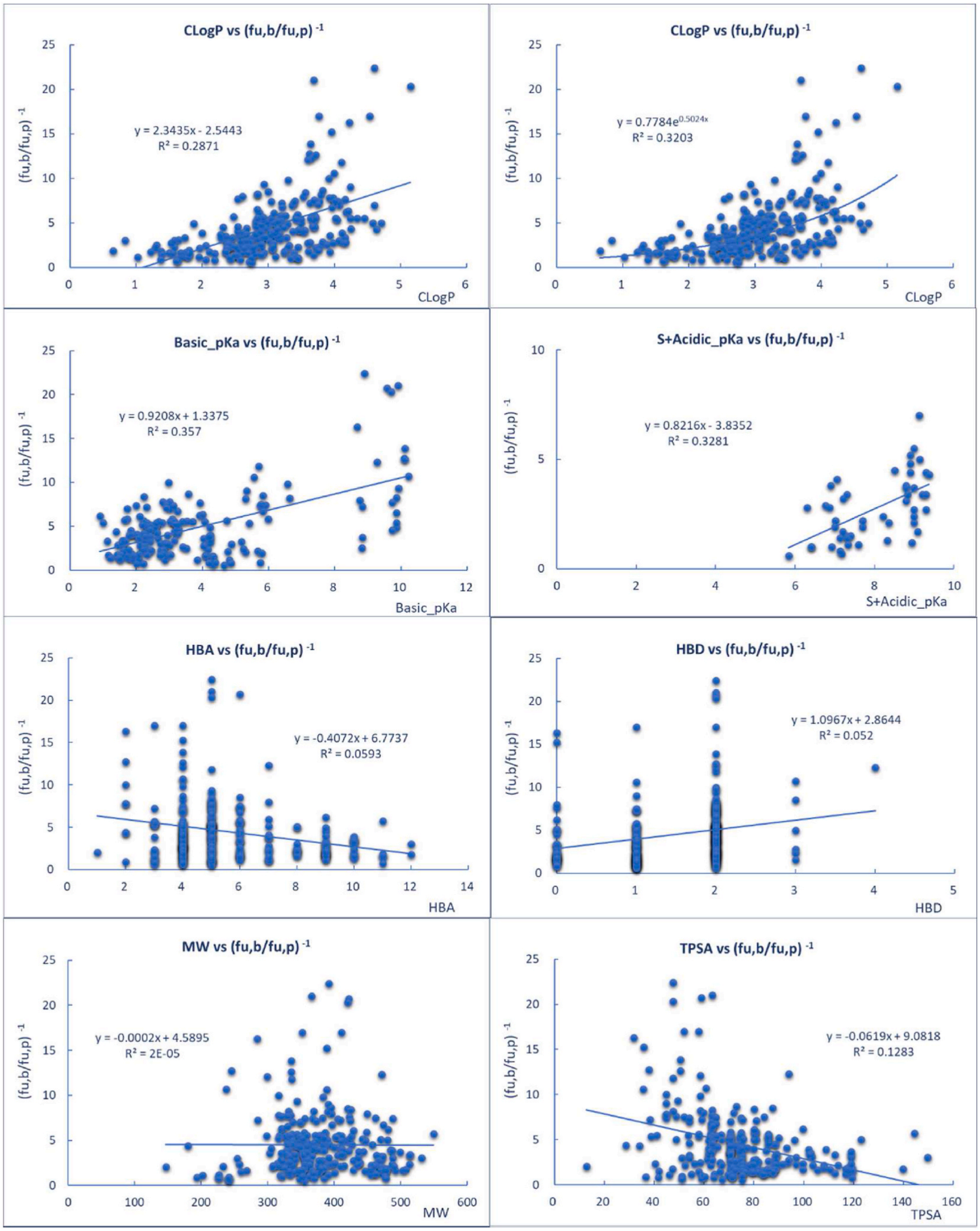

Fig. 4. Single-parameter linear regression analysis results for studied compounds. A total of 256 compounds including acidic, basic, neutral, and zwitterionic compounds with diverse structures and physicochemical properties were involved. The $f_{u,b}$ and $f_{u,p}$ values of studied compounds were determined in-house, and the physicochemical properties were calculated by ADMET Predictor v10.3.0.0 (Simulations Plus, Inc, Lancaster, CA, USA). The ratio of $f_{u,b}/f_{u,p}$ was negatively correlated with CLogP, especially in an exponential trend ($R^2 = 0.32$, $n = 256$), negatively correlated with basic pKa/f_{cation} ($R^2 = 0.36$, $n = 182$), and negatively/positively correlated with acidic pKa/f_{anion} ($R^2 = 0.33$, $n = 49$), respectively. While TPSA, HBA, HBD, and MW were less relevant ($R^2 < 0.13$, $n = 256$).

3.7. Development and performance of “black box” QSAR models

Machine learning approaches have become popular in recent years. It is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions. In this study, an attempt was also made to use machine learning approaches to build QSAR models based on the same dataset. A total of

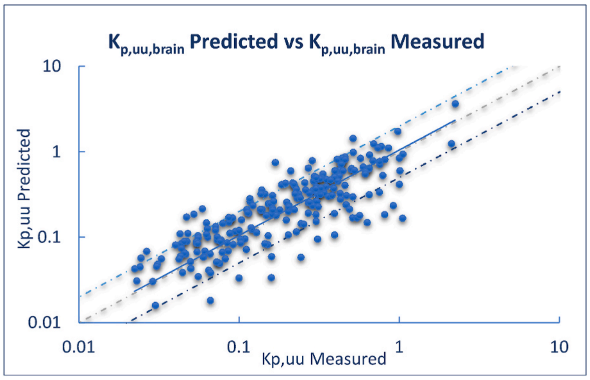

Fig. 5. The plot of predicted $K_{p,uu,brain}$ vs measured $K_{p,uu,brain}$ for 226 compounds in the training dataset. The percentage of compounds with $K_{p,uu,brain}$ predicted within 2-fold variability was 82.3 %, and 92.5 % within 3-fold variability. The dotted lines represent the 2-fold window of linear regression. Solid lines are the result of linear regression analysis of log-transformed data.

