

Correlation of blood–brain penetration using structural descriptors

Alan R. Katritzky,^{a,*} Minati Kuanar,^a Svetoslav Slavov,^{a,b} Dimitar A. Dobchev,^{a,b}
Dan C. Fara,^a Mati Karelson,^{b,c} William E. Acree, Jr.,^d
Vitaly P. Solov'ev^e and Alexandre Varnek^f

^aCenter for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, FL 32611, USA

^bDepartment of Chemistry, University of Tartu, Jakobi Street 2, Tartu 51014, Estonia

^cInstitute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia

^dDepartment of Chemistry, University of North Texas, Denton, TX 76203-5070, USA

^eInstitute of Physical Chemistry, Russ. Ac. Sci., Leninskiy prospect 31a, 119991 Moscow, Russia

^fFaculte de Chimie, 4, rue B. Pascal, Strasbourg 67000, France

Received 13 January 2006; revised 8 March 2006; accepted 11 March 2006

Available online 11 May 2006

Abstract—Experimental blood–brain partition coefficients (log BB) for a diverse set of 113 drug molecules are correlated with computed structural descriptors using CODESSA-PRO and ISIDA programs to give statistically significant QSAR models based respectively, on molecular and on fragment descriptors. The linear correlation CODESSA-PRO five-descriptor model has correlation coefficient $R^2 = 0.781$ and standard deviation $s^2 = 0.123$. The ‘consensus model’ of ISIDA gave $R^2 = 0.872$ and $s^2 = 0.047$. The developed models were successfully validated using the central nervous system activity data of an external test set of 40 drug molecules. © 2006 Elsevier Ltd. All rights reserved.

1. Introduction

Predicting properties such as absorption, distribution, metabolism or excretion (ADME) for large virtual libraries is of growing interest for pharmaceutical companies, in particular for those companies focusing on the discovery and development of central nervous system medicines where crossing the blood–brain barrier is a mandatory step for drug distribution. Thus, blood–brain barrier (BBB) permeability is an important ADME property and plays a key role in drug design. Numerous drug targets are located in the central nervous system (CNS) within the brain. The blood–brain barrier is a unique membranous barrier that tightly segregates the brain from the circulating blood.¹ The ability of a drug to penetrate the blood–brain barrier is of fundamental importance in drug design. Thus, the blood–brain partition coefficient (log BB) is a determining factor for the efficacy of central nervous system-acting drugs.

The experimental determination of blood–brain permeation is often complex and time consuming, requiring animal experiments and sometimes even the synthesis in radio-labeled form of the compounds to be tested. However, for modeling purposes, large data sets of BBB permeation can be derived from databases of known drugs by equating a compound's activity (CNS⁺) or inactivity (CNS[−]) against a CNS target with its brain permeation, with the assumption that CNS⁺/CNS[−] is due to compound's ability and/or inability to penetrate BBB. The necessity of a compound to cross the barrier is evident, although the mechanism of passage across the BBB can vary, depending on the compound; in particular, some compounds might be substrates for active transport mechanisms. Hence, non-permeation of the BBB is not the only reason for lack of observed CNS activity. For instance, compounds might be rapidly metabolized or removed from the brain by systems such as P-glycoprotein^{2,3} or might simply be inactive against the relevant molecular target in the brain.

Literature sources for BBB indicate that species differences are often ignored. Albert⁴ emphasized the importance of considering species differences for the in vitro modeling of BBB. The author pointed out that many publications compare in vivo BBB properties in one

Keywords: Blood–brain penetration; Central nervous system activity; Computational descriptors; Partition coefficients.

* Corresponding author. Tel.: +1 352 392 0554; fax: +1 352 392 9199; e-mail: katritzky@chem.ufl.edu

species (e.g., rat) with an in vitro model using cells from a different species (e.g., bovine or porcine) and then draw conclusions on the utility of the in vitro system to predict BBB drug permeability for the third species (e.g., human).⁴ Although the differences among species in drug metabolizing enzymes (e.g., P450 isoforms) have been well studied, the differences in BBB functions are yet to be clearly defined. The development of a reproducible and practical BBB model that can accurately predict the BBB permeability of drugs in humans is a challenging research area in the drive to enhance the efficiency of drug development.

The permeability (logPS) of the blood–brain barrier is an in vivo measure of BBB penetration. LogPS is measured using a short-duration vascular perfusion method, from which a permeability–surface area product is calculated. Gratton et al.⁵ have modelled the logPS data for 18 compounds with five solute descriptors: (i) R , excess molar refraction; (ii) π_{H}^2 , solute dipolarity or polarizability; (iii) α_{H}^2 ; (iv) β_{H}^2 , the hydrogen bond acidity/ basicity, and (v) V_{X} , the solute McGowan volume.

A common measure of the degree of BBB penetration is the ratio of the steady-state molar concentrations of the drug molecule in the brain and in the blood, expressed as BB (blood–brain distribution) (Eq. 1).

$$\text{BB} = C_{\text{brain}}/C_{\text{blood}} \quad (1)$$

Thus, logBB data are most commonly used for the BBB penetration and in silico prediction of BBB permeation.⁶ Over the years, data from diverse sources have been compiled and widely used in QSAR modeling. Recently, Abraham's research group compiled logBB data for 157 rat and human logBB values;⁷ some indirect experimental values from 22 literature sources were included, for example data points from Salminen et al.⁸ as plasma brain ratios. The logBB value of ibuprofen was obtained from a post mortem analysis.⁹ The authors⁷ had difficulties in modeling the compounds containing a carboxylic acid group, and thus used an indicator variable for carboxylic acid (I_1) together with the LFER descriptors to achieve correlation ($R^2 = 0.745$) for 148 compounds. So far, the logBB models derived have been limited to a maximum of approximately 150 data points (logBB values in human and rat),⁶ which is hardly representative of all chemicals, or even the smaller subset of drug-like molecules.

Over the past 15 years, many models for the prediction of BBB penetration have been reported and comprehensively reviewed.^{10–12,6}

Computational approaches have been employed and successfully applied by many research groups^{12–23} to predict the BBB penetration and ADME properties. These studies resulted in valuable virtual screening tools and enhanced the design of large library of drugs. Hence, elucidation of the features that differentiate CNS⁺ active and CNS[−] inactive drugs is found as a novel approach to predict the blood–brain barrier penetration.

In 1988, Young et al.²⁴ developed a physicochemical model for brain penetration while investigating histamine H_2 receptor antagonists. Initially the authors investigated a small data set (logBB in rat) of six compounds and found a good correlation Eq. 2 ($R^2 = 0.960$, $s^2 = 0.062$) with the partition parameter $\Delta \log P$.²⁵

$$\Delta \log P = \log P(\text{octanol/water}) - \log P(\text{cyclohexane/water}). \quad (2)$$

The same article²⁴ also reports an expanded data set of 20 compounds; here also the most significant best correlation ($R^2 = 0.691$) was achieved with the partition parameter $\Delta \log P$. Correlations with $\log P_{\text{octanol/water}}$ or $\log P_{\text{cyclohexane/water}}$ alone were 0.190 and 0.536, respectively. The authors suggested that the combination of the two quantities in $\Delta \log P$ might represent two different processes involved in the distribution of a drug between blood and brain. Partitioning into cyclohexane could reflect the tendency of a compound to occupy non-polar regions of the brain, while $\log P_{\text{octanol/water}}$ might indicate the amount of protein binding into the brain. The first computational approach to logBB prediction was that of Van de Waterbeemd and Kansy²⁶ who studied the same set of 20 compounds as Young et al.²⁴ and found a significant correlation ($R^2 = 0.697$) between logBB partitioning and the computed descriptors polar surface area (PSA) and molar volume (V_{M}). However, the practical use of this method seems to be limited in its present state since the standard deviation of the logBB prediction is rather high ($s = 0.45$ log units),²⁶ and the equation overestimated the blood–brain distribution for a test set of six compounds.²⁷

Development of a reliable predictive model generally requires a much wider data set. Abraham et al. analyzed in several papers^{28–30} the data set of Young et al.²⁴ (consisting of 30 rat logBB) and used this more extended data set to predict logBB values. Abraham et al.²⁸ extended the Young data set²⁴ by an additional 35 compounds, including small organic compounds and several gases. The logBB values of the additional compounds were calculated as the logarithm of the product of blood–air and brain–air partition coefficient values in human. Such indirect determinations, using in vitro data, would not necessarily yield the same results as in vivo measurements. In recent years a large number of data sets were compiled by several researchers to predict the blood–brain partition coefficients using computational methods. Many pharmaceutical industries, including Pfizer³¹ and Albany³² are also actively engaged in the development of computational models for the prediction of logBB values (see Table 1). Some important QSAR predictions are summarized in Table 1.

