Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 43 (2008) 1581–1592
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

In silico prediction of brain and CSF permeation of small molecules using PLS regression models

Stefanie Bendels [a,*], Manfred Kansy [a], Björn Wagner [a], Jörg Huwyler [b]

[a] *F. Hoffmann-La Roche Ltd., Pharmaceutical Research, Discovery Chemistry, Grenzacherstrasse, CH-4070 Basel, Switzerland*
[b] *University of Applied Sciences Northwestern Switzerland, Institute of Pharma Technology, Gruendenstrasse 40, CH-4132 Muttenz, Switzerland*

Received 3 July 2007; received in revised form 13 November 2007; accepted 19 November 2007
Available online 26 November 2007

Abstract

Computational partial least square (PLS) regression models were developed, which can be applied to predict central nervous system (CNS) penetration of drug-like organic molecules. For modeling, a dataset of 77 structurally diverse compounds was used with reported steady-state rat brain to plasma ratios (BPR). Information on steady-state cerebrospinal fluid distribution (CSF to plasma ratio or CSFPR) was available for 37 of these compounds. The molecules were from different chemical series and included bases, acids, zwitterions and neutral molecules. They were CNS active and were therefore assumed to penetrate the blood–brain barrier and/or the blood–liquor barrier. Using these PLS models, the dataset could be described accurately ($r^2 = 0.78$, StErrorEst = 0.30 and $r^2 = 0.75$, StErrorEst = 0.28 for BPR and CSFPR, respectively). Molecular descriptors used for the prediction of passive membrane transport were lipophilicity, polar and hydrophobic surface areas as well as structural parameters and net charge at physiological pH. There was no apparent correlation between experimental brain and CSF exposure. Consequently, different PLS models and guiding rules were developed and discussed for the prediction of BPR or CSFPR. The present models provide a cost-effective and efficient strategy to guide synthetic efforts in medicinal chemistry at an early stage of the drug discovery and development process.

Keywords: Blood–brain barrier; CSF-plasma and brain-plasma ratios; PLS model; CNS active; Passive transcellular transport

1. Introduction

There is an increasing interest in predicting the process of passive translocation of drugs from the blood stream to the brain, in particular for pharmaceutical companies focusing on the development of drugs that act on targets in the central nervous system (CNS). The pharmacological activity of such CNS medicines not only depends on receptor affinity but also on the achieved compound concentration in brain. In many instances, however, access of chemicals to the brain is restricted at the level of the brain capillary endothelial wall that forms the blood–brain barrier (BBB). Direct measurements of BBB permeability or brain uptake of drugs is difficult and time-consuming and requires sophisticated *in vitro* experimental systems or animal experiments. Some examples of used techniques include *in vitro* models of the BBB [1,2], *in vivo* pharmacokinetic and tissue distribution studies [3,4] or *in situ* brain perfusion and capillary depletion experiments [5,6]. In a clinical setting, access to brain tissue is not possible. It has therefore been suggested to use cerebrospinal fluid (CSF) as a surrogate marker for drug concentrations in brain tissue since this CNS compartment is accessible by lumbar puncture in human or ventricular puncture in experimental animals [7]. All mentioned techniques can be applied to detailed mechanistic studies with selected test compounds, however, their use for routine drug screening is not possible due to their limited throughput. This situation has created an interest in predictive *in silico* permeability models, which can be used to analyze compound libraries in an industrial setting and to guide medicinal chemists in drug discovery and development [8–11].

It was the aim of the present study to develop and validate a computational blood–brain barrier permeation model, which

* Corresponding author. Tel.: +41 61 68 85254.
E-mail address: stefanie.bendels@roche.com (S. Bendels).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2007.11.011

can be used to predict the extent of passive uptake of drug-like organic molecules into brain tissue as well as CSF. We could thereby make use of a proprietary in-house database, which contains brain permeation data for 77 drug-like molecules and CSF exposure data for a subset of 37 molecules. All compounds were CNS active, derived from several structurally unrelated chemical series and were assumed not to be substrates of drug transporting proteins such as P-glycoprotein [12]. It was not possible to describe the dataset with existing brain permeation models, such as the quantitative structure–activity relationship (QSAR) model by Clark [13]. This as well as our interest in CNS permeation of drugs led to the development of a new and improved partial least square (PLS) regression model.

2. Materials and methods

2.1. In vivo dataset

The proprietary in-house dataset in this study consists of CNS active compounds. These drug-like compounds are from different chemical classes covering a broad range of physicochemical and structural properties (Table 1). In the beginning, 91 compounds with reported brain to plasma ratios (BPR) were analyzed and a subset of 77 high-quality data was selected for this study (Table 2). Exclusion criteria were, for example, a poor aqueous solubility of the tested compounds (below 1 μg/mL), contributions of active transport, or reported instabilities or impurities. Information on CSF exposure (CSF to plasma ratio or CSFPR) was available for 37 of these compounds. Chemical structures of 11 compounds from the training and 15 from the test dataset are shown in Appendixes A and B, respectively. It should be noted that all compounds used in the present work are proprietary Roche compounds. The 26 chemical structures revealed in Appendixes A and B represent thus molecules, for which no licenses were granted to third parties nor other potential conflict of interest might prevent their publication.

2.2. Diversity analysis

The molecular diversity of the dataset was analyzed using average pairwise Tanimoto coefficients. The coefficients were calculated with MACCS fingerprints implemented in the software package MOE (Chemical Computing Groups Inc., MOE version 2006.08, Montreal, Quebec, Canada).

Table 1
Minimal and maximal values of important molecular parameters for 77 compounds with in-house measured log BPR values

Parameter	Range of values
log BPR (77 compounds)	−1.3–1.7
log CSFPR (37 compounds)	−2.7 to −0.3
Polar surface area (PSA)	7–110 Å^2
Hydrophobic surface	190–400 Å^2
Absolute value of the net charge at pH 7.4 (ANC 7.4)	0–1
Highest basic pKa	9.6
Lowest acidic pKa	6.7
clog P	0.9–6.2
Molecular weight (MW)	260–500 Da
Number of rings	2–5
Number of rotatable bonds	0–8

2.3. In vivo studies to investigate brain and CSF penetration

The used *in vivo* measure for brain uptake is log BPR, which is defined as the ratio between the steady-state concentrations of a test compound in brain tissue and the corresponding terminal plasma concentrations (log BPR = log 10 ([brain]/[plasma])). CSF penetration is defined accordingly as log CSFPR = log 10 ([CSF]/[plasma]). In these studies, two different experimental protocols were used to ensure that steady-state conditions were reached by the end of the *in vivo* experiments [14]: Wistar rats were dosed orally with test compound followed by terminal collection of brain tissue and plasma at defined time points. In most cases, $n = 2$ animals per group were sacrificed at 0.5, 1, 2 and 4 h after dosing. Alternatively, compound was administered intravenously ($n \geq 2$ rats) by continuous infusion by an indwelling cannula implanted into the jugular vein. At defined time points during the infusion, plasma samples were taken followed by collection of terminal plasma and brain tissue at 5 h. Both methods allow for a monitoring of tissue and/or plasma concentrations over time. Data were used for the present study only if apparent steady-state conditions were reached by the end of the experiment. In some studies, cerebrospinal fluid (CSF) was collected prior to removal of brain tissue by insertion of a collection needle (0.7 × 19 mm) into the cerebellomedullary cistern (cisterna magna) [15,16]. CSF was drained by a silicon tubing (ID 0.5 mm) by capillary force into pre-weighted vials. Using this technique, it is possible to obtain ~0.1 mL of CSF from a rat. After visual inspection for contaminating blood, the CSF samples were spiked with one aliquot of plasma from untreated rats prior to analytics. All experiments were conducted in accordance with current Cantonal and Federal legislation on the welfare of experimental animals.