Table 3
Model performance statistics.

	Number of Compounds	AFE	RMSE	% Within 2-fold	% Within 3-fold
Training dataset	226	1.13	0.23	82.3 %	92.5 %
Internal validation	30	0.99	0.28	80.0 %	93.3 %
External validation	36	0.80	0.30	83.3 %	91.7 %

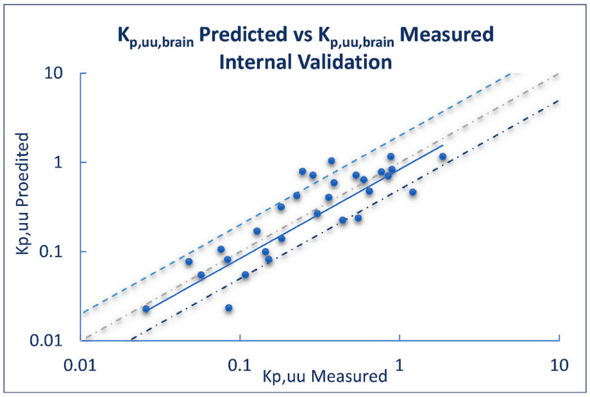

Fig. 6. The plot of predicted $K_{p,uu,brain}$ vs measured $K_{p,uu,brain}$ for 30 additional compounds in the internal validation dataset. The result showed that 80.0 % of the compounds fell within a 2-fold variability, and 93.3 % of the compounds fell within a 3-fold variability. The dotted lines represent the 2-fold window of linear regression. Solid lines are the result of linear regression analysis of log-transformed data.

292 compounds (the same dataset as the Mathematical Equation Model, 256 in-house compounds, and 36 marketed drugs) were randomly divided into two groups. Using the training set of 262 compounds to analyze the structure-$f_{u,p}/f_{u,b}$ ratio relationships and derive statistical models, the other 30 compounds in the test set (see Table S3) were used for external validation of the models. Several machine learning techniques including ANN, SVM, MLR, and PLS, were employed to build "black box" QSAR Models. The internal consistency of the resulting models was evaluated by k-fold cross-validation performed in Python, and taking the mean absolute error (MAE), the average across experiments, as the scoring parameter of model quality to report. The results are shown in Table 4, the ANN model displayed the best performance with an MAE value of 0.16.

The results obtained from the test set observations are reported in Table 5. Overall, these models had good predictive power, with around 80.0 % of the test drugs falling within 2-fold variability. Here, the ANN model displayed the highest accuracy in the predictions with an RMSE value of 0.27 and 86.7 % of the test drugs falling within 2-fold variability. The correlation of the observed and predicted values for $f_{u,p}/f_{u,b}$ in the ANN model is shown in Fig. 8.

3.8. *Comparison of $K_{p,uu,brain}$ between mice and rats*

Different brain penetration between rats and non-rodent animals (monkey/dog) has been reported [35]. As mice and rats are the

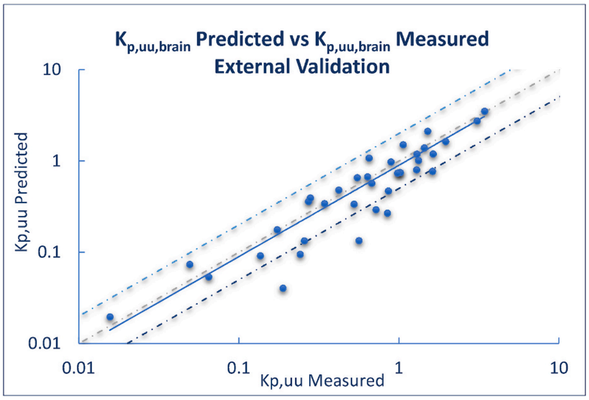

Fig. 7. The plot of predicted $K_{p,uu,brain}$ vs observed $K_{p,uu,brain}$ for the 36 marketed published drugs in the external validation dataset. The result showed that 83.3 % of the drugs fell within a 2-fold window of linear regression, and 91.7 % within a 3-fold variability. The dotted lines represent the 2-fold window of linear regression. Solid lines are the result of linear regression analysis of log-transformed data.

Table 4
K-fold (K = 10) cross-validation results for different QSAR models.

Models	Number of Molecular descriptors	MAE
ANN	68	0.16
SVM	51	0.21
MLR	85	0.32
PLS	102	0.23

Table 5
Test set results for different QSAR models.