The main target of this work is to develop QSAR models using reliable data sets for the prediction of blood–brain partition coefficient (logBB) values using computational models. Two different multilinear regression (MLR) approaches based on: (i) molecular (CODES-SA-PRO)⁵² and (ii) molecular fragment (ISIDA)⁵³

Table 1. Summary of QSAR studies on blood–brain distribution

Descriptors used (<i>n</i>)	Training set	R^2	<i>s</i>	References
$\Delta \log P$	20	0.691	0.439	24
$\log P$	20	0.190	0.711	24
$\log P_{\text{cyc}}$	20	0.536	0.538	24
PSA, V_M	20	0.697	0.448	26
R_2 , π_2^H , $\alpha_2^H \beta_2^H$	22	0.89	0.270	28
R_2 , π_2^H , $\alpha_2^H \beta_2^H$	57	0.91	0.197	28
R_2 , π_2^H , $\alpha_2^H \beta_2^H$	65	0.75	0.397	28
E , S , A , B , V_X , I_1	148	0.745	0.343	7
ΔG_W^0	55	0.672	0.41	31
D, ONH, NQ4	60	0.776	0.377	32
$\log P$, M_m	20, 33	0.642, 0.897	0.486, 0.126	33
$\log P_{\text{cyh}}$, M_m	20	0.845	0.321	33
$\log P_{\text{cyh}}$, V_{wav}	20	0.889	0.271	33
K_{IAM} , I_2 , V_M	26, 21	0.581, 0.848	0.56, 0.27	8
K_{oct} , I_3 , V_M	26, 23	0.705, 0.848	0.47, 0.32	8
Molsurf (14), Clog P , HBAn	56, 28	0.862, 0.834	0.311, 0.312	34
HBAo, HBD	45, 70	0.72, 0.76		35
GRID (S_2_W)	20	0.723	0.001	36
18 (TIs + CDs)	58	0.850	0.318	37
PSA	45	0.841		16
PSA	57, 55	0.671, 0.707	0.455, 0.410	38
PSA + Clog P	55	0.787	0.354	
$N_{\text{acc,Solv}}$, log P , A_{pol}	61	0.730	0.424 (rms)	39
TPSA	45,57	0.78, 0.66		40
SASA (Monte Carlo)	76	0.984	0.173	41
G_{Solv} (GB/SA)	55	0.722	0.37	42
E-state (3)	102	0.66	0.45	43
5 Str. descriptors (GA)	59,72	0.757,0.785	0.41, 0.358	44
MI-QSAR (5)	56	0.845		45
Structural desc. (5)	48	0.837	0.26	22
log P , PSA, TIs	58	0.81		20
Atom type (3)	57	0.897	0.259	46
TOPS-MODE (2)	114, 81	0.697, 0.740	0.422	47
HAS (5)	47	0.781	0.375	48
V , Q_H , $Q_{O,N(\text{MRA})}$	56	0.817	0.330	48
V , Q_H , $Q_{O,N(\text{ANN})}$	56		0.236 (rmse)	49
Clog P , TPSA	150	0.69		50
E-state index, AlogP98, van der Waals surface area, Kappa shape index	88	0.746	0.392	51

descriptors are considered for the prediction of blood–brain distribution of organic molecules.

2. Results

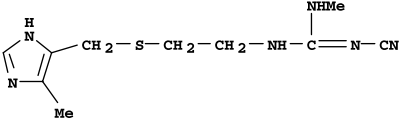
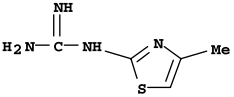
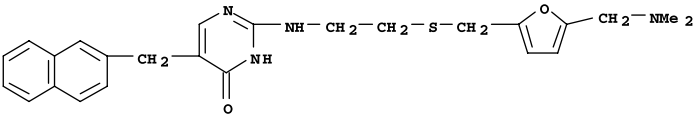
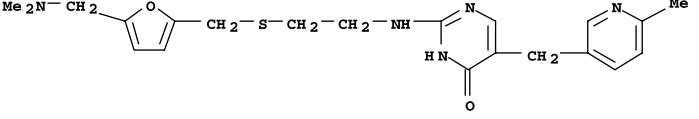
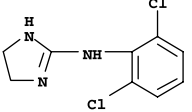
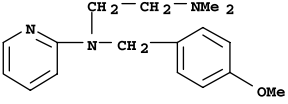
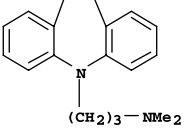
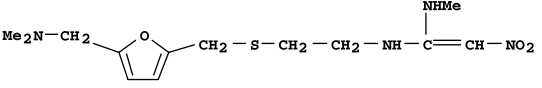
2.1. Data set

The logBB values studied by various authors, and the test set data, are listed in [Supplementary material \(SM1\)](#). In the present paper, we used the experimental values for blood–brain partition coefficient (logBB) values in rat for 127 compounds collected from several literature sources as listed in [Table 2](#). In building the database we tried to the extent possible to select data that had been determined by roughly the same experimental procedure. The drugs were administered intravenously and the animals sacrificed after a specified equilibration interval. The blood and brain samples were immediately harvested and frozen. The concentration of the drug in both the blood and whole brain was determined by standard analytical methods—usually

GC–MS, HPLC or scintillation counting in the case of radio-labeled drug samples. Hence, a preliminary analysis was made for the distribution of logBB data.

Many of the reported articles use experimental values of logBB together with indirect logBB values calculated using the simple algorithm of Abraham.²⁸ In the first part of our study, we paid attention to the experimental values of logBB in rat collected for 127 compounds from 15 different literature sources as shown in [Table 2](#). The data set consists of 30 values from Young et al.,²⁴ 21 from Salminen et al.,⁸ 20 from Kelder et al.,¹⁶ and 42 compiled by Platts et al.;⁷ values for the remaining 14 structures were collected from 11 different literature sources (see [Table 2](#)). The tabulated logBB values pertain to drugs or drug-like molecules except for hexane. In this set of 127 compounds, the experimental values for logBB range from –2.15 to 1.64; although most lie between –1.5 and 1.5, nine compounds (15, 16, 29, 30, 56, 61, 81, 116, and 123) deviate more widely. Since, we compiled the data from several different literature sources, we checked the normal

Table 2. Experimental (with references) and calculated logBB values in rat for 127 compounds

No.	Compound	Structure	logBB			References
			Exp	CODESSA ^a	ISIDA ^b	
1	Cimetidine (YM-1)		−1.42	−1.25	−1.27	24
2	YM-2		−0.04	−0.71	−0.54	24
3	SSKF 93619(YM-4)		−1.3	−1.01	−0.93	24
4	Lupitidine (YM-5)		−1.06	−1.64	−1.57	24
5	Clonidine (YM-6)		0.11	0.18	−0.05	24
6	Mepyramine (YM-7)		0.49	−0.17	0.21	24
7	Imipramine (YM-8)		0.83	0.85	0.91	24
8	Ranitidine(YM-9)		−1.23	−1.63	−0.88	24

(continued on next page)

Table 2 (continued)