2.4. Analytics

For sample preparation, 50 μL of plasma-spiked CSF was mixed with 150 μL of methanol containing an internal standard. Brain tissue (i.e. one rat brain) was diluted with three volumes of water and homogenized in an ice-water bath using an ultrasonic probe. 50 μL of the homogenized brain sample was mixed with 150 μL methanol containing an internal standard. Brain and plasma samples were centrifuged, 100 μL of the upper organic phase was diluted with one volume of water and analyzed by liquid chromatography (LC) and tandem mass spectrometry [17].

2.5. Determination of HT-log D values

The applied high throughput (HT) method for the determination of distribution coefficients (HT-log D) is based on a micro plate technique and derived from the conventional

Table 2
Calculated molecular descriptors and experimental log BPR and log CSFPR values for 77 CNS active compounds used as training or test set for modeling

Identifier[a]	log BPR	Charge class[b]	ANC 7.4[c]	clog P[d]	Nb (rings)	Nb (rotatable bonds)	MW	PSA [Å^2]	log CSFPR
Training **1**	−1.01	A	0.33	3.18	5	5	455	81.41	
Training **2**	−0.25	A	0.28	3.28	4	4	383	62.47	
Training **3**	−0.69	A	0.30	3.89	5	5	443	73.44	
Training **4**	−0.64	A	0.03	2.35	4	3	378	68.62	
Training **5**	−0.59	A	0.04	3.82	3	1	300	82.46	−1.52
Training **6**	−1	AB	0.06	2.64	5	6	483	82.26	
Training **7**	0.65	AB	0.44	3.37	3	6	375	60.97	−1.40
Training **8**	0.01	AB	0.82	4.50	5	6	452	63.32	
Training **9**	−0.38	AB	0.68	3.43	4	8	456	79.00	
Training **10**	1.2	B	0.78	3.13	4	5	499	25.51	
Training **11**	0.74	B	0.97	2.78	3	3	276	45.68	
Training **12**	1.7	B	1.00	4.58	2	6	263	6.89	−1.70
Training **13**	0.78	B	0.74	2.47	3	6	343	62.02	−0.40
Training **14**	0.18	B	0.58	2.78	4	2	278	27.83	−1.00
Training **15**	0.41	B	0.63	2.42	3	2	270	27.81	−0.74
Training **16**	0.04	n	0.01	3.15	3	0	291	24.98	−2.00
Training **17** (ZM-241385)	−1.3	n	0.00	2.82	4	3	337	109.10	
Training **18**	0.34	n	0.00	6.21	3	5	494	36.52	
Training **19**	0.89	n	0.00	5.82	3	5	472	26.76	
Training **20**	0	n	0.00	1.50	3	3	293	87.99	
Training **21**	−0.14	n	0.00	2.95	3	1	301	54.44	−2.00
Training **22**	0.74	n	0.00	4.68	2	8	295	24.49	−1.70
Training **23**	−0.44	n	0.01	4.79	5	7	486	57.64	
Training **24**	−0.66	n	0.00	3.74	4	3	486	50.45	
Training **25**	−0.3	n	0.00	3.23	4	1	356	66.59	−2.40
Training **26**	−1	n	0.00	2.50	3	7	368	98.35	
Test **1**	−0.7	A	0.49	2.59	4	5	414	76.40	
Test **2**	−0.68	A	0.24	3.15	5	5	454	75.54	
Test **3**	−0.47	A	0.25	2.08	4	4	385	70.91	
Test **4**	−0.77	A	0.35	1.36	4	5	415	79.24	
Test **5**	−0.1	A	0.16	3.67	4	4	400	75.49	
Test **6**	−0.1	A	0.40	3.96	4	3	360	52.62	−1.70
Test **7**	−0.41	A	0.38	3.31	4	4	384	68.16	
Test **8**	−0.85	AB	0.10	2.75	4	8	471	82.77	
Test **9** (Besonprodil)	−0.52	AB	0.09	3.20	4	6	402	56.83	
Test **10**	−0.11	AB	0.70	4.66	5	6	470	62.95	
Test **11**	0.09	AB	0.58	4.03	4	6	444	63.34	
Test **12**	−0.45	AB	0.49	3.59	4	8	474	78.64	
Test **13**	−0.15	AB	0.25	2.91	3	6	375	67.82	−1.30
Test **14**	−0.1	AB	0.26	2.91	3	6	375	68.15	−1.00
Test **15**	0	AB	0.25	3.41	3	6	389	67.82	−1.52
Test **16**	0.46	AB	0.83	4.93	4	7	454	62.58	
Test **17**	0.28	B	0.62	2.64	3	2	290	27.83	−0.74
Test **18**	0.99	B	0.41	3.48	3	3	335	25.26	−1.05
Test **19**	0.64	B	0.32	3.18	3	2	286	38.34	−0.70
Test **20**	−0.1	B	0.28	3.02	3	3	324	38.37	
Test **21**	0.15	B	0.29	2.61	3	2	270	38.41	
Test **22**	−0.22	B	0.30	1.41	3	3	318	36.62	−0.52
Test **23**	−0.1	B	0.32	1.94	3	3	332	36.69	−1.00
Test **24**	0.9	B	0.89	3.28	3	6	341	45.87	−0.70
Test **25**	0.9	B	0.96	3.86	3	4	294	46.67	−0.77
Test **26**	1.26	B	0.94	4.70	3	3	349	28.80	−2.00
Test **27**	1.26	B	0.88	3.79	3	3	322	42.66	−1.30
Test **28**	0.71	B	0.90	3.10	3	5	352	58.86	−0.77
Test **29**	1.38	B	0.90	3.50	3	5	352	52.10	−0.89
Test **30**	0.65	B	0.57	2.71	3	2	266	27.83	
Test **31**	0.49	B	0.77	2.47	3	6	343	62.02	−0.30
Test **32**	0.7	B	0.13	2.72	3	2	272	28.72	−1.00
Test **33**	0.46	B	0.56	2.99	3	2	286	27.83	−1.40
Test **34**	−0.22	B	0.10	2.50	3	5	340	62.19	−1.15
Test **35**	−0.12	n	0.00	2.64	4	3	335	74.74	−1.46
Test **36**	−0.15	n	0.00	2.88	3	7	400	81.04	−1.40
Test **37**	0.2	n	0.00	2.79	3	5	390	73.54	

(*continued on next page*)

Table 2 (*continued*)

Identifier[a]	log BPR	Charge class[b]	ANC 7.4[c]	clog *P*[d]	Nb (rings)	Nb (rotatable bonds)	MW	PSA [Å^2]	log CSFPR
Test **38**	0.11	n	0.01	4.43	4	3	404	59.53	
Test **39**	0.45	n	0.00	3.37	3	0	283	35.47	−1.70
Test **40**	−0.57	n	0.01	2.96	5	4	418	67.68	
Test **41**	0.18	n	0.00	4.37	3	2	319	29.79	
Test **42**	0.68	n	0.00	5.55	3	5	452	27.27	
Test **43**	0.3	n	0.00	4.06	4	1	361	56.01	−2.70
Test **44**	0.04	n	0.00	2.75	3	2	328	64.15	−1.52
Test **45**	0.46	n	0.00	4.37	3	2	319	30.16	−2.00
Test **46**	0	n	0.00	0.89	3	3	310	64.61	−1.00
Test **47**	0.6	n	0.00	3.21	3	3	362	48.44	
Test **48**	−0.1	n	0.00	2.27	4	0	302	71.41	
Test **49**	−0.4	n	0.01	3.09	4	2	307	66.13	
Test **50**	−0.19	n	0.00	4.35	4	4	499	48.07	
Test **51**	−0.8	n	0.00	2.33	4	3	321	75.59	−1.30

[a] Identifier according to training or test set for log BPR model.
[b] A: acid with ApKa < 8.9; B: base with BpKa > 6; AB: zwitterion with ApKa < 8.9 and BpKa > 6; n: neutral (less than 3% dissociation at pH 7.4).
[c] Absolute value of the net charge at pH 7.4.
[d] clog *P* v4.71 Daylight.