Models	AFE	RMSE	% Within 2-fold	% Within 3-fold
ANN	0.99	0.27	86.7 %	100 %
SVM	0.91	0.31	76.7 %	93.3 %
MLR	1.03	0.32	80 %	100 %
PLS	0.95	0.31	80 %	100 %

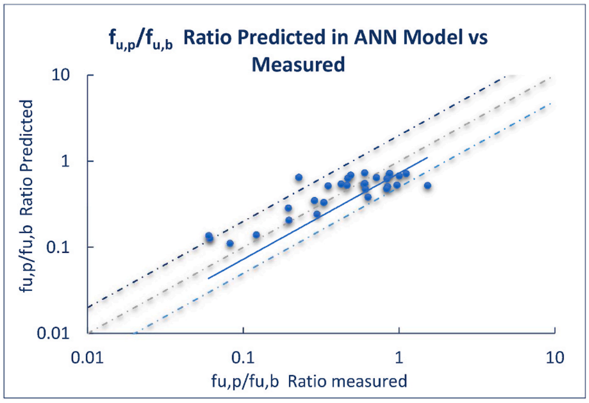

Fig. 8. The plot of predicted $f_{u,p}/f_{u,b}$ ratio vs observed $f_{u,p}/f_{u,b}$ ratio for 30 compounds in the test set in the ANN model. The result showed that 86.7 % of the drugs fell within a 2-fold window of linear regression, and 100 % within a 3-fold variability. The dotted lines represent the 2-fold window of linear regression. Solid lines are the result of linear regression analysis of log-transformed data.

most frequently used animals for evaluation of in vivo pharmacological effects, it is interesting to compare whether drug partition to mouse and rat brain tissues to different degrees. The default assumption is that mouse and rat $K_{puu,\ brain}$ should be the same. In this study, the $K_{p,uu,brain}$ values were measured in both mice and rats for the same set of compounds with a $K_{p,uu,brain}$ range from 0.03 to 1.95 as described in section 2.3. The results are shown in Table S4 and Fig. 9, a tight correlation (RMSE = 0.24, $R^2 = 0.92$, n = 15) of $K_{p,}$

$_{uu,brain}$ between mice and rats was observed, indicating that limited difference between mice and rats based on these 15 compounds dataset, which may support $K_{p,uu,brain}$ data transfer for mice and rats in pre-clinical studies. It is necessary to collect more data and extend the dataset to validate this conclusion in the future.

3.9. Comparison of $K_{p,brain}$ value across timepoints in mice

74 compounds with both 1hr and 4hr $K_{p,brain}$ values in the internal database were available to perform the analysis. As shown in Fig. 10, 23 % of the compounds were out of a 2-fold window of linear regression. By analyzing the physicochemical parameters and PK data, we found that these compounds have one or more of the following characteristics: 1) basic compounds with high LogP result in extensive brain binding and large distribution volume; 2) high plasma CL; 3) unstable in brain homogenate. These properties may contribute to the slow equilibrium and non-parallel concentration-time (c-t) profiles in plasma and brain. Thus, the AUC ratio or steady-state concentrations ratio is considered an appropriate approach for the $K_{p,brain}$ determination of these compounds. While consistent $K_{p,brain}$ values were observed for most compounds with high permeability and no/little BBB efflux, which facilitates rapid equilibrium and thus obtained parallel c-t profiles in plasma and brain.

4. Discussion

Developing in silico models to predict the brain penetration of drugs is a difficult task owing to the intricate involvement of multiple transport systems in the blood-brain barrier, and the necessity to consider a combination of multiple pharmacokinetic parameters. However, the models developed in this study performed exceptionally well when compared to other models reported in the past decade (as shown in Table 6). This is due to the fact that the dataset used in this work possesses reasonable structural diversity, and the physicochemical parameters, LogP and ionized fraction, were explored to be the most correlated parameters to the $f_{u,b}/f_{u,p}$ ratio. This solid data foundation and robust correlation contribute to the model's good generalization ability. Additionally, the experimentally measured $K_{p,brain}$ eliminated concerns about incomplete investigation of the transporters. Finally, building a relationship between the $f_{u,b}/f_{u,p}$ ratio and the computed physicochemical properties may reduce prediction bias compared to predicting $f_{u,p}$ and $f_{u,b}$ separately. Upon analyzing the data of compounds whose accuracy is out of 2- or 3-fold variation, it was found that some of them have very high protein binding rates, with a $f_{u,b}$ or $f_{u,p}$ value less than 0.5 %, which may be beyond the precision of the assay itself. This could be a possible reason for the discrepancy in predictions.

One point to note here is that the mice $K_{p,brain}$ values used in the model development were determined by a single point of plasma and brain concentration after a single dose, not based on the AUC or steady-state concentrations. The major goal of the present study is to build a relationship between the $f_{u,b}/f_{u,p}$ ratio and the computed physicochemical properties or molecular descriptors. In the process of model development, $K_{p,brain}$ is not a dependent or independent variable, and plays a role that does not affect the model building. In addition, the determination of AUC or steady-state concentrations is time-consuming, labor-intensive, and may not be necessary at the early screening stage.

Unsurprisingly, CLogP is the most relevant parameter in the statistical models described in this work due to its evident relationship with f_u. The statistical analysis indicates a strong relationship between CLogP and the $f_{u,b}/f_{u,p}$ ratio, and therefore CLogP can be effectively used as a guideline to roughly judge the numerical gap between $K_{p,brain}$ and $K_{p,uu,brain}$. The higher the CLogP value, the greater the gap between $K_{p,brain}$ and $K_{p,uu,brain}$. Compounds with a CLogP value smaller than zero deviate significantly from the 2-fold window seen during modeling, with more than a 4-fold deviation from measured values (as shown in Table S5). Numerous studies have described the physicochemical properties required for optimal brain exposure, medicinal chemists working in CNS drug discovery attempt to modify the LogP of a compound to greater than one (generally between two and four) [36–39]. No additional analysis was performed because such compounds should be relatively few in CNS drug discovery. Compounds with a CLogP <0 were deemed not applicable and were removed from the model building.