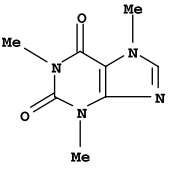
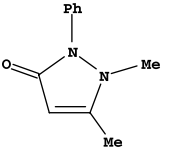
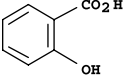
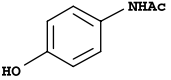
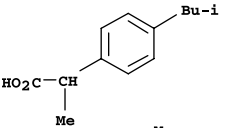
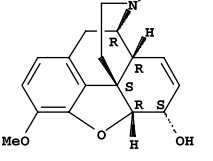
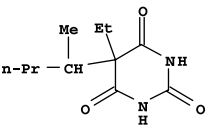
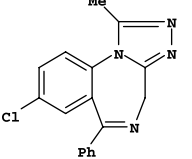
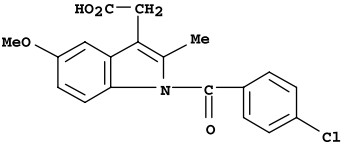
No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
9 ^c	Tiotidine (YM-10)		-0.82	-2.18	^d	24
10	SKF 94445 (YM-15)		-0.67	-0.82	-0.76	24
11	(YM-16)		-0.66	-0.91	-0.76	24
12	(YM-17)		-0.12	-0.72	-0.42	24
13	(YM-19)		-0.18	-0.44	-0.61	24
14	(YM-20)		-1.15	-1.08	-1.07	24
15	(YM-22)		-1.57	-1.14	-1.05	24
16	ICI 127032 (YM-23)		-1.54	-1.45	-1.54	24

17	(YM-24)		-1.12	-1.06	-0.85	24
18	(YM-25)		-0.73	-0.72	-0.56	24
19	(YM-26)		-0.27	-0.24	-0.49	24
20	(rntd ang) (YM-29)		-0.28	-0.50	-0.42	24
21	(rntd ang) (YM-30)		-0.46	-0.39	-0.18	24
22	SKF 94826 (YM-31)		-0.24	0.07	0.03	24
23	(rntd ang) (YM-34)		-0.02	-0.04	0.03	24
24 ^c	SKF 94674 (YM-36)		0.69	-0.12	^d	24
25	(rntd ang) (YM-37)		0.44	-0.02	0.07	24

(continued on next page)

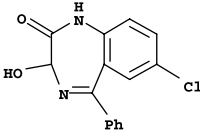
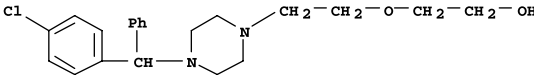
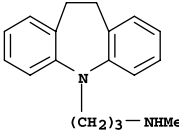
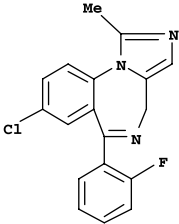
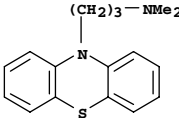
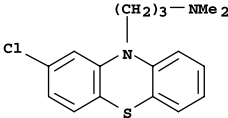
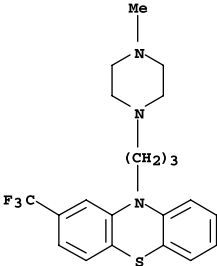
Table 2 (continued)

No.	Compound	Structure	logBB			References
			Exp	CODESSA ^a	ISIDA ^b	
26	Zolantidine (YM-41)		0.14	0.47	0.20	24
27	(YM-42)		0.22	0.56	−0.06	24
28	(YM-12)		−1.17	−1.02	−1.21	24
29	Icotidine (YM-3)		−2	−1.54	−1.68	24
30 ^c	(cmt d ang) (YM-13)		−2.15	−1.07	^d	24
31	Aspirin		−0.5	−0.27	−0.38	8
32	Valproic acid		−0.22	−0.07	0.05	8
33	Theophylline		−0.29	−0.30	−0.31	8

34	Caffeine		-0.055	-0.07	-0.11	8
35	Antipyrine		-0.097	-0.07	-0.15	8
36 ^c	Salicylic acid		-1.1	0.31	^d	8
37	Acetaminophen		-0.31	-0.19	-0.41	8
38 ^c	Ibuprofen		-0.18	0.78	^d	8
39	Codeine		0.55	0.32	0.01	8
40	Pentobarbital		0.12	-0.36	-0.17	8
41	Alprazolam		0.044	0.13	0.17	8
42 ^c	Indomethacin		-1.26	0.00	^d	8

(continued on next page)

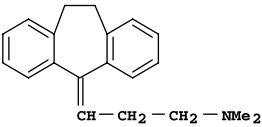
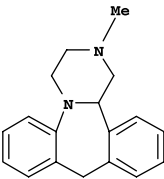
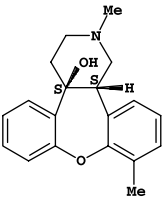
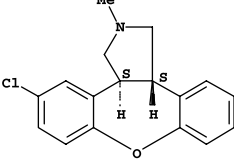
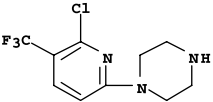
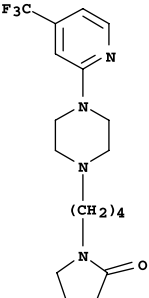
Table 2 (continued)

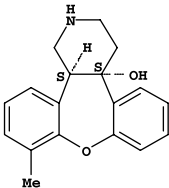
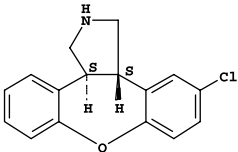
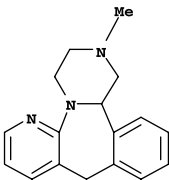
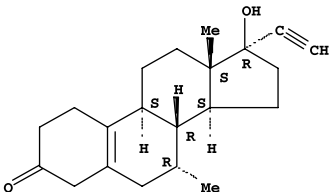
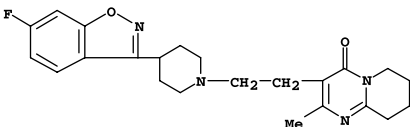
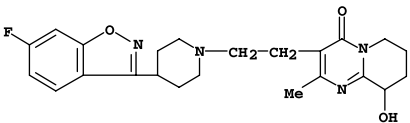
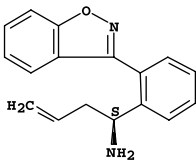
No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
43 ^c	Oxazepam		0.61	−0.33	^d	8
44	Hydroxyzine		0.39	−0.11	0.37	8
45	Desipramine		1.2	0.72	0.74	8
46	Midazolam		0.36	0.41	0.25	8
47	Promazine		1.23	0.87	0.82	8
48	Chlorpromazine		1.06	1.05	1.01	8
49	Trifluoroperazine		1.44	1.37	1.10	8

50	Thioridazine		0.24	1.01	0.58	8
51	Verapamil		−0.7	−0.29	−0.49	8
52	Carbamazepine		0	0.12	0.16	31
53	Carbamazepine 10,11-epoxide		−0.337	−0.15	−0.21	31
54	L663581		−0.301	−0.12	−0.60	31
55	M1L663-581		−1.337	−0.65	−1.14	31
56	M2L663-581		−1.824	−1.19	−1.37	31

(continued on next page)

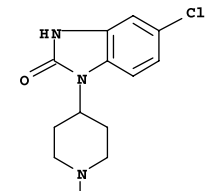
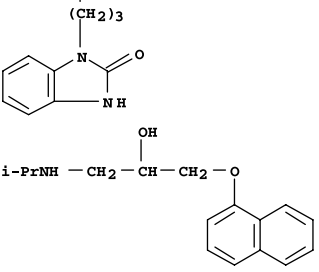
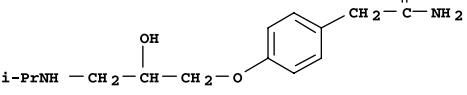
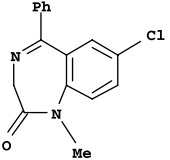
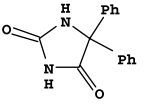
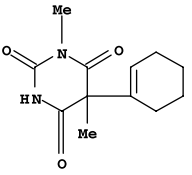
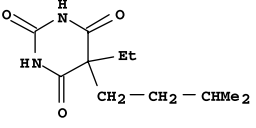
Table 2 (continued)