'shake flask' method [18,19]. In brief, a 0.5 mM DMSO solution of the test compound dispensed in aqueous buffer at pH 7.4 is analyzed by UV spectroscopy. The obtained optical density is equal to the concentration of the substance in the aqueous phase before partitioning. After addition of one volume of 1-octanol, the sample is incubated for 2 h at room temperature while shaking. The emulsion is allowed to settle overnight. When partition equilibrium is reached, the layers are separated by centrifugation at 3000 rpm for 10 min and the concentration of the substance in the aqueous phase is determined by UV absorption.

2.6. Determination of pKa values

The pKa values were determined by potentiometric titration using a Sirius GLpKa instrument (Sirius Analytical Instruments, Forest Row, UK). By this pH-metric method, a solution of the test compound in 0.1 M KNO_3 is titrated over a wide pH range by addition of 0.5 M KOH. pKa values are calculated by analyzing the shape of the titration curve. The titration is running under argon atmosphere to minimize absorption of atmospheric CO_2. This procedure results in a dataset containing a single potentiometric titration curve.

2.7. Calculated molecular descriptors

Partition coefficients were calculated using clogP v4.71 (Daylight Chemical Information Systems, Irvine, CA) and Kowwin v1.57 [20]. For three-dimensional descriptors, single molecular conformations were generated with Corina v2.4 [21]. Surface, volume, size and form descriptors as well as the numbers of hydrogen bond donors and acceptors were calculated with Moloc (Gerber Molecular Design, Amden, Switzerland). Numbers of rotatable bonds, different ring systems and possible internal hydrogen bonds were counted [22]. When counting rotatable bonds, conjugated single bonds (e.g. peptide bonds) were excluded. In addition, conformational restrictions were considered. Charge descriptors were determined considering the dissociation level of ionizable groups at a pH of 7.4. For calculation of the net charge (NC) at pH 7.4 the sum of the charged fraction of every ionizable group within the molecule was taken. Positive and negative charges were considered corresponding to basic and acidic groups.

Only measured pKa values were considered for compounds with ionizable groups. In summary 30 descriptors were used.

2.8. Data analysis

Multivariate data analysis (principle component analysis, PCA and partial least square regression, PLS) was performed with Simca-P+10 (Umetrics, Umea, Sweden). For training set selection Modde 7 was used (Umetrics). For log BPR modeling, a principal component analysis (PCA) with 30 descriptors was performed for the studied 77 compounds ($r^2 = 0.80$, $q^2 = 0.65$). The 30-dimensional space was thereby reduced to a 4-dimensional space. With the four score vectors, a d-optimal design was applied for maximizing the volume spanned in space. This corresponds to maximizing the information content of the chosen compounds. A training set of 26 compounds (including three center points) was retrieved. For the selection of a training set for a log CSFPR model the same procedure was used with 37 compounds. 18 compounds were compiled for the training set.

For the PLS models (log BPR, log CSFPR), all 30 descriptors were used in the beginning with the corresponding training set. The number of descriptors was reduced during model refinement to keep only the most descriptive parameters. Thus, redundant and less important descriptors were eliminated and not further considered. One selection criterion was thereby the variable influence on projection (VIP). Cross-validation was performed with default settings in Simca-P+10 (seven cross-validation groups). Over fitting was checked via randomly permuted log BPR and log CSFPR. The final validation of the model was performed with the remaining test set. Standard

error of estimate and F-value were determined with Statistica version 7.1 (StatSoft, Inc. (2005)).

3. Results and discussion

The distribution of molecules between the blood compartment and the central nervous system (CNS) is a complex process, which is the consequence of active and passive transport across two cellular barriers in the CNS. These are the blood–brain barrier (BBB), located at the endothelial cells of the brain vasculature, and the blood–liquor barrier (choroid plexus) [23]. One of the distinguishing properties of these barriers is the presence of high-resistance tight junctions, which seal the endothelial cells of the BBB as well as the epithelial cells of the choroid plexus. Paracellular transport across these barriers is therefore limited, leaving the transcellular route to the majority of compounds entering the CNS. Transcellular permeability depends thereby on many factors (Fig. 1) such as, for example, passive diffusion, active uptake and efflux, brain metabolism and binding to both plasma as well as tissue proteins. At present, *in silico* modeling of active transport is almost impossible due to the complexity of the situation and insufficient data [8]. We therefore focused in the present work on the prediction of CNS permeability by passive diffusion.

For the presented study, we could assemble a large *in vivo* dataset from a Roche proprietary database applying stringent quality criteria. All compounds are drug-like organic molecules from different chemical series (property ranges shown in Table 1). The dataset consists of bases, acids, zwitterions and neutral molecules. The calculated average Tanimoto coefficient for the 77 compounds is 0.46 ± 0.16 (average $\pm$ SD) which shows that the overall diversity of the dataset is given. All compounds are CNS active and are therefore assumed to penetrate the BBB and/or the blood–liquor barrier. Compounds with poor aqueous solubility (below 1 μg/mL) or reported instabilities were excluded from the dataset to reduce the risk of technical artefacts. Known substrates of drug transporting proteins were not considered. Table 2 provides information on molecular descriptors and brain permeation (log BPR) of 77 molecules used for this project. Information on cerebrospinal fluid permeation (log CSFPR) was available for a subset of 37 of these compounds. The log BPR and log CSFPR values were determined in animal experiments, where apparent steady-state conditions were reached (e.g. continuous infusion of test compound over 5 h in the rat).

Initially, we tried to analyze the log BPR dataset using published multiple linear regression models by Clark [13], which use as descriptors polar surface area (PSA) and lipophilicity (calculated log P). PSA and clog P calculation results are dependent on the underlying software packages. Therefore, the Clark model was adapted by using Moloc for PSA and Kowwin v1.57 for lipophilicity calculations. Using the 55 compounds of the original training set from Clark et al., a model was developed with a regression coefficient of $r^2 = 0.77$ and a StErrorEst of 0.37. We observed a good correlation between measured values of the Clark dataset and the prediction by the corresponding model proposed by Clark.

In the next step, this model was applied to the 77 in-house measured compounds from this study. As demonstrated in Fig. 2B, the result was disappointing in that the model was not predictive for the present dataset ($r^2 = 0.45$, StErrorEst = 0.47). Analysis of the residuals (Fig. 3) revealed that 52% of the compounds showed more than a ± 0.3 log unit difference

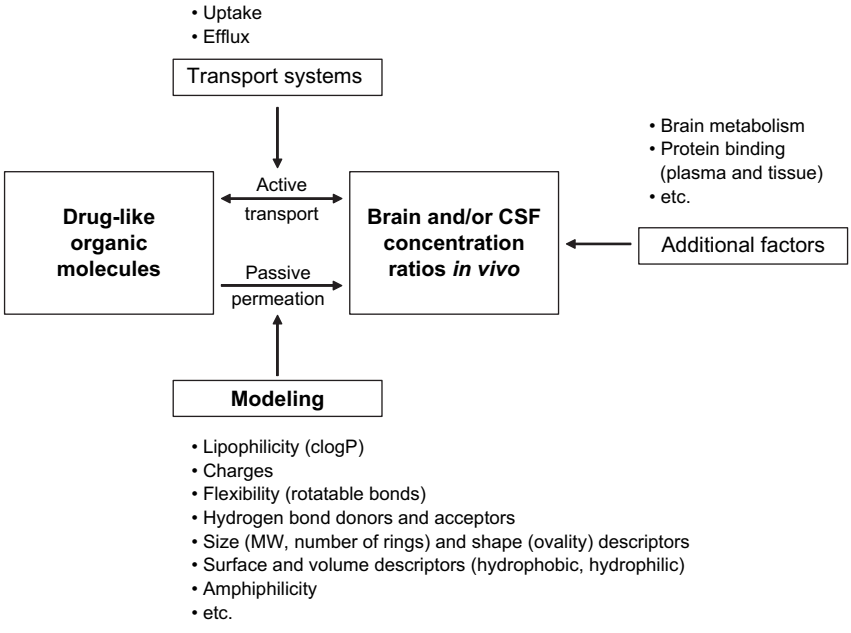


Fig. 1. Factors contributing to brain and CSF permeation of organic chemicals. Descriptive capability of modeling using physicochemical properties is at present limited to passive transport mechanisms. Active and passive drug transport into the central nervous system (CNS) are observed almost exclusively at two cellular barriers, the blood–brain barrier (BBB) and the blood–liquor barrier (choroid plexus).