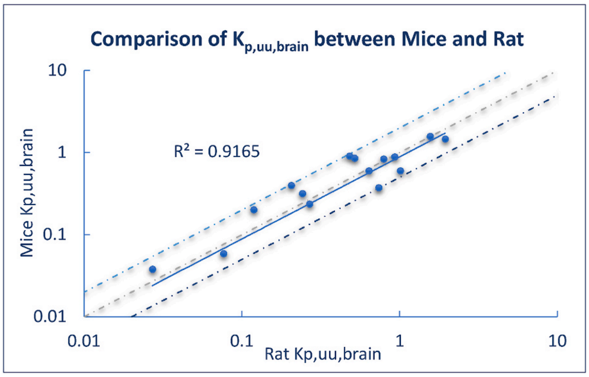

Fig. 9. The plot of rats $K_{p,uu,brain}$ vs mice $K_{p,uu,brain}$ for the 15 compounds. A tight correlation of $K_{p,uu,brain}$ between mice and rats was observed. The dotted lines represent the 2-fold window of linear regression. Solid lines are the result of linear regression analysis of log-transformed data.

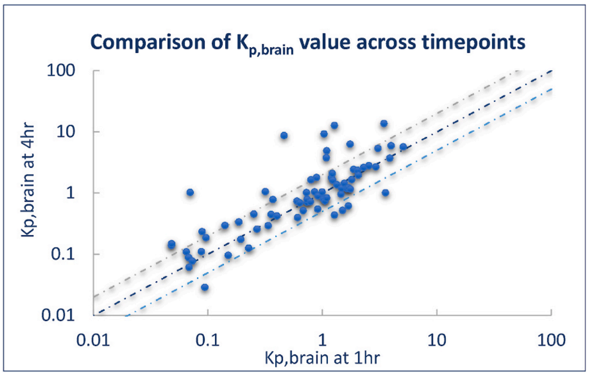

Fig. 10. The plot of mice $K_{p,brain}$ at 1hr vs 4hr for 74 compounds. 23 % of the compounds were out of a 2-fold window of linear regression. The dotted lines represent the 2-fold window of linear regression.

The most extensive binding to brain tissue observed in this study was observed among the basic compounds with $f_{u,b}$ smaller than 0.01. As such, the range of $f_{u,b}$ estimates were much wider among basic compounds, which in turn resulted in a wider range of $f_{u,b}/f_{u,p}$ ratios for basic compounds than was observed for the weakly acidic or neutral compounds. The degree of ionization can make a big difference in the protein binding and distribution of a molecule. Basic compounds, being positively charged at physiological pH, have favorable interactions with anionic phospholipid head groups leading to higher tissue affinity [40]. Therefore, basic compounds tend to have a higher tissue distribution (K_p) than neutral and acidic compounds. Ion trapping in lysosomes (a condition characterized by the accumulation of phospholipids and drugs in lysosomes, cationic amphiphilic drugs with high LogP and basic pKa are well-known structural features) is a typical example [41–43]. Acidic drugs exhibit extensive binding to albumin in plasma, typically have a low volume of distribution (V_d) value (<1 L*kg^{-1}), thus a lower $K_{p,brain}$ value [10,44], and show a relatively high $f_{u,b}/f_{u,p}$ ratio (acidic pKa/f_{anion} is negatively/positively correlated with $f_{u,b}/f_{u,p}$ ratio). This study involved 52 acidic compounds (43 in the training dataset, 6 in the internal validation group, and 3 in the external validation group). Many of them were weak acids (nine compounds with an acidic pKa between 5.0 and 7.0, while others were between 7.0 and 9.4), and few strong acids were available. Therefore, prediction accuracy is limited and a more extensive validation is required for strong acidic compounds. Further research to collect a larger amount of relatively strong acidic drugs will be necessary for both prediction accuracy and to validate the model. In this study, basic and acidic compounds are defined as having a basic pKa >5.4 and an acidic pKa <9.4, respectively. All charges were considered separately to ensure that, for example, for zwitterionic molecules (basic pKa >5.4 and acidic pKa <9.4), they would not cancel out.

Of relevance to drug design, the equation in the model can provide a line of sight to chemists in assessing the numerical gap between $K_{p,brain}$ and $K_{p,uu,brain}$, and achieving a balance of properties that is necessary for brain penetration success. A basic compound, with high LogP and a basic pKa will lead to a large gap between the value of $K_{p,brain}$ and $K_{p,uu,brain}$. Take chlorpromazine as an example, a strong basic drug with a basic pKa of 9.4 and a high CLogP value of 5.3, the value of $K_{p,brain}$ is 35 times that of $K_{p,uu,brain}$. While if a neutral or weak acidic compound (acidic pKa >7), the numerical gap between $K_{p,brain}$ and $K_{p,uu,brain}$ is mainly determined by the LogP, the higher the LogP, the larger the gap between the value of $K_{p,brain}$ and $K_{p,uu,brain}$. For the acidic compounds with acidic pKa <7, which generally show relatively lower $K_{p,brain}$ values, however, the ratio of $f_{u,b}/f_{u,p}$ is positively correlated with the ionized fraction, and the difference between $K_{p,brain}$ and $K_{p,uu,brain}$ are relatively small, sometimes the $K_{p,brain}$ values are even smaller than $K_{p,uu,brain}$. For example, warfarin, an acidic drug with an acidic pKa value of 5.2 and a moderate CLogP value of 3.3, showed a low $K_{p,brain}$ value of 0.07, but the $K_{p,uu,brain}$ value was 2.89 times that of $K_{p,brain.}$

In addition, a total of 292 compounds were studied in this work, of which 279 compounds had a $K_{p,uu,brain}$ value smaller than the $K_{p,brain}$ value, and only 4.8 % (14 compounds) had a $K_{p,uu,brain}$ value slightly greater than the $K_{p,brain}$ value (1.04–2.89 folds, the acidic drug warfarin produced the fold value of 2.89). This suggests that compounds with a lower $K_{p,brain}$ value (e.g., smaller than 0.2), notably for the basic and neutral compounds, may be less significant to proceed with $K_{p,uu,brain}$ determinations in CNS drug discovery. Alternately, the $K_{p,brain}$ cutoff value may be changed to better suit the demands of each project.