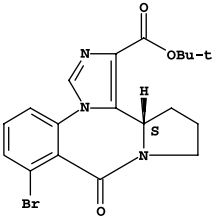
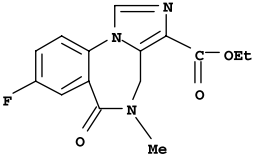
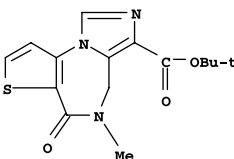
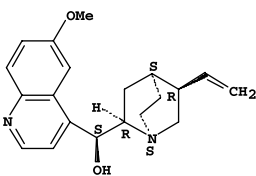
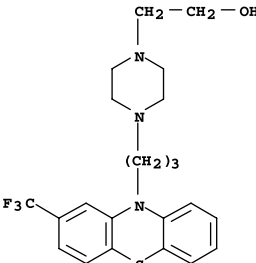
No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
57	Amitriptyline	 <chem>CN(C)CCC=C1C2=CC=CC=C2C3=CC=CC=C13</chem>	0.98	0.70	1.11	16
58	Mianserin	 <chem>CN1CC2C(C1)C3=CC=CC=C3C4=CC=CC=C24</chem>	0.99	0.81	0.97	16
59	Org 4428	 <chem>CN1CC2C(C1)C3=CC=CC=C3C4=CC=CC=C24O</chem>	0.82	0.88	0.52	16
60	Org 5222	 <chem>CN1CC2C(C1)C3=CC=CC=C3C4=CC=CC=C24Cl</chem>	1.03	0.70	0.86	16
61 ^c	Org 12962	 <chem>C1=CC=C(C=C1N2CCNCC2)C3=CC=CC=C3C(F)(F)F</chem>	1.64	0.43	^d	16
62	Org 13011	 <chem>C1CCN(C1)CCCC2CCN2</chem>	0.16	−0.21	0.08	16

63	Org 32104		0.52	0.69	0.27	16
64	Org 30526		0.39	0.87	0.57	16
65	Mirtazepine		0.53	0.38	0.57	16
66	Tibolone		0.4	0.35	0.71	16
67	Risperidone		-0.02	-0.18	-0.24	16
68	9-Hydroxyrisperidone		-0.67	-0.98	-0.70	16
69 ^c	Org 34167		0	0.18	-0.02	16

(continued on next page)

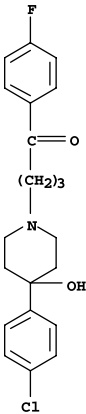
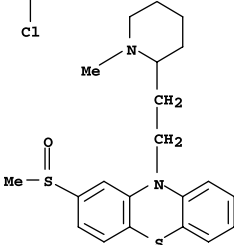
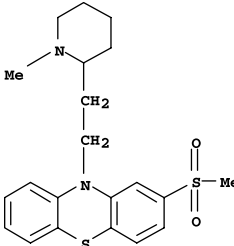
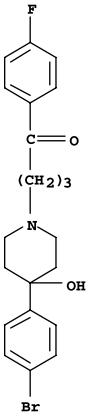
Table 2 (continued)

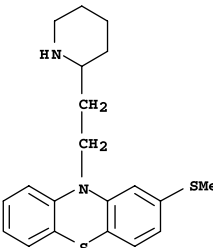
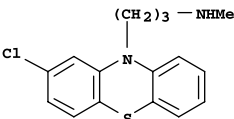
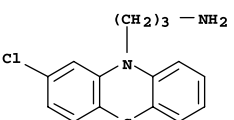
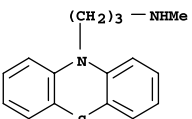
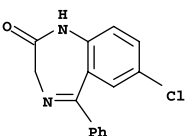
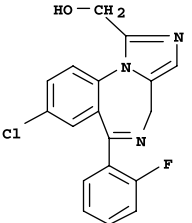
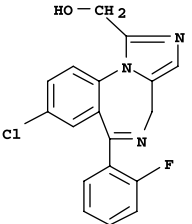
No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
70	Domiperidone		−0.78	−0.08	−0.39	16
71	Propranolol		0.64	0.19	0.16	7
72	Atenolol		−1.42	−1.37	−0.82	7
73	Diazepam		0.52	0.39	0.48	7
74	Phenytoin		−0.04	0.04	−0.09	7
75	Hexobarbital		0.1	−0.16	−0.07	7
76	Amobarbital		0.04	−0.35	−0.26	7

77 ^c	Bretazenil		-0.09	0.19	-0.18	7
78	Flumazenil		-0.29	-0.11	-0.34	7
79	Ro 19-4603		-0.25	-0.31	-0.32	7
80 ^c	Quinidine		-0.46	-0.07	-0.25	7
81 ^c	Fluphenazine		1.51	0.29	^d	7

(continued on next page)

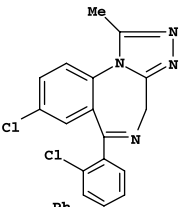
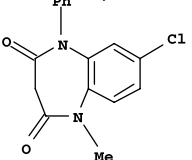
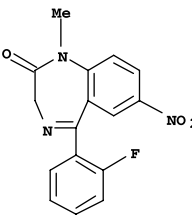
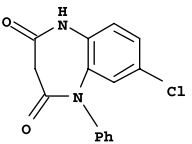
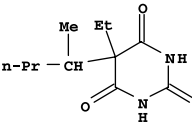
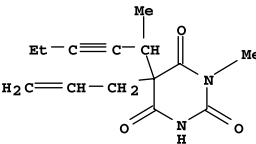
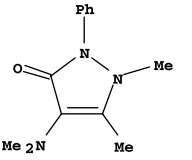
Table 2 (continued)

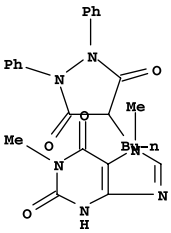
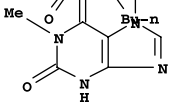
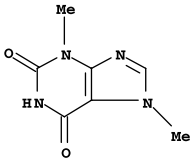
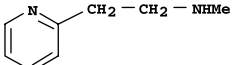
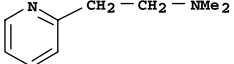
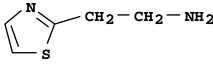
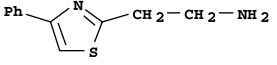
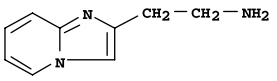
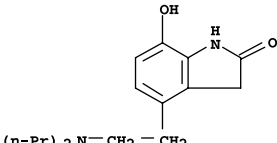
No.	Compound	Structure	logBB			References
			Exp	CODESSA ^a	ISIDA ^b	
82 ^c	Haloperidol		1.34	0.19	^d	7
83 ^c	Mesoridazine		-0.36	0.54	^d	7
84	Sulforidazine		0.18	0.44	0.58	7
85 ^c	Bromoperidol		1.38	-0.04	^d	7

86	Northioridazine		0.75	0.93	0.46	7
87 ^c	Norchlorpromazine		1.37	0.55	^d	7
88	Nor2chlorpromazine		0.97	0.59	0.78	7
89	Desmonomethylpromazine		0.59	0.72	0.76	7
90	Desmethyldiazepam		0.5	0.23	0.27	7
91	1-Hydroxymidazolam		-0.07	-0.18	0.11	7
92	4-Hydroxymidazolam		-0.3	0.02	0.03	7

(continued on next page)

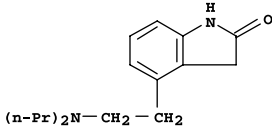
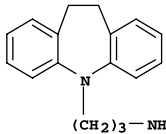
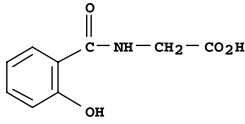
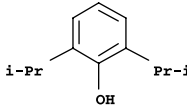
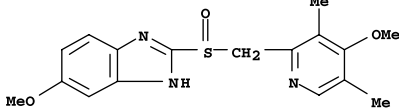
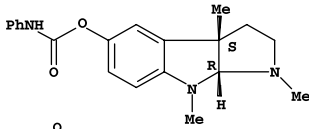
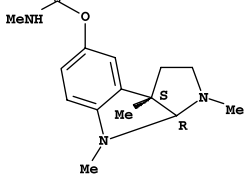
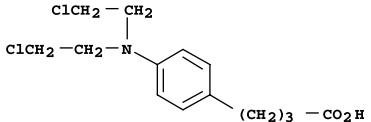
Table 2 (continued)

No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
93	Triazolam		0.74	0.36	0.30	7
94	Clobazam		0.35	0.26	0.43	7
95	Flunitrazepam		0.06	0.01	0.05	7
96	Desmethyclobazam		0.36	0.05	0.22	7
97	Thiopental		−0.14	−0.23	−0.07	7
98	Methohexital		−0.06	−0.44	0.01	7
99	Aminopyrine		0	0.04	−0.04	7