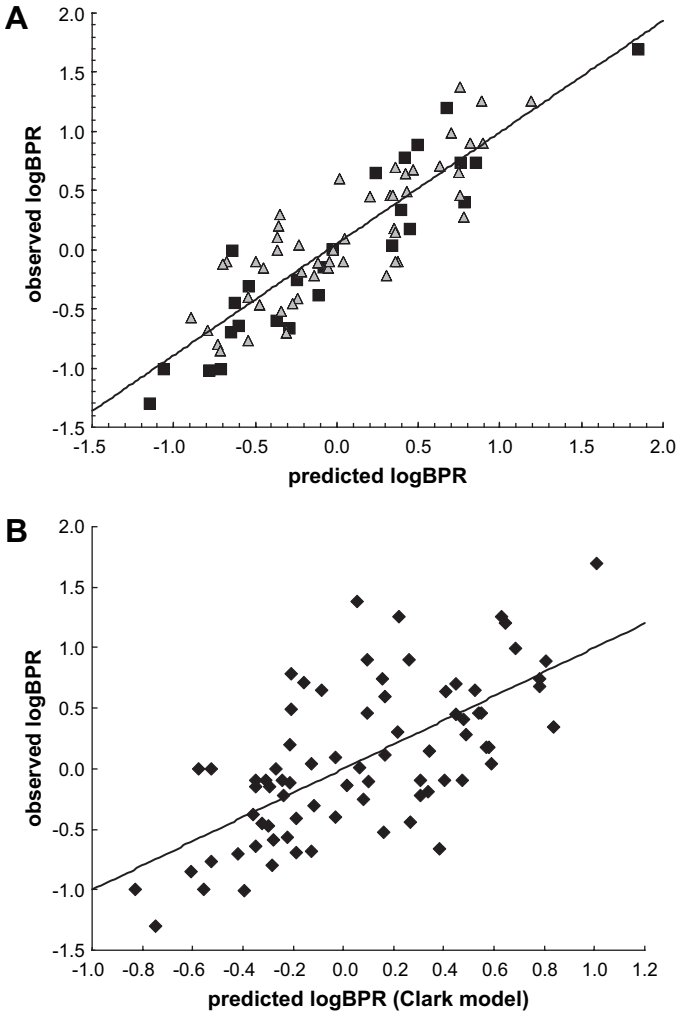


Fig. 2. Comparison between two different computational models to predict brain to plasma concentration ratios (log BPR). (A) Improved PLS regression model of the present study. Observed log BPR for 77 drug-like molecules (Table 2) against predicted log BPR ($r^2 = 0.78$, StErrorEst = 0.30, $F = 263.9$). Black symbols: training dataset. Grey triangles: test dataset (test set statistics: $n = 51$, $r^2 = 0.71$, StErrorEst = 0.30, $F = 119.5$). (B) Published Clark linear regression model. Observed log BPR for the same dataset of 77 drug-like molecules against predicted log BPR values ($r^2 = 0.45$, StErrorEst = 0.47).

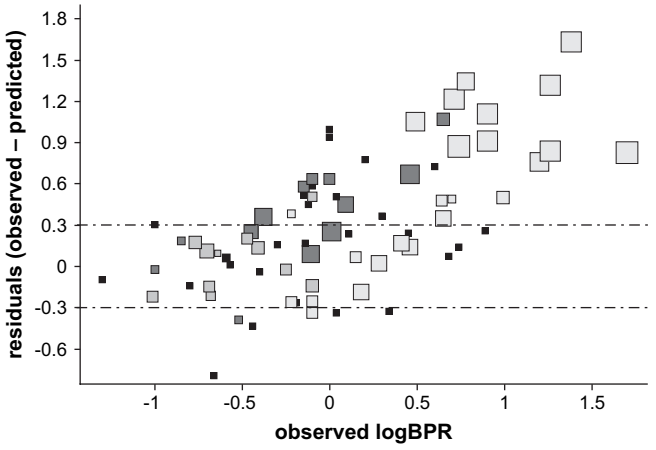


Fig. 3. Differences between observed and predicted log BPR. For the prediction the published Clark model was used after adaptation to in-house programs. Colors according to pKa properties: bases (BpKa > 6.0, more than 3% dissociation at pH 7.4) in white, acids (ApKa < 8.9, more than 3% dissociation at pH 7.4) in pale-gray, zwitterions (BpKa > 6.0 and ApKa < 8.9) in dark-gray, neutral compounds in black. The size reflects the absolute value of the net charge at pH 7.4 (values from 0 to 1).

between experimental and predicted values. Note that such a difference in the range of ±0.3 log unit can be considered to be still an acceptable error in view of experimental uncertainty in log BPR data [8,11]. In addition, an interesting trend is obvious. With increasing experimental log BPR the residuals are also increasing, which means that the general Clark model under-predicts the values for compounds with high brain-plasma ratios. In this area especially basic compounds are found. Within the 55 compounds, that were the basis for the Clark model, most of the molecules are basic or small and lipophilic. For this dataset the descriptors PSA and log *P* are sufficient, but they are obviously not sufficient for the description of a set including bases, acids, zwitterions and neutral compounds. Therefore, we decided to develop an extended log BPR model including additional descriptors for charge, other physicochemical and structural properties (examples are listed in Table 1). Using a d-optimal design, 26 out of the 77 compounds were selected for a training set. The calculated average Tanimoto coefficient based on MACCS fingerprints for the 26 compounds of the training set is 0.44 ± 0.16 (average ± SD).

For the training set, a PLS model with a regression coefficient r^2 of 0.86, StErrorEst of 0.29 and *F*-value of 149.0 was developed (Fig. 2A, black symbols). Cross-validation resulted in $q^2 = 0.80$. The four descriptors providing a basis for the obtained model are the absolute value of the net charge at pH 7.4 (ANC 7.4), lipophilicity (clog *P*), polar surface area (PSA) and size (reflected by the total number of aromatic and aliphatic ring systems (Nb(rings)). Although a coefficient of 0.014 for the PSA seems to be quite small, it should be noted that the coefficients displayed are unscaled. The PSA values are in a range of 7–110 Å^2, which shows that the influence on the predicted log BPR can be quite important.

$$\text{Predicted log BPR} = 1.397 + 0.853 \cdot \text{ANC 7.4} + 0.070 \cdot \text{clog } P - 0.014 \cdot \text{PSA} - 0.317 \cdot \text{Nb(rings)} \quad (1)$$

Using the model (Eq. (1)), the present log BPR dataset could be described ($r^2 = 0.78$, StErrorEst = 0.30). Sixty-eight percentage of the compounds were now within a ±0.3 log unit difference between experimental and predicted values.

Analysis of the model led to the following insights: First, the absolute value of the net charge at pH 7.4 (ANC 7.4) is an important descriptor. This can be explained by the fact, that hydrophobic interactions as well as electrostatic and hydrogen bonding components determine the interaction with phospholipid bilayers [24]. Molecules with a certain polarity have therefore advantages over pure lipophilic compounds with respect to membrane penetration and permeation. This positive influence of especially weakly basic groups for

partitioning of ionizing molecules between aqueous buffers and phospholipid membranes was discussed previously [25]. It is important to note that strong bases or acids are not contained within the present dataset of CNS active compounds (Table 1). Such compounds are likely to be inactive due to their very poor membrane partitioning according to the pH partition hypothesis described by Brodie and coworkers [26]. Second, the brain-plasma ratio decreases with increasing PSA (Fig. 4A). Compounds that cross the blood–brain barrier readily (log BPR > 0.3) are characterized by a PSA < 65 Å^2. Molecules above a threshold PSA of 90 Å^2 have a low probability to cross the BBB, which is in line with previous reports [8,27]. Third, high log BPR values (>0.3) can be reached with a balanced lipophilicity, i.e. clog *P* values between 2 and 6 (Fig. 4 B) or log *D* (pH 7.4) values between 1 and 3.5 (Fig. 4C).