5. Conclusion

A mathematical model through building the relationship between the $f_{u,b}/f_{u,p}$ ratio and the computed physicochemical properties to project $K_{p,uu,brain}$ based on experimental $K_{p,brain}$ values is presented. A total of 256 compounds and 36 marketed published drugs including acidic, basic, neutral, and zwitterionic compounds with diverse structures and physicochemical properties were involved in this study. The model showed good performance in both internal and external validations. Some 79.3 %–81.1 % of the compounds with predicted $K_{p,uu,brain}$ values were within 2-fold variability, and more than 90 % were within 3-fold variability. The statistical analysis in this study indicates that CLogP and the ionized fraction of the compounds showed a strong correlation with the $f_{u,b}/f_{u,p}$ ratio. ClogP can be effectively used as a guideline to roughly judge the numerical gap between $K_{p,brain}$ and $K_{p,uu,brain}$. The effect of the ionization degree of basic and acidic compounds on the affinity to plasma proteins and tissue phospholipids is discussed, and the $f_{u,b}/f_{u,p}$ ratio was negatively correlated with FCation and positively correlated with FAnion.

Meanwhile, “black box” QSAR models based on the chemical descriptors are also presented, and the ANN model displayed the

Table 6
Comparisons of in silico models of $K_{p,uu,brain}$ prediction within the years of 2013–2023.

Datasets	Experimental approaches to generate $K_{p,uu,brain}$	In silico models	Model performance	Refs/Year
Training set: 40 compounds **Test set: 93 literature compounds**	$K_{p,brain}$: mice or rat. $f_{u,b}$: mice or rat, brain homogenate/brain slice/microdialysis methods $f_{u,p}$: mice or rat plasma equilibrium dialysis	Indirect regression QSAR model using experimental $K_{p,brain}$, and in silico predictions for $f_{u,b}$ and $f_{u,b}$	Poor predictive performance of the model as the accuracy within 10-fold error: $R^2 = 0.74–0.89$ against different test sets.	[14]/2013
Training set: 29 compounds **Internal test set: 11 compounds** **External test: 41 literature compounds.**	$K_{p,brain}$: rats or mice $V_{u,brain}$: rat brain slice $f_{u,p}$: rat or mouse plasma	Direct regression QSAR model using the multivariate PLS analysis	The best model against the test set: $R^2 = 0.82$ and RMSE = 0.31. However, the model performed poorly in the external validation.	[12]/2014
Training set: 242 compounds **Test set: 104 compounds**	$K_{p,brain}$: rat, 4h intravenous infusion. $f_{u,b}$: rat brain slice $f_{u,p}$: rat plasma equilibrium dialysis	Direct and indirect regression QSAR models using two nonlinear machine learning algorithms (RF and SVM)	In the best consensus model: $R^2 = 0.60$ and RMSE = 0.53.	[11]/2015
Training set: 677 compounds **Test set: 169 compounds**	$K_{p,brain}$ from different designs $f_{u,b}$: rat brain homogenate $f_{u,p}$: rat blood or plasma	Direct binary classification ($K_{p,uu,brain}$ > or < 0.3) QSAR models using nonlinear categorical model-building algorithms.	*Classification* model: accuracy = 0.75–0.79.	[8]/2016
Training set: 1030 compounds **Test set: 91 compounds**	$K_{p,brain}$: mice $f_{u,b:}$ mouse brain homogenate $f_{u,p:}$ mouse plasma	Regression QSAR Models.	Regression model: $R^2 = 0.53$ and RMSE = 0.57.	[13]/2016
Training set/Test set: **$f_{u,b}$ model: 505/46 compounds** **$f_{u,p}$ model: 462/45 compounds** **P-gp NER model: 397/50 compounds** **$K_{p,uu,brain}$: 42 compounds**	Datasets of $f_{u,brain}$, $f_{u,p}$, P-gp NER, were constructed from in-house experiments, publicly available data in ChEMBL, and previous study findings	Built three prediction models for in vitro P-gp NER, $f_{u,b}$ and $f_{u,p,}$ and validated using additional in-house experiment data. $K_{p,uu,brain}$ was calculated based on predicted P-gp NER, $f_{u,b}$ and $f_{u,p.}$	The percentage of compounds that fell within 5- and 10-fold errors were 66.7 % and 73.8 %, respectively.	[15]/2021
Training set: 88 published compounds **External test: 38 in-house compounds**	$K_{p,uu,brain}$: mice or rats, published data	An RF model was used for the predictive QSAR model.	Accuracy: RMSE = 0.455, $R^2 = 0.726$. External validation: RMSE = 0.491, $R^2 = 0.438$.	[16]/2022
Rat dataset: 512 training data and 128 test data. **Monkey dataset 51 compounds.** **Human dataset 14 compounds.**	$K_{p,brain}$: rats or monkeys fast iv $f_{u,b}$ and $f_{u,p}$: equilibrium dialysis MDCK-MDR1 or MDCK-BCRP: Permeability Assay	Establish machine-learning models of monkeys and rats. Human $K_{p,uu,brain}$ prediction: consider appropriate scaling methods based on the animal models.	Accuracy within 2-fold error was 71 % and 64 % based on rat and monkey machine learning models, respectively.	[17]/2023
241 compounds from published articles and internal programs.	$K_{p,brain}$: mice or rats $f_{u,b}$ and $f_{u,p}$: mice or rats Efflux ratios: MDCKII-MDR1	A strong relationship emerged between Solvation-free energy (E-sol) and $K_{p,uu,brain}$.	A *categorical* accuracy of 79 % and an R^2 of 0.61 from a linear regression model.	[18]/2023
Mathematical equation model: **Training set: 226 compounds** **Internal test set: 30 compounds** **External test: 36 literature compounds.** **ANN model:** **Training set: 262 compounds** **Test set: 30 compounds**	$K_{p,brain}$: mice, single timepoint $f_{u,b}$ and $f_{u,p}$: equilibrium dialysis	Mathematical equation model using experimental $K_{p,brain}$, and in silico predicted ratio of $f_{u,b}/f_{u,b.}$ ANN model based on the chemical descriptors.	Mathematical equation model: RMSE = 0.455, the percentage of compounds that fell within 2-fold errors were 80 % and 83.3 % for internal and external tests, respectively. ANN model: RMSE = 0.27, external test accuracy within 2-fold error was 86.7 %	Current Work