100	Phenylbutazone		−0.52	0.16	−0.18	7
101	Paraxanthine		0.06	−0.39	−0.14	7
102	Theobromine		−0.28	−0.47	−0.21	7
103	Betahistine (<i>N</i> -methyl-2-pyridineethanamine) YG-14		−0.3	−0.21	−0.21	7
104	<i>N,N</i> -Dimethyl-2-pyridineethanamine (YG-15)		−0.06	−0.02	−0.04	7
105	2-Thiazolylethylamine (YG-16)		−0.42	−0.49	−0.18	7
106 ^c	2-Thiazoleethanamine, 4-phenyl-(9CI)YG-19		−1.3	−0.18	^d	7
107	2-(2-Aminoethyl)imidazo[1,2- <i>a</i>]pyridine, Y-G20,SKF 71473		−1.4	−0.61	−0.47	7
108	SKF 89124		−0.43	−0.39	−0.18	7

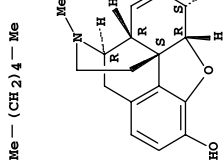
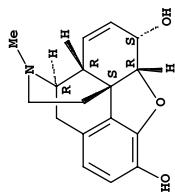
(continued on next page)

Table 2 (continued)

No.	Compound	Structure	log BB			References
			Exp	CODESSA ^a	ISIDA ^b	
109	Ropinirole (SKF 101468)		0.25	−0.06	0.10	7
110	N-Desmethyldesipramine		1.06	0.50	0.70	7
111	Salicyluric acid		−0.44	−0.73	−0.69	7
112	Propofol		0.48	1.05	0.47	55
113	Omeprazole		−0.82	−0.45	−0.81	56
114	Phenserine		1	0.77	0.54	9
115	Physostigmine		0.079	0.35	0.35	12
116	Chloroambucil		−1.7	−1.04	^d	43

117	BCNU		-0.52	-0.61	-0.23	12
118	Thiopramide		-0.161	-0.23	-0.24	12
119	Phenyl <i>N-tert</i> -butylnitrone (PBN)		0.17	0.49	0.08	57
120	α -(4-Pyridyl 1-oxide)- <i>N-tert</i> -butylnitrone (POBN)		-0.15	-0.46	-0.17	57
121	SB-222200		0.3	0.60	0.29	12
122	Terbutylchloroambucil		1	0.10	0.65	12
123	Temelastine		-1.88	-1.31	-1.42	28,58,16
124	Cotinine		-0.32	0.16	-0.21	59
125	1-Butyl-3-phenylthiourea		0.037	-0.06	0.04	60

Table 2 (continued)

No.	Compound	Structure	logBB		References
			Exp	CODESSA ^a ISIDA ^b	
126	Hexane	Me-(CH ₂) ₄ -Me 	0.42	0.43	0.15 28,61
127	Morphine		-0.16	-0.24	0.01 7

NB: the structure names are noted as YM1–YM30, taken from Young et al. [88JMC656-24].

^aThe logBB values are calculated using model Table 5.

^bThe logBB values are calculated using the 'consensus model'.

^cOutliers excluded from the model by CODESSA-PRO.

^dOutliers excluded from the model by ISIDA.

distribution of the data by using STATISTICA⁵⁴ prior to building the models. The plot of number of observations versus the logBB values in Figure 1 shows that the data are normally distributed.

2.2. Methodology

2.2.1. Computational methods.

2.2.1.1. CODESSA approach. CODESSA (comprehensive descriptors for structural and statistical analysis)-PRO⁵² is a program for developing QSAR/QSPR models. CODESSA-PRO includes diverse statistical structure–property correlation techniques that can be used for the analysis in combination with the calculated molecular descriptors. In particular, various algorithms based on stepwise statistical multilinear regression analysis are applicable for searching 'the best' multiparameter correlation in the large spaces of the natural molecular descriptors. Since only theoretically calculated descriptors are used in the resulting multiparameter correlation equations, the value of the property of interest can be predicted for compounds not yet synthesized. CODESSA-PRO methodology has shown promising results in correlations and predictions of physicochemical and biological properties of diverse sets of organic compounds. We successfully correlated aqueous solubility,⁶² partitioning of organic molecules in aqueous biphasic system,⁶³ and water–air partition coefficients⁶⁴ of structurally diverse compounds. Correlations⁶⁵ of rat blood–air, saline–air, and olive oil–air partition coefficients of 100 diverse organic compounds with molecular descriptors followed by tests of the predictive power of these models using 33 compounds not included in the training set gave fitted squared correlation coefficient values of 0.791, 0.794, and 0.846, respectively. Using molecular descriptors solely calculated from structural characteristics our group also developed QSAR models for the genotoxicity of heteroaromatic and aromatic amines,⁶⁶ aquatic toxicity of environmental pollutants,⁶⁷ β -cyclodextrin complexation free energies,⁶⁸ antibacterial activity of 3-aryloxazolidin-2-one,⁶⁹ HIV-1 protease inhibitory activity of substituted tetrahydropyrimidinones,⁷⁰ blood and tissue air partition coefficients of organic solutes,⁷¹ prediction of partition of drugs in human milk and plasma,⁷² and antimalarial activity of drugs.⁷³ The use of CODESSA-PRO methodology in medicinal chemistry has been reviewed elsewhere.⁷⁴

2.2.1.2. CODESSA-PRO calculations. The chemical structures were drawn by using ISIS Draw 2.4⁷⁵ and pre-optimized using molecular mechanics force field method included in Hyperchem 7.0.⁷⁶ Final optimization was performed with MOPAC 7.0⁷⁷ (implemented in CODESSA-PRO) using the AM1 semiempirical method.⁷⁸ Molecular descriptors of different kinds like constitutional, topological, geometrical, charge-related, semiempirical, and thermodynamic were calculated using CODESSA-PRO program.⁵² The gradient norm of 0.01 kcal/Å was applied to the geometry optimization.

The best multilinear regression (BMLR) procedure^{79,80} available in the framework of the CODESSA-PRO was used to find the best correlation models from

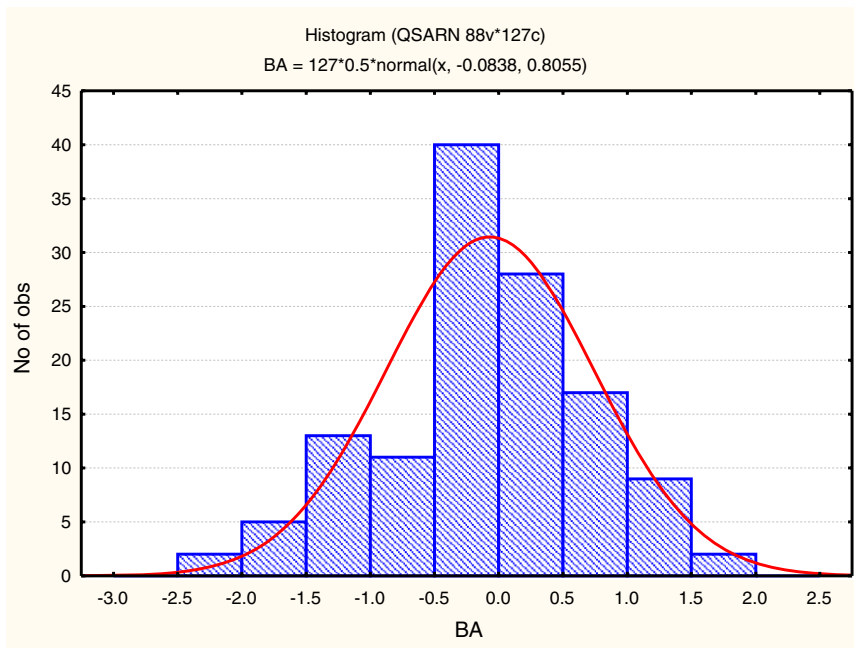


Figure 1. Histogram of the data distribution of the experimental logBB values.

selected non-collinear descriptors. The BMLR selects the best two-parameter regression equations, the best three-parameter regression equations, etc., on the basis of the highest R^2 value in the stepwise regression procedure. During the BMLR procedure, the descriptor scales are normalized and centered automatically, and the final result is given in natural scales. The result obtained by BMLR is the best representation of the property in the given descriptors' pool.