Compounds with published brain-plasma ratios were collected to establish an external test dataset. The log BPR prediction by our model was not successful for several of these compounds. The encountered difficulties and our conclusions can be summarized as follows. First, some literature compounds (e.g. halothane, pentane, methane) represent small organic molecules, which are structurally very different compared to the drug-like molecules analyzed in this study. It is thus not recommended to use PLS regression models in an extrapolation setup outside the parameter space described by the initial model. Second, the published log BPR values were determined in different laboratories using different methods and experimental protocols. These differences may lead to inconsistencies within the dataset, which cannot be resolved. Third, the PLS model presented in this paper depends on molecular charges and therefore high-quality pKa values are required for accurate predictions. For many literature compounds, no or limited information on pKa values are available. Fourth, active drug transporters, such as P-glycoprotein, may interfere with brain uptake of their substrates. While the proprietary compounds used in the present study were confirmed to be no substrates of P-glycoprotein, corresponding information were not available for most compounds from the external dataset. Taken together, these findings indicate that the successful implementation of a predictive PLS regression model critically depends on the availability of a reliable and consistent experimental dataset.

It should be emphasized that computational models are in many cases not global and therefore not universally applicable. Such models should only be used within the limits defined by the chemical space of the used training dataset.

As expected, there is no apparent correlation between experimental log BPR values and experimental log CSFPR values (Fig. 5) [28,29]. In particular, the contribution of lipophilicity towards BPR and CSFPR seems to be very different. Comparison of two compounds with identical chemical scaffold, similar PSA and pKa values but different clog *P* reveals that a decreasing clog *P* increases CSFPR but decreases BPR. As a consequence, the established BPR prediction model cannot be used for the prediction of CSF exposure. We therefore decided to establish a PLS regression model for the

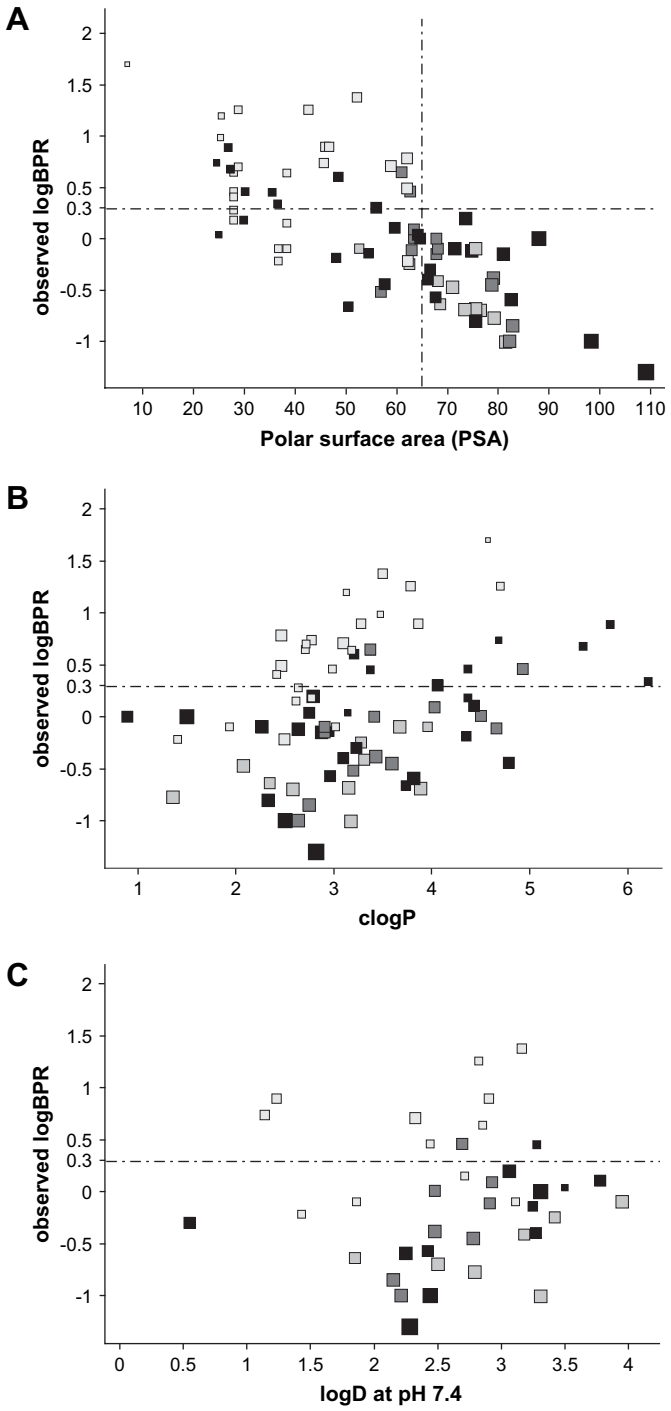


Fig. 4. Comparison between observed log BPR, polar surface area (PSA) and lipophilicity (clog *P* and log *D*). (A) Observed log BPR against polar surface area (PSA). (B) Observed log BPR against calculated clog *P* for 77 compounds. (C) Observed log BPR against experimentally determined log *D* (for 39 compounds at pH 7.4). Colors according to ionization: bases (BpKa > 6.0, more than 3% dissociation at pH 7.4) in white, acids (ApKa < 8.9, more than 3% dissociation at pH 7.4) in pale-gray, zwitterions (BpKa > 6.0 and ApKa < 8.9) in dark-gray, neutral compounds in black. The size reflects the PSA (values from 6.9 to 109.1 Å^2). Horizontal lines separate compounds that cross the BBB readily (log BPR > 0.3) from compounds that are only poorly distributed to the brain.

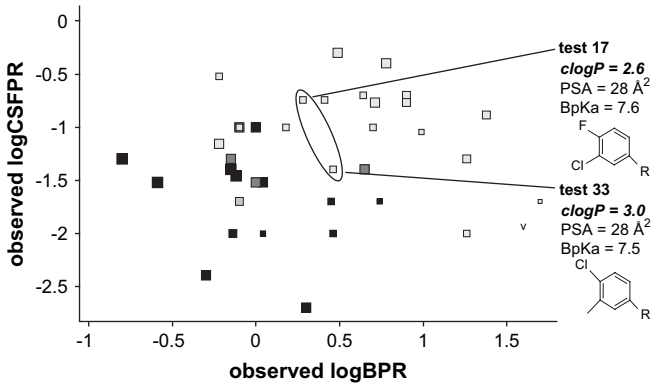


Fig. 5. Scatter plot of experimental brain to plasma ratios (log BPR) and CSF to plasma ratios (log CSFPR) for 37 drug-like organic molecules. Colors according to ionization: bases (BpKa > 6.0, more than 3% dissociation at pH 7.4) in white, acids (ApKa < 8.9, more than 3% dissociation at pH 7.4) in pale-gray, zwitterions (BpKa > 6.0 and ApKa < 8.9) in dark-gray, neutral compounds in black. The size reflects the PSA (values from 6.9 to 82.5 Å^2). Two molecules with an identical molecular scaffold R but a clog P difference of 0.4 are shown.