Abbreviations: unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$), brain-to-plasma concentration ratio ($K_{p,brain}$), unbound fraction in plasma ($f_{u,p}$), unbound fraction in the brain ($f_{u,b}$), unbound volume of distribution ($V_{u,brain}$), Quantitative Structure-Activity Relationship (QSAR), random forest (RF), Support Vector Machine (SVM), Kernel Partial Least Squares (PLS), Artificial Neural Network (ANN), net efflux ratio (NER), residual mean squared error (RMSE), linear coefficient of correlation (R^2).

highest accuracy with an RMSE value of 0.27 and 86.7 % of the test set drugs fell within a 2-fold window of linear regression. Another contribution of this work is that the observed consistent $K_{p,uu,brain}$ in mice and rats based on 15 compounds dataset, which may support $K_{p,uu,brain}$ data transfer for the two commonly used species in pre-clinical pharmacology studies.

To the best of our knowledge, this is the first time that a relationship between the $f_{u,b}/f_{u,p}$ ratio and computed physicochemical properties has been demonstrated, thereby predicting the $K_{p,uu,brain}$ based on a single experimental $K_{p,brain}$ value. The author believes that a mathematical equation model with competitive prediction accuracy has better practicability than "black box" QSAR models that rely on statistical software, especially for research institutions where commercial software is unavailable. In summary, the proposed new in silico approaches for the rapid assessment of brain penetration for candidate compounds should benefit CNS target projects.

Ethics statement

The animal experiments were performed in accordance with the National Institutes of Health – Office of Laboratory Animal Welfare policies and laws and approved by the Institutional Animal Use and Care Committee of Sironax Co., Ltd. Ethics approval number: SIRONAX-2022001, SIRONAX-2022002, and SIRONAX-2022003.

Data availability statement

The experimental data and computed physicochemical parameters of in-house compounds that were utilized for model development can be found in the "Supporting Information" document. However, due to confidentiality reasons, the structures of these compounds are not provided. The data information along with the structures of 36 marketed drugs used in the external test were provided in the manuscript.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Yongfen Ma: Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Mengrong Jiang:** Writing - review & editing, Validation, Software, Formal analysis. **Huma Javeria:** Writing - review & editing. **Dingwei Tian:** Software. **Zhenxia Du:** Writing - review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Many thanks to Sironax (Beijing) Co., Ltd, for providing and sharing the structure information of the compounds studied in this work. We would like to thank Dr. Zhaolan Zhang and Dr. Charles Liu, for their valuable comments on the manuscript and kind help with the analysis of molecular structure diversity.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e24304.

References

[1] A. Doran, R.S. Obach, B.J. Smith, N.A. Hosea, S. Becker, E. Callegari, C. Chen, X. Chen, E. Choo, J. Cianfrogna, L.M. Cox, J.P. Gibbs, M.A. Gibbs, H. Hatch, C. E. Hop, I.N. Kasman, J. Laperle, J. Liu, X. Liu, M. Logman, D. Maclin, F.M. Nedza, F. Nelson, E. Olson, S. Rahematpura, D. Raunig, S. Rogers, K. Schmidt, D. K. Spracklin, M. Szewc, M. Troutman, E. Tseng, M. Tu, J.W. Van Deusen, K. Venkatakrishnan, G. Walens, E.Q. Wang, D. Wong, A.S. Yasgar, C. Zhang, The impact of P-glycoprotein on the disposition of drugs targeted for indications of the central nervous system: evaluation using the MDR1A/1B knockout mouse model, Drug Metab. Dispos. 33 (1) (2005) 165–174.

[2] X. Liu, C. Chen, B.J. Smith, Progress in brain penetration evaluation in drug discovery and development, J. Pharmacol. Exp. Therapeut. 325 (2) (2008) 349–356.

[3] M. Hammarlund-Udenaes, M. Friden, S. Syvanen, A. Gupta, On the rate and extent of drug delivery to the brain, Pharm. Res. (N. Y.) 25 (8) (2008) 1737–1750.

[4] L. Di, H. Rong, B. Feng, Demystifying brain penetration in central nervous system drug discovery, Miniperspective, J Med Chem 56 (1) (2013) 2–12.

[5] M. Friden, A. Gupta, M. Antonsson, U. Bredberg, M. Hammarlund-Udenaes, In vitro methods for estimating unbound drug concentrations in the brain interstitial and intracellular fluids, Drug Metab. Dispos. 35 (9) (2007) 1711–1719.