To develop QSAR/QSPR models, it is important to decide when to stop the addition of descriptors during the stepwise regression procedure. An excessive number of descriptors lead to over-correlated equations that are difficult to interpret in terms of interactions and mechanisms. A simple procedure to control the proliferation of descriptor is the 'break point'. From the analyses of the plot of number of descriptors involved (n) versus the squared correlation coefficient (R^2), and the cross-validated square correlation coefficient (R_{CV}^2), using the values corresponding to the property models, it appears that the statistical improvement of the model is higher (steeper ascent of the relationship) until one point (the 'break point') and after that the improvement is negligible. Consequently, the model corresponding to the break point shows the optimum number of descriptors to be used in modeling that property.

2.2.2. ISIDA approach. In silico design and data analysis (ISIDA) program realizes substructural molecular fragments (SMF) method, where substructural fragments (sequences of atoms and bonds or augmented atoms) are used as descriptors in QSPR studies.^{68,81–87} The modeled property can be presented as a linear or non-linear combination of selected fragment descriptors. Besides the QSAR/QSPR module, ISIDA also includes

clustering and combinatorial modules as well as some supplementary tools including the editor of 2D structures EdChemS and the editor of SD files EdiSDF.⁵³

2.2.2.1. ISIDA calculations. The substructural molecular fragments (SMF) method is based on the representation of a molecular graph as a superposition of fragments (subgraphs) and on the calculation of their contributions to a given property Y . Two different classes of fragments are used: 'sequences' and 'augmented atoms'. The sequences may contain atoms and bonds, atoms only, or bonds only. Only shortest paths from one atom to the other are used. For each type of sequence, the minimal (n_{\min}) and maximal (n_{\max}) number of constituent atoms is defined. In the current version of ISIDA, $n_{\min} \geq 2$ and $n_{\max} \leq 15$, thus the total number of all possible types of sequences containing from n_{\min} to n_{\max} atoms and including either atoms, or bonds or atoms together with bonds is 315. An 'augmented atom' represents a selected atom with its environment including either neighboring atoms and bonds, or atoms only, or bonds only. Atomic hybridization can be taken into account. The total number of the types of augmented atoms is four.

At the training stage, ISIDA builds up to 1300 structure–property models involving two linear (with and without fragment independent term) and two non-linear fitting equations and 319 different types of fragmentation. Usually not one but several models provide high internal cross-validation parameters. Therefore, instead of using one particular model, we apply simultaneously M best models including the values of internal cross-validation correlation coefficient $R_{CV}^2 \geq R_{CV,lim}^2$, where $R_{CV,lim}^2$ is a user, defined threshold. Thus, for each compound, the program consensus computes the property as

arithmetic mean of values obtained with all *M* models excluding outlying values according to Grubbs's test.⁸⁸ Our experience^{83–85} podands⁸⁵ as well as previously reported data^{89,90} show that such an ensemble modeling smoothes inaccuracies of particular individual models, thus improving the robustness of predictions.

Backward stepwise variable selection procedure has been used to select pertinent fragment descriptors. This procedure is based on the elimination of variables with low values of *t* statistic criterion $t_i = a_i/s_i$, where a_i and s_i are the coefficient at *i* variable and its standard deviation, respectively.⁸⁶ First, the program selects the variable with the smallest $t < t_0$, then it performs a new fitting excluding that variable. This procedure is repeated until $t \geq t_0$ for selected variables or if the number of variables reaches the user's defined value. Here, tabulated value of Student's t_0 criterion is a function of the number of data points, the number of variables, and the significance level.

The EdChemS editor of 2D structures has been used to modify the bond types in the aromatic fragments of compounds originally presented as Kekule structures. A special program utility was prepared to convert the CODESSA-PRO data presented as the *.MOL (MDL) and text *.PRP files into the MDL SDF file.⁹¹ The hydrogen atoms were omitted in the calculations.

2.3. QSAR models for log BB using CODESSA-PRO and ISIDA approach

CODESSA-PRO software was used to calculate structural descriptors for the entire data set consisting of 127 compounds. A large POOL of 877 molecular descriptors of different types as constitutional, topological, geometrical, electrostatic, and thermodynamic was calculated. We have selected the descriptor pool by eliminating those descriptors not expected to be significant link to the blood–brain partition coefficient. Eliminated descriptors included the structural descriptors consisting of specific atom types, thermodynamic descriptors linked only to the stability of the compounds and their physical properties, and descriptors with small variances and/or non-Gaussian distributions. A set of 474 descriptors was thus obtained for the development of the QSAR model for rat log BB values.

As lipophilicity is an important parameter affecting the brain penetration¹¹ we added calculated Clog *P*⁹² values, as an external descriptor in the CODESSA-PRO treatment. For the calculation of Clog *P* values the ChemDraw version 8.0.3⁹² program was used. The method is based on fragment approach for log *P* (octanol–water) prediction.⁹³

To model 127 log BB values in rat, we built multiple-parameter correlations with up to ten descriptors using the selected descriptor POOL of 475 descriptors including Clog *P*. The application of 'break point' procedure to these preliminary QSAR models indicates that five is optimum number of parameters for the QSAR model

from the POOL of 475 descriptors (see plot in (SM-2)). The statistical characteristics of the optimum descriptor regression model (number of data points (*n*), number of descriptors (*k*), squared correlation coefficient (R^2), cross-validated squared correlation coefficient (R_{CV}^2), Fisher ratio (*F*), squared standard deviation (s^2) together with descriptor coefficients, '*t*' test values, and the descriptors involved in the model) are listed in Table 3. The five descriptor values for 127 compounds are listed in (SM-3).

The statistical characteristics are found to be poor with standard error value $s^2 \approx 0.25$ for the log BB model of 127 compounds as shown in Table 3. Some of the experimental data are outliers to the 2σ rule, as implemented in the BMLR method: 9, 24, 30, 36, 38, 42, 43, 61, 81, 82, 83, 85, 87, and 106 (Table 2). It is observed that most models exclude a number of compounds from the analysis; invariably these are compounds that have observed log BB values much more negative than calculated. There are several possible reasons for such outliers: (a) experimental errors due to different experimental protocols used in the measurement of log BB⁷ values; if the analytical method were radiochemical detection, any biological degradation will lead to much smaller observed values than calculated; (b) efflux mechanisms most notably by P-glycoprotein⁹⁴ will also result in more negative observed values than calculated. The compounds 9 and 24 are derived from the radiochemical detection measurements²⁴ which were found as outliers. Also some of the outliers (30, 36, 38, 61, 81, and 106) are similar to outliers reported by Platts et al.,⁷ several causes are also discussed therein.

Further, based on the structural pattern of outlier compounds we then grouped the outliers into four different types. The first set includes the compounds 9, 106, 81, 83, and 87. The compound tiotidine (no. 9) and 4-phenyl-2-thiazole ethanamine (no. 106) contain thiazole ring and three compounds fluphenazine (no. 81), mesoridazine (no. 3), and norchlorpromazine (no. 87) with similar structural and substitution patterns. The second set of outliers, salicylic acid (no. 36), ibuprofen (no. 38), indomethacin (no. 42), and oxazepam (no. 3), contains carboxylic acid or similar groups. Platts et al.⁷ used an indicator variable for the acid functionality to model these compounds. However, the plasma–brain distribution ratios have been taken for the log BB values for salicylic acid (no. 36) and ibuprofen (no. 38). The third set

Table 3. The multilinear regression model of log BB for 127 compounds using CODESSA-PRO ($R^2 = 0.631$, $R_{CV}^2 = 0.586$, $F = 41.4$, and $s^2 = 0.249$)

No.	<i>X</i>	$\pm\Delta X$	<i>t</i> -test	Descriptor
0	0.179	0.187	0.954	
1	−0.148	0.023	−6.437	Number of double bonds, <i>D</i> ₁
2	0.204	0.034	5.972	Clog <i>P</i> , <i>D</i> ₂
3	−0.161	0.030	−5.31	H-donors CPSA (version 2), <i>D</i> ₃
4	−0.138	0.028	−4.941	Kier flexibility index, <i>D</i> ₄
5	6.771	1.618	4.181	Max partial charge (Zefirov) for all atom types, <i>D</i> ₅

of outliers includes compounds piperazine, 1-6-chloro-5-(trifluoromethyl)-2-pyridinyl-monohydrochloride (no. 61), haloperidol (no. 82), and bromoperidol (no. 85). Haloperidol (no. 82) and bromoperidol (no. 85) have similar structures, which may lead to specific interactions in the nerve cells. The fourth type outlier is represented by the single compound (no. 24), which has an ether linkage.