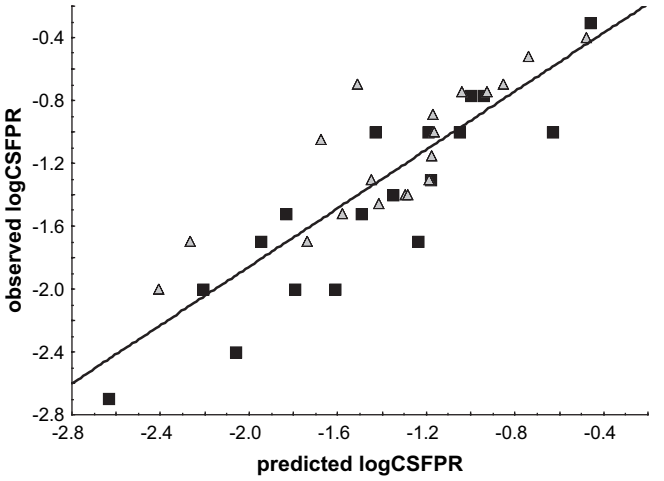


Fig. 6. Correlation between observed CSF to plasma ratios (log CSFPR) and predicted log CSFPR. Experimental log CSFPR data were obtained from 37 drug-like molecules. The PLS regression model for the prediction of log CSFPR was characterized by r^2 of 0.75, StErrorEst of 0.28 and F-value of 106.6. Black symbols: training dataset. Grey symbols: test dataset (test set statistics: $n = 19$, $r^2 = 0.71$, StErrorEst = 0.24, $F = 42.3$).

prediction of CSF partitioning of drugs. Based on the available log CSFPR dataset (Table 2), 18 compounds were selected as training set leading to a PLS model with a regression coefficient r^2 of 0.82 and a StErrorEst of 0.28 ($F = 70.6$). Cross-validation resulted in $q^2 = 0.69$. The four descriptors were the absolute value of the net charge at pH 7.4 (ANC 7.4), number of rotatable bonds, hydrophobic surface and lipophilicity (clog P) (Eq. (2)). Although a coefficient of 0.006 for the hydrophobic surface seems to be quite small, it should be noted that the coefficients displayed are unscaled. The hydrophobic surface values are in a range of 190–300 Å^2, which shows that the influence on the predicted log CSFPR can be quite important.

$$\begin{aligned}\text{Predicted log CSFPR} = {} & 0.817 + 0.831 \cdot \text{ANC 7.4} \\ & + 0.129 \cdot \text{Nb(rotatable bonds)} \\ & - 0.006 \cdot \text{hydrophobic surface} \\ & - 0.416 \cdot \text{clog } P \end{aligned} \quad (2)$$

Including the test set a correlation with $r^2 = 0.75$ and StErrorEst = 0.28 was obtained for the 37 compounds (Fig. 6).

The number of descriptors can be reduced with a slight decrease of the overall predictivity. Exclusion of the parameter hydrophobic surface resulted in a PLS model with $r^2 = 0.72$, $q^2 = 0.56$, StErrorEst = 0.34 and $F = 41.3$ for the training set. The model is described by Eq. (3).

$$\begin{aligned}\text{Predicted log CSFPR} = {} & -0.587 + 0.754 \cdot \text{ANC 7.4} \\ & + 0.103 \cdot \text{Nb(rotatable bonds)} \\ & - 0.458 \cdot \text{clog } P \end{aligned} \quad (3)$$

Including the test dataset $r^2 = 0.70$, StErrorEst = 0.31 and $F = 80.9$ were retrieved for the 37 compounds.

It was interesting to observe that net charge increases CSFPR whereas lipophilicity (hydrophobic surface and clog P in Eq. (2)) reduces the CSF exposure. The last point is different from the conditions for brain-plasma partitioning and can be explained by the fact that BPR reflects distribution within brain tissue including e.g. accumulation in membranes. Lipophilicity is one important factor that has a positive impact on tissue distribution. On the other hand CSFPR indicates partitioning into an aqueous fluid compartment which could be reduced when lipophilic properties outweigh polar features.

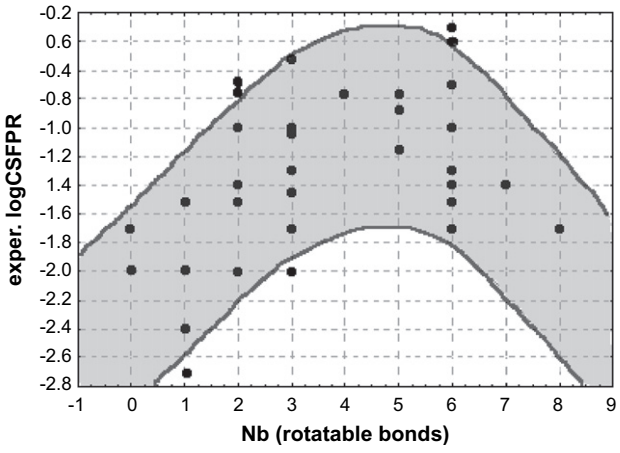


Fig. 7. Correlation between experimental log CSFPR values and the number of rotatable bonds. A parabolic relationship between these parameters is indicated (grey area within graph).

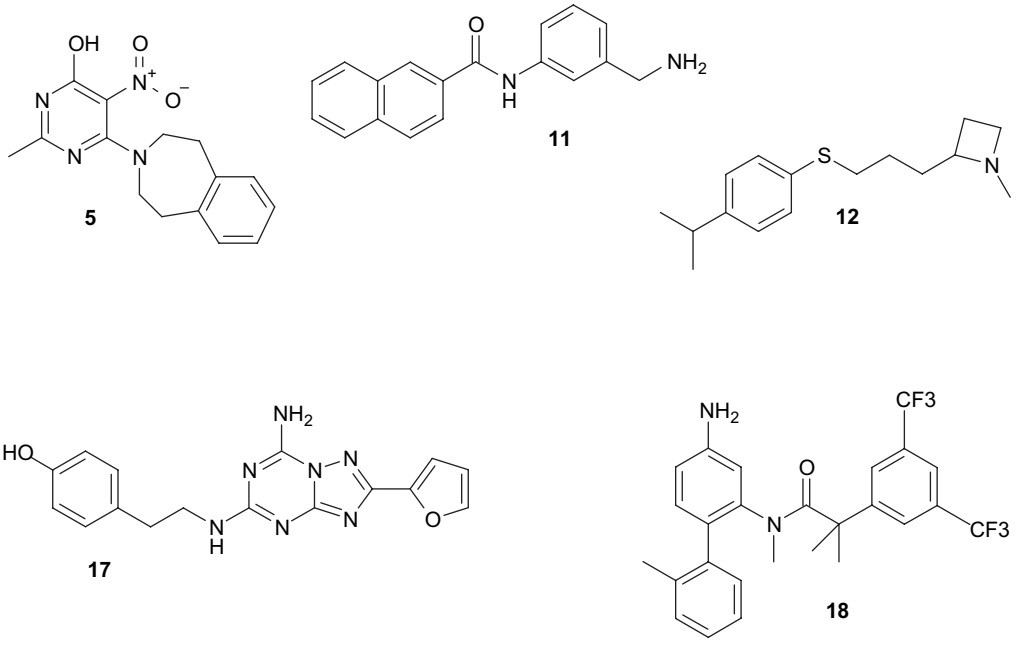
In the present PLS model, the number of rotatable bonds (and thus the flexibility of the molecule) has a beneficial influence on the CSF-plasma ratio. This observation is in line with previous BBB partitioning studies. Iyer et al. report a positive correlation between molecular flexibility and BBB permeability [30]. On the contrary, Veber and coworkers have shown a decrease in oral bioavailability with increasing solute molecular flexibility [31]. Iyer et al. discussed several explanations for these opposite findings. The group proposed that there might exist a parabolic relationship between log BPR and molecular flexibility [30]. Our dataset seems to support this thesis (Fig. 7). The log CSFPR increases up to three to four rotatable bonds, for more than five flexible bonds the CSF-plasma ratio is decreasing. However, statistically significant differences were not observed due to the small number of compounds with a high number of rotatable bonds in our dataset.