[6] M. Hammarlund-Udenaes, Active-site concentrations of chemicals - are they a better predictor of effect than plasma/organ/tissue concentrations? Basic Clin. Pharmacol. Toxicol. 106 (3) (2010) 215–220.

[7] T.S. Maurer, D.B. Debartolo, D.A. Tess, D.O. Scott, Relationship between exposure and nonspecific binding of thirty-three central nervous system drugs in mice, Drug Metab. Dispos. 33 (1) (2005) 175–181.

[8] Y.Y. Zhang, H. Liu, S.G. Summerfield, C.N. Luscombe, J. Sahi, Integrating in silico and in vitro approaches to predict drug accessibility to the central nervous system, Mol. Pharm. 13 (5) (2016) 1540–1550.

[9] M. Friden, S. Winiwarter, G. Jerndal, O. Bengtsson, H. Wan, U. Bredberg, M. Hammarlund-Udenaes, M. Antonsson, Structure-brain exposure relationships in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids, J. Med. Chem. 52 (20) (2009) 6233–6243.

[10] H. Chen, S. Winiwarter, M. Friden, M. Antonsson, O. Engkvist, In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms, J. Mol. Graph. Model. 29 (8) (2011) 985–995.

[11] S. Varadharajan, S. Winiwarter, L. Carlsson, O. Engkvist, A. Anantha, T. Kogej, M. Friden, J. Stalring, H. Chen, Exploring in silico prediction of the unbound brain-to-plasma drug concentration ratio: model validation, renewal, and interpretation, J. Pharmaceut. Sci. 104 (3) (2015) 1197–1206.

[12] I. Loryan, V. Sinha, C. Mackie, A. Van Peer, W.H. Drinkenburg, A. Vermeulen, D. Heald, M. Hammarlund-Udenaes, C.M. Wassvik, Molecular properties determining unbound intracellular and extracellular brain exposure of CNS drug candidates, Mol. Pharm. 12 (2) (2014) 520–532.

[13] E. Dolgikh, I.A. Watson, P.V. Desai, G.A. Sawada, S. Morton, T.M. Jones, T.J. Raub, QSAR model of unbound brain-to-plasma partition coefficient, Kp,uu,brain: incorporating P-glycoprotein efflux as a variable, J. Chem. Inf. Model. 56 (11) (2016) 2225–2233.

[14] M. Spreafico, M.P. Jacobson, In silico prediction of brain exposure: drug free fraction, unbound brain to plasma concentration ratio and equilibrium half-life, Curr. Top. Med. Chem. 13 (7) (2013) 813–820.

[15] R. Watanabe, T. Esaki, R. Ohashi, M. Kuroda, H. Kawashima, H. Komura, Y. Natsume-Kitatani, K. Mizuguchi, Development of an in silico prediction model for P-glycoprotein efflux potential in brain capillary endothelial cells toward the prediction of brain penetration, J. Med. Chem. 64 (5) (2021) 2725–2738.

[16] Y. Umemori, K. Handa, S. Sakamoto, M. Kageyama, T. Iijima, QSAR model to predict Kp,uu,brain with a small dataset, incorporating predicted values of related parameter, SAR QSAR Environ. Res. 33 (11) (2022) 885–897.

[17] S. Liu, Y. Kosugi, Human brain penetration prediction using scaling approach from animal machine learning models, AAPS J. 25 (5) (2023).

[18] M. Lawrenz, M. Svensson, M. Kato, K.H. Dingley, J. Chief Elk, Z. Nie, Y. Zou, Z. Kaplan, H.R. Lagiakos, H. Igawa, E. Therrien, A computational physics-based approach to predict unbound brain-to-plasma partition coefficient, Kp,uu, J. Chem. Inf. Model. 63 (12) (2023) 3786–3798.

[19] H. Liu, K. Dong, W. Zhang, S.G. Summerfield, G.C. Terstappen, Prediction of brain:blood unbound concentration ratios in CNS drug discovery employing in silico and in vitro model systems, Drug Discov. Today 23 (7) (2018) 1357–1372.

[20] C. International Transporter, K.M. Giacomini, S.M. Huang, D.J. Tweedie, L.Z. Benet, K.L. Brouwer, X. Chu, A. Dahlin, R. Evers, V. Fischer, K.M. Hillgren, K. A. Hoffmaster, T. Ishikawa, D. Keppler, R.B. Kim, C.A. Lee, M. Niemi, J.W. Polli, Y. Sugiyama, P.W. Swaan, J.A. Ware, S.H. Wright, S.W. Yee, M.J. Zamek-Gliszczynski, L. Zhang, Membrane transporters in drug development, Nat. Rev. Drug Discov. 9 (3) (2010) 215–236.

[21] Y.T.K. Nguyen, H.T.T. Ha, T.H. Nguyen, L.N. Nguyen, The role of SLC transporters for brain health and disease, Cell. Mol. Life Sci. 79 (1) (2021).

[22] M.M. Parvez, A. Sadighi, Y. Ahn, S.F. Keller, J.O. Enoru, Uptake transporters at the blood–brain barrier and their role in brain drug disposition, Pharmaceutics 15 (10) (2023).

[23] K.M. Huttunen, T. Terasaki, A. Urtti, A.B. Montaser, Y. Uchida, Pharmacoproteomics of brain barrier transporters and substrate design for the brain targeted drug delivery, Pharmaceut. Res. 39 (7) (2022) 1363–1392.

[24] X. Liu, K. Van Natta, H. Yeo, O. Vilenski, P.E. Weller, P.D. Worboys, M. Monshouwer, Unbound drug concentration in brain homogenate and cerebral spinal fluid at steady state as a surrogate for unbound concentration in brain interstitial fluid, Drug Metab. Dispos. 37 (4) (2009) 787–793.