According to Stouch et al.⁹⁵ report on the several case history of failure of models, it is important to choose a proper data set for modeling. Thus, our training set was chosen selectively based on the data measured roughly by the same experimental procedure. In the present work, in our training set the indirect experimental values of logBB used by Platt et al.⁷ are not included.

Table 4. Statistical characteristics of the models involving up to ten descriptors: modeling of logBB for 113 compounds

No.	R^2	R_{CV}^2	ΔR^2	ΔR_{CV}^2
2	0.573	0.548		
3	0.684	0.657	0.111	0.109
4	0.757	0.730	0.073	0.073
5	0.781	0.752	0.024	0.022
6	0.799	0.767	0.018	0.015
7	0.810	0.776	0.012	0.009
8	0.827	0.795	0.017	0.018
9	0.835	0.799	0.008	0.005
10	0.840	0.803	0.005	0.003

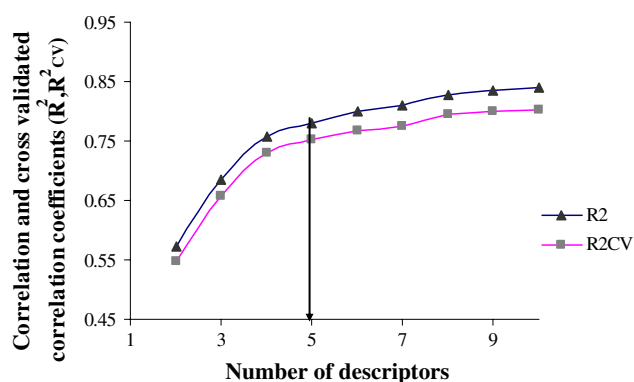


Figure 2. Correlation and cross-validated correlation coefficients (R^2 and R_{CV}^2) versus number of descriptors for logBB values of 113 compounds.

In view of these considerations, a data set of experimental values of logBB for 113 compounds was chosen to build the final regression model. Using the same set of 475 descriptors we redeveloped multilinear regression equations using the BMLR method included in CODESSA-PRO. The statistical characteristics for those models involving up to ten descriptors are given in Table 4. The five-descriptor regression model was chosen as the optimum according to the break point in the plot of correlation coefficient (R^2) and cross-validated correlation coefficient (R_{CV}^2) against the number of descriptors in the model (Fig. 2).

This optimal regression model and the statistical characteristics including 't' test values of the descriptors, descriptor coefficients, and correlation coefficients are shown in Table 5.

3. Discussions

3.1. Discussion on the QSAR models for the prediction of logBB penetration

3.1.1. CODESSA model. The multicollinearity of the five descriptors involved in the regression model (Table 5) is shown in Table 6. The lower values of the variance inflation factor (VIF) show that there are no high multicollinearity effects present in the proposed logBB model for 113 compounds. The experimental and calculated logBB values (using Table 5) are given in Table 2. The predicted ranges of a five-descriptor logBB model for 113 compounds are in good agreement with experimental data with minimum standard error ($s^2 = 0.123$). The predicted range is -1.64 to 1.37 compared with the experimental range -2 to 1.44 . Our model consists of one structural, one topological, two charge-related, and the calculated ClogP values. The most significant descriptor in Table 5 based on 't' test values is the lipophilicity parameter (ClogP), D_2 . The positive contribution of ClogP indicates that an increase of this descriptor leads to higher brain penetration. It is evident from the Atkinson's paper¹¹ that lipophilicity is a significant parameter affecting the brain penetration. The second significant descriptor is the Kier flexibility index D_4 , defined as in Eq. 3⁹⁶

$$\varphi = \frac{{}^1k^2k}{N_{SA}}, \quad (3)$$

where 1k and 2k are Kier shape indices and N_{SA} is the number of non-hydrogen atom in the molecule.

Table 5. The multilinear QSAR model of logBB for 113 compounds using CODESSA Pro. ($R^2 = 0.781$, $R_{CV}^2 = 0.752$, $F = 76.1$, $s^2 = 0.123$)

No.	X	$\pm\Delta X$	t -test	R^2	R_{CV}^2	Descriptor
0	0.378	0.143	2.647			Intercept
1	0.224	0.025	8.88	0.279	0.254	ClogP, D_2
2	-0.176	0.021	-8.18	0.539	0.512	Kier flexibility index, D_4
3	-0.131	0.017	-7.932	0.666	0.638	Number of double bonds, D_1
4	-0.162	0.023	-7.062	0.753	0.726	H-donors CPSA (version 2), D_3
5	4.744	1.378	3.442	0.781	0.752	Maximum partial charge (Zefirov) for all atom types, D_5

Table 6. Multicollinearity of descriptors of logBB model (Table 5) using CODESSA-PRO

Dependent variable	Independent variable	Tolerance	Variance inflation factor
D1	D2	0.734	1.363
	D3	0.850	1.176
	D4	0.864	1.157
	D5	0.899	1.113
D2	D1	0.959	1.042
	D3	0.935	1.069
	D4	0.967	1.034
	D5	0.943	1.061
D3	D1	0.834	1.199
	D2	0.702	1.425
	D4	0.789	1.267
	D5	0.916	1.092
D4	D1	0.920	1.087
	D2	0.787	1.270
	D3	0.856	1.168
	D5	0.909	1.100
D5	D1	0.837	1.195
	D2	0.671	1.490
	D3	0.869	1.151
	D4	0.795	1.257

This topological descriptor contains information concerning molecular shape and complexity of the molecule. Thus, the structural features of a compound like molecular branching, degree of cyclicality and spatial density may be attributed to the brain penetration. The third descriptor is the number of double bonds D_1 , which relates to the structural rigidity of the molecule. The fourth descriptor, the H-donors CPSA (version 2) D_3 , describes the donor–acceptor interactions between the solute and solvents that influence the logBB partition coefficient. The fifth descriptor is the maximum partial charge (Zefirov) for all atom types D_5 which can be related to the solvation of compounds and describes their capability to dissolve in different environments (intracellular vs blood). The plot of calculated versus observed logBB values in rat for 113 compounds according to the five-descriptor regression model (Table 5) is shown in Figure 3.

3.1.2. ISIDA model. The first series of calculations has been performed on the set of 113 compounds taken to

derive the optimal regression model (Table 5) with CODESSA-PRO. Most of the ISIDA's models reasonably reproduce the logBB values for all compounds except the compound 116 (Table 2) for which the difference between the calculated and experimental values was larger than three standard deviations for selected individual models. Therefore, this compound was excluded from the data set. The descriptor pool contained more than 2800 fragment descriptors including atoms, bonds, and atoms/bonds sequences from 2 to 8 atoms and augmented atoms. Individual models with correlation coefficients $R^2 > 0.8$ and cross-validation correlation coefficients $R_{cv}^2 > 0.7$ have poor statistical characteristics on internal validity (see Section 3.2). In most cases, the 'consensus models' (see Section 2.2) are more predictive. For the initial set of 112 compounds, the developed 'consensus model' includes the 48 best models, for which the values of internal cross-validation correlation coefficient $R_{cv}^2 \geq 0.6$ (see (SM4)). Statistical characteristics of the 'consensus model' are $R^2 = 0.872$ and $s^2 = 0.047$ for the linear correlation between experimental and calculated blood–brain partition coefficient (Fig. 3).