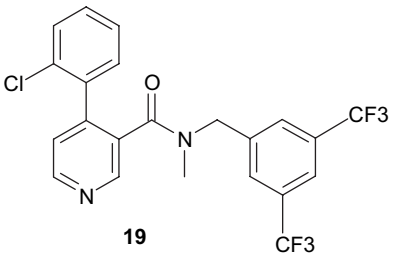
Permeation into the central nervous system is a prerequisite for the pharmacological action of CNS medicines. The molecular targets of these drugs are either accessible from within the brain tissue compartment or from the cerebrospinal fluid compartment. Therefore, investigations on compound distribution to the CNS as well as their partitioning into tissue compartments within the CNS are of central interest during the drug discovery and development process. Computational modeling, such as the strategies presented in this work, provides a cost-effective and efficient alternative to time-consuming and complex *in vitro* or *in vivo* experimentation. Present *in silico* tools, however, are still based on biophysical or physicochemical models of passive cellular permeability and do not yet consider additional factors such as catalyzed membrane transport. Future challenges for the development of the next generation of computational models will be an integration of the emerging knowledge on active transport processes in the CNS. Our progress will thereby build on the availability of robust experimental data, generated *in vitro* as well as *in vivo*. Considering the limitations of existing computational models, we recommend to interpret data from *in silico* models in the context of complementary experimental data from mechanistic *in vitro* or *in vivo* studies. Such a refinement and validation of computational models for a defined chemical space offers the possibility to explore and identify parameters responsible for brain and/or CSF permeation and to streamline and guide early synthetic efforts in medicinal chemistry.

Acknowledgements

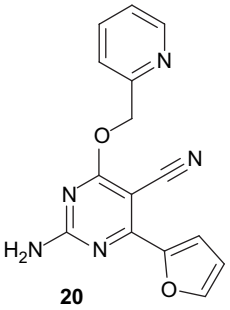
We thank Dr. Sonia Poli and Dr. Philippe Coassolo for their advice and support. For the determination of distribution coefficients and solubility values we would like to thank Pia Warga, Virginie Micallef and Isabelle Parrilla.

Appendix A. Eleven compounds from the training dataset [32]

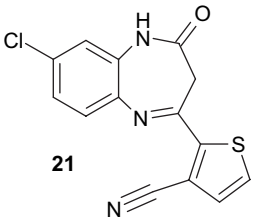




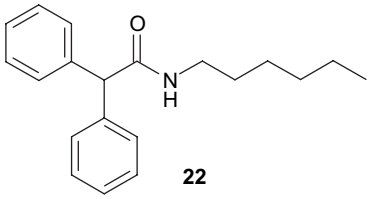
19



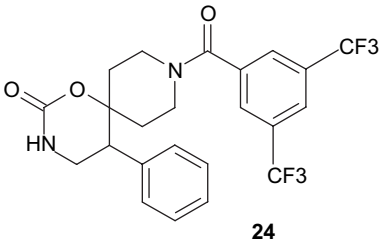
20



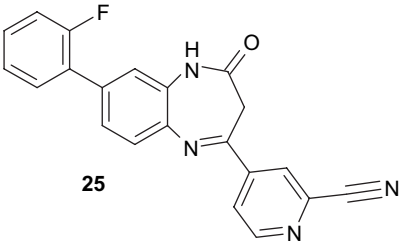
21



22

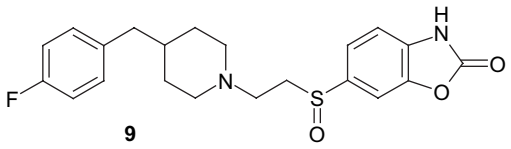


24

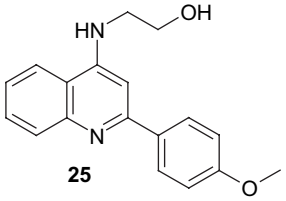


25

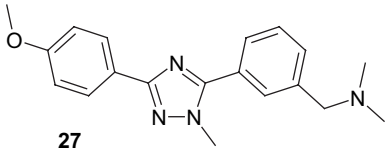
Appendix B. Fifteen compounds from the test dataset [33]