[25] P. Poulin, F.P. Theil, A priori prediction of tissue:plasma partition coefficients of drugs to facilitate the use of physiologically-based pharmacokinetic models in drug discovery, J. Pharmaceut. Sci. 89 (1) (2000) 16–35.

[26] Z. Rankovic, CNS drug design: balancing physicochemical properties for optimal brain exposure, J. Med. Chem. 58 (6) (2015) 2584–2608.

[27] M. Gibaldi, P.J. McNamara, Apparent volumes of distribution and drug binding to plasma proteins and tissues, Eur. J. Clin. Pharmacol. 13 (5) (1978) 373–380.

[28] M. Gertz, P.J. Kilford, J.B. Houston, A. Galetin, Drug lipophilicity and microsomal protein concentration as determinants in the prediction of the fraction unbound in microsomal incubations, Drug Metab. Dispos. 36 (3) (2008) 535–542.

[29] F. Lombardo, R.S. Obach, M.Y. Shalaeva, F. Gao, Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data, J. Med. Chem. 45 (13) (2002) 2867–2876.

[30] G. Berellini, F. Lombardo, An accurate in vitro prediction of human VDss based on the øie–tozer equation and primary physicochemical descriptors. 3. Analysis and assessment of predictivity on a large dataset, Drug Metabol. Dispos. 47 (12) (2019) 1380–1387.

[31] X. Liu, X. Ding, G. Deshmukh, B.M. Liederer, C.E. Hop, Use of the cassette-dosing approach to assess brain penetration in drug discovery, Drug Metab. Dispos. 40 (5) (2012) 963–969.

[32] H. Wan, M. Rehngren, F. Giordanetto, F. Bergström, A. Tunek, High-throughput screening of Drug−Brain tissue binding and in silico prediction for assessment of central nervous system drug delivery, J. Med. Chem. 50 (19) (2007) 4606–4615.

[33] S. Sato, K. Matsumiya, K. Tohyama, Y. Kosugi, Translational CNS steady-state drug disposition model in rats, monkeys, and humans for quantitative prediction of brain-to-plasma and cerebrospinal fluid-to-plasma unbound concentration ratios, AAPS J. 23 (4) (2021) 81.

[34] A. Yan, J. Gasteiger, Prediction of aqueous solubility of organic compounds by topological descriptors, QSAR Comb. Sci. 22 (8) (2003) 821–829.

[35] J.G. Kettle, S.K. Bagal, D. Barratt, M.S. Bodnarchuk, S. Boyd, E. Braybrooke, J. Breed, D.J. Cassar, S. Cosulich, M. Davies, N.L. Davies, C. Deng, A. Eatherton, L. Evans, L.J. Feron, S. Fillery, E.S. Gleave, F.W. Goldberg, M.A. Cortés González, C. Guerot, A. Haider, S. Harlfinger, R. Howells, A. Jackson, P. Johnström, P. D. Kemmitt, A. Koers, M. Kondrashov, G.M. Lamont, S. Lamont, H.J. Lewis, L. Liu, M. Mylrea, S. Nash, M.J. Niedbala, A. Peter, C. Phillips, K. Pike, P. Raubo, G. R. Robb, S. Ross, M.G. Sanders, M. Schou, I. Simpson, O. Steward, Discovery of AZD4747, a potent and selective inhibitor of mutant GTPase KRASG12C with demonstrable CNS penetration, J. Med. Chem. 66 (13) (2023) 9147–9160.

[36] B.B. Freeman 3rd, L. Yang, Z. Rankovic, Practical approaches to evaluating and optimizing brain exposure in early drug discovery, Eur. J. Med. Chem. 182 (2019) 111643.

[37] S.A. Hitchcock, L.D. Pennington, Structure-brain exposure relationships, J. Med. Chem. 49 (26) (2006) 7559–7583.

[38] M.J. Waring, Defining optimum lipophilicity and molecular weight ranges for drug candidates-Molecular weight dependent lower logD limits based on permeability, Bioorg Med Chem Lett 19 (10) (2009) 2844–2851.

[39] H. van de Waterbeemd, G. Camenisch, G. Folkers, J.R. Chretien, O.A. Raevsky, Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors, J. Drug Target. 6 (2) (1998) 151–165.

[40] D.A. Smith, L. Di, E.H. Kerns, The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery, Nat. Rev. Drug Discov. 9 (12) (2010) 929–939.

[41] J.P. Ploemen, J. Kelder, T. Hafmans, H. van de Sandt, J.A. van Burgsteden, P.J. Saleminki, E. van Esch, Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential: a case study with structurally related piperazines, Exp. Toxicol. Pathol. 55 (5) (2004) 347–355.

[42] H. Fischer, E.A. Atzpodien, M. Csato, L. Doessegger, B. Lenz, G. Schmitt, T. Singer, In silico assay for assessing phospholipidosis potential of small druglike molecules: training, validation, and refinement using several data sets, J. Med. Chem. 55 (1) (2012) 126–139.

[43] U.M. Hanumegowda, G. Wenke, A. Regueiro-Ren, R. Yordanova, J.P. Corradi, S.P. Adams, Phospholipidosis as a function of basicity, lipophilicity, and volume of distribution of compounds, Chem. Res. Toxicol. 23 (4) (2010) 749–755.

[44] M. Lobell, L. Molnar, G.M. Keseru, Recent advances in the prediction of blood-brain partitioning from molecular structure, J. Pharmaceut. Sci. 92 (2) (2003) 360–370.