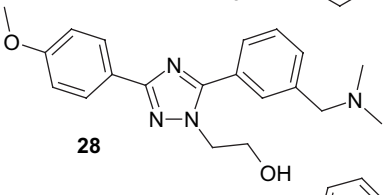
9



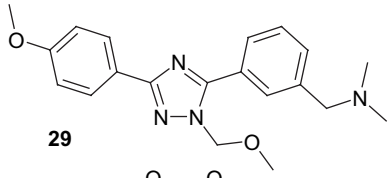
25



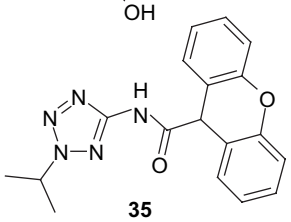
27



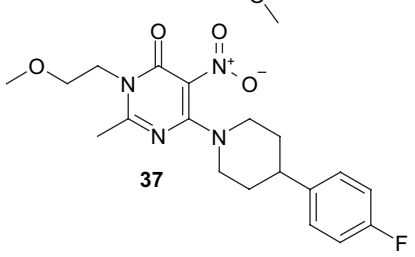
28



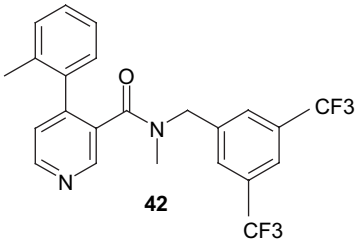
29



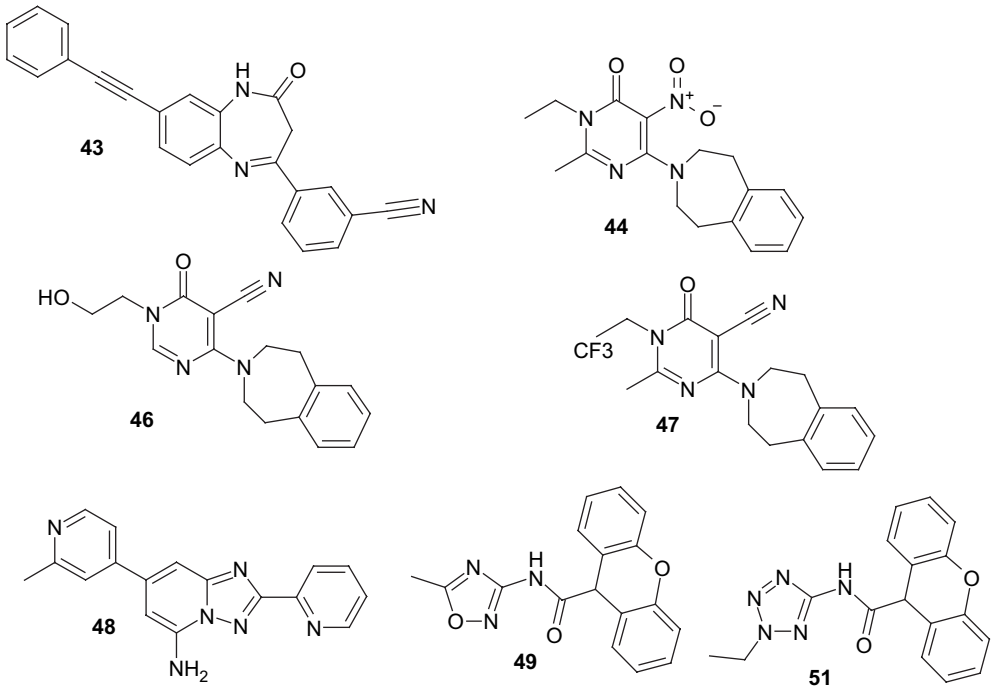
35



37



42



References

[1] M.A. Deli, C.S. Abraham, Y. Kataoka, M. Niwa, Cell. Mol. Neurobiol. 25 (2005) 59–127.
[2] M. Török, J. Huwyler, H. Gutmann, G. Fricker, J. Drewe, Exp. Brain Res. 221 (2003) 356–365.
[3] G.Z. Zheng, P. Bhatia, T. Kolasa, M. Patel, O.F. El Kouhen, R. Chang, M.E. Uchic, L. Miller, S. Baker, S.G. Lehto, P. Honore, J.M. Wetter, K.C. Marsh, R.B. Moreland, J.D. Brioni, A.O. Stewart, Bioorg. Med. Chem. Lett. (2006).
[4] T. Ooie, T. Terasaki, H. Suzuki, Y. Sugiyama, J. Pharmacol. Exp. Ther. 283 (1997) 293–304.
[5] D. Triguero, J. Buciak, W.M. Pardridge, J. Neurochem. 54 (1990) 1882–1888.
[6] A. Cerletti, J. Drewe, G. Fricker, A.N. Eberle, J. Huwyler, J. Drug Target. 8 (2000) 435–447.
[7] D.D. Shen, A.A. Artru, K.K. Adkison, Adv. Drug Deliv. Rev. 56 (2004) 1825–1857.
[8] D.E. Clark, Drug Discov. Today 8 (2003) 927–933.
[9] Y.N. Kaznessis, Curr. Med. Chem. Cent. Nerv. Syst. Agents 5 (2005) 185–191.
[10] J.T. Goodwin, D.E. Clark, J. Pharmacol. Exp. Ther. 315 (2005) 477–483.
[11] M. Lobell, L. Molnar, G.M. Keseru, J. Pharm. Sci. 92 (2003) 360–370.
[12] D. Schwab, H. Fischer, A. Tabatabaei, S. Poli, J. Huwyler, J. Med. Chem. 46 (2003) 1716–1725.
[13] D.E. Clark, J. Pharm. Sci. 88 (1999) 815–821.
[14] A.J. Grottick, G. Trube, W.A. Corrigall, J. Huwyler, P. Malherbe, R. Wyler, G.A. Higgins, J. Pharmacol. Exp. Ther. 294 (2000) 1112–1119.
[15] L.C. Hudson, C.S. Hughes, N.O. Bold–Fletcher, S.L. Vaden, Lab. Anim. Sci. 44 (1994) 358–361.
[16] Y.L. Huang, A. Säljö, A. Suneson, H.A. Hansson, Brain Res. Bull. 41 (1996) 273–279.
[17] T.M. Ballard, M.L. Woolley, E. Prinssen, J. Huwyler, R. Porter, W. Spooren, Psychopharmacology 179 (2005) 218–229.
[18] K. Takacs-Novak, A. Avdeef, J. Pharm. Biomed. Anal. 14 (1996) 1405–1413.
[19] K. Takacs-Novak, A. Avdeef, K.J. Box, B. Podanyi, G. Szasz, J. Pharm. Biomed. Anal. 12 (1994) 1369–1377.
[20] W.M. Meylan, P.H. Howard, J. Pharm. Sci. 84 (1995) 83–92.
[21] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, J. Comput. Aided Mol. Des. 19 (2005) 453–463.
[22] J. Zuegge, U. Fechner, O. Roche, N.J. Parrott, O. Engkvist, G. Schneider, Quant. Struct. Act. Relat. 21 (2002) 249–256.
[23] W.M. Pardridge, Adv. Drug Deliv. Rev. 15 (1995) 5–36.
[24] A. Avdeef, Curr. Top. Med. Chem. 1 (2001) 277–351.
[25] R.P. Austin, A.M. Davis, C.N. Manners, J. Pharm. Sci. 84 (1995) 1180–1183.
[26] P.A. Shore, B.B. Brodie, C.A. Hogben, J. Pharmacol. Exp. Ther. 119 (1957) 361–369.
[27] J. Kelder, P.D. Grootenhuis, D.M. Bayada, L.P. Delbressine, J.P. Ploemen, Pharm. Res. 16 (1999) 1514–1519.
[28] J.M. Collins, R.L. Dedrick, Am. J. Physiol. 245 (1983) R303–R310.
[29] E.C. de Lange, M. Danhof, Clin. Pharmacokinet. 41 (2002) 691–703.
[30] M. Iyer, R. Mishru, Y. Han, A.J. Hopfinger, Pharm. Res. 19 (2002) 1611–1621.
[31] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, J. Med. Chem. 45 (2002) 2615–2623.
[32] References for the molecules used as training dataset: (training **5**) G. Adam, A. Binggeli, H.P. Maerki, V. Mutel, M. Wilhelm, W. Wostl, Eur. Pat. Appl. EP 1074549 A2 20010207, 2001. (**17**) P.W. Caulkett, G. Jones, M.G. Collis, S.M. Poucher, Eur. Pat. Appl. EP 459702 A1 19911204, 1991. (**18**) M. Boes, G. Galley, T. Godel, T. Hoffmann, W. Hunkeler, P. Schnider, H. Stadler, PCT Int. Appl. WO 2000053572 A1 20000914, 2000. (**19**) M. Boes, Q. Branca, G. Galley, T. Godel, T. Hoffmann, W. Hunkeler, P. Schnider, H. Stadler, Ger. Offen. DE 10008042 A1 20000831, 2000. (**20**) E.M. Borroni, G. Huber-Trottmann, G.J. Kilpatrick, R.D. Norcross, PCT Int. Appl. WO 2001062233 A2 20010830, 2001. (**21**, **25**) G. Adam, A. Alanine, E. Goetschi, V. Mutel, T.J. Woltering, PCT Int. Appl. WO 2001029012 A2 20010426, 2001. (**22**) V. Mutel, E. Vieira, J. Wichmann, PCT Int. Appl. WO 2001027070 A1 20010419, 2001. (**24**) H.Y. Cai, M.P. Dillon, G. Galley, A. Goergler, S. Kolczewski, D.M. Muszynski-Barsy, PCT Int. Appl. WO 2002092604 A1 20021121, 2002.

[33] References for the molecules used as test dataset: (test **9**) J.L. Wright, S.R. Kesten, R.B. Upasani, N.C. Lan, PCT Int. Appl. WO 2000000197 A1 20000106, 2000. (**25**) A. Alanine, S. Burner, B. Buettelmann, M.P. Heitz Neidhart, G. Jaeschke, E. Pinard, R. Wyler, Eur. Pat. Appl. EP 1088818 A1 20010404, 2001. (**27**, **28**, **29**) A. Alanine, B. Buettelmann, M.P. Heitz Neidhart, G. Jaeschke, E. Pinard, R. Wyler, Eur. Pat. Appl. EP 1070708 A1 20010124, 2001. (**35**, **49**, **51**) E. Vieira, J. Huwyler, S. Jolidon, F. Knoflach, V. Mutel, J. Wichmann, Bioorg. Med. Chem. Lett. 15 (2005) 4628–4631. (**37**) A. Binggeli, H.P. Maerki, T. Masquelin, V. Mutel, M. Wilhelm, W. Wostl, PCT Int. Appl. WO 2002098864 A1 20021212, 2002. (**42**) T. Hoffmann, M. Boes, H. Stadler, P. Schnider, W. Hunkeler, T. Godel, G. Galley, T.M. Ballard, G.A. Higgins, S.M. Poli, A.J. Sleight, Bioorg. Med. Chem. Lett. 16 (2006) 1362–1365. (**43**) G. Adam, A. Alanine, E. Goetschi, V. Mutel, T.J. Woltering, PCT Int. Appl. WO 2001029012 A2 20010426, 2001. (**44**, **46**, **47**) G. Adam, A. Binggeli, H.P. Maerki, V. Mutel, M. Wilhelm, W. Wostl, Eur. Pat. Appl. EP 1074549 A2 20010207, 2001. (**48**) G. Huber-Trottmann, W. Hunkeler, R. Jakob-Roetne, G.J. Kilpatrick, M.H. Nettekoven, C. Riemer, PCT Int. Appl. WO 2001017999 A2 20010315, 2001.