3.2. Cross-validation

3.2.1. Internal validation of the proposed model. An important aspect of any QSAR study is the validation of the model. For the internal validation, the parent data set was divided into three subsets: the first, fourth, seventh, etc., entries form the first subset (no. 1), the second, fifth, eighth, etc., entries form the second subset (no. 2), and the third, sixth, ninth, etc., form the third subset (no. 3). Three training sets were prepared as a combination of two subsets, set I, (nos. 1 and 2), set II, (nos. 1 and 3), and set III (nos. 2 and 3). For each training set the remaining subset (nos. 3, 2, and 1, respectively) constituted the test set (Tables 7a and 7b). For each training set the correlation equation was derived by CODESSA-PRO with the same descriptors as shown in Table 3. Then the equations obtained for the training set were used to predict the logBB values of the test set data. Similar methods have been used elsewhere.^{68,71,72}

Both CODESSA-PRO and ISIDA predict reasonably well the logBB values from three test sets: for the linear correlation between experimental and 'fitted predicted'

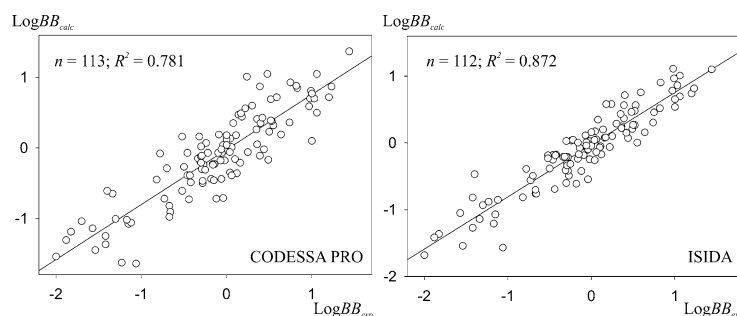
**Figure 3.** The plot of observed versus calculated blood–brain partition coefficient logBB values in rat using CODESSA-PRO and ISIDA programs.

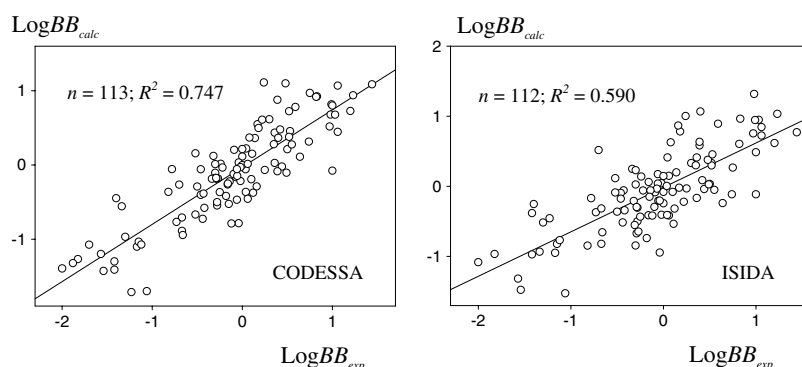
Table 7a. Internal validation of the QSAR models of logBB (Table 5) using CODESSA-PRO approach

Training set	<i>N</i>	R^2 (fit)	R_{CV}^2 (fit)	s^2 (fit)	Test set	<i>N</i>	R^2 (Pred)	s^2 (Pred)
CODESSA-PRO								
A + B	76	0.797	0.756	0.105	C	37	0.757	0.164
A + C	75	0.807	0.769	0.124	B	38	0.709	0.120
B + C	75	0.753	0.700	0.140	A	38	0.829	0.098
Average		0.786	0.742	0.123			0.765	0.127

Table 7b. Internal validation of the QSAR models of logBB using consensus model, ISIDA approach

Training set	<i>N</i>	R^2 (fit)	<i>k</i> ^a	s^2 (fit)	Test set	<i>N</i>	R^2 (Pred)	s^2 (Pred)
ISIDA								
A + B	74	0.896	36	0.038	C	38	0.604	0.157
A + C	75	0.895	37	0.042	B	37	0.526	0.153
B + C	75	0.908	23	0.031	A	37	0.672	0.110
Average		0.900		0.037			0.601	0.140

^a For each training set the ‘average model’ includes *k* best models, for which the values of internal cross-validation correlation coefficient $R_{CV}^2 \geq 0.6$.

**Figure 4.** Predicted versus experimental blood–brain partition coefficient (logBB) values in rat. Internal validation of the CODESSA and ISIDA models.**Table 8.** The experimental and calculated logBB values for external test set compounds

No.	Compound	logBB _{exp}	logBB _{CODESSA}	logBB _{ISIDA}
1	Benzene	0.37	0.79	0.13
2	Trichloroethylene	0.34	0.40	0.00
3	Acetone	−0.15	0.10	−0.10
4	2-Butanone	−0.08	0.06	−0.10
5	Ethanol	−0.16	−0.02	−0.08
6	1-Propanol	−0.16	−0.05	−0.10
7	2-Methyl-1-propanol	−0.17	0.17	−0.07
8	2-Propanol	−0.15	0.24	−0.09
9	Chloroform	0.29	0.52	0.00
10	Pentane	0.76	0.49	0.07
11	2-Methyl pentane	0.97	0.72	0.12
12	3-Methyl pentane	1.01	0.73	0.13
13	2,2-Dimethyl butane	1.04	0.97	0.14
14	Heptane	0.81	0.38	0.18
15	3-Methyl hexane	0.9	0.68	0.18
16	Methylcyclopentane	0.93	1.00	0.12
17	Cyclohexane	0.92	0.93	0.10
18	Diethyl ether	0	0.03	−0.11
19	Divinyl ether	0.11	0.01	−0.16

data, the correlation coefficient R^2 (Pred.) and standard deviation s (Pred.) are 0.747 and 0.33 (CODESSA-PRO), 0.590 and 0.37 (ISIDA), see Figure 4.

3.2.2. External validation. The real predictive ability of any QSAR model cannot be judged solely by using internal validation, it has to be validated on the basis of predictions for compounds not included in the training set.⁹⁷ A test set of logBB values for 19 compounds is chosen for the validation (see Table 8). These are indirect experimental logBB values for human⁹⁸ obtained from log(brain–air) and log(blood–air) by using the simple algorithm as shown in Abraham’s papers.^{28–30}

The fitness of the test set data is shown using our models, Table 5 and the ‘consensus model’, and is listed in Table 8. The correlation coefficient R^2 and standard deviation s^2 for the linear correlation between experimental and predicted values were found to be 0.766 and 0.032 (CODESSA-PRO), 0.827 and 0.0025 (ISIDA) as shown in Figure 5.

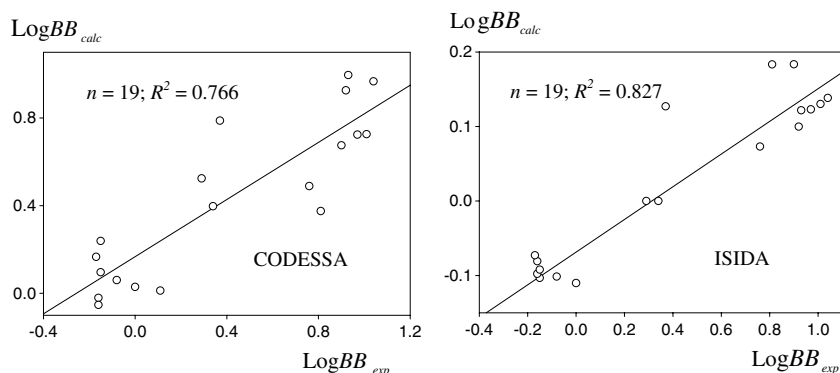


Figure 5. The plot of observed versus calculated log BB values for 19 external test set compounds fitted using Table 5 (CODESSA-PRO) and the ‘consensus model’ (ISIDA).

3.2.2.1. Prediction of brain penetration data. The predictability of the models was tested for the CNS observed activity data taken from Crivori et al.¹⁷ The

log BB values were predicted for external test of 40 compounds using the model Table 5 as shown in Table 9. A ‘+’ value was assigned if log BB predicted > 0; otherwise,

Table 9. CNS activity prediction for external validation set

Compound	CNS activity (obsd)	CNS _{CODESSA}	log BB _{CODESSA}	CNS _{ISIDA}	log BB _{ISIDA}
Diphenhydramine	+	+	0.55	+	0.56
Estradiol	+	+	0.5	+	0.62
Perphenazine	+	+	0.22	+	0.87
Progesterone	+	+	0.31	+	0.81
Rivastigmine	+	+	0.21	+	0.34
Roxindole	+	+	0.15	+	0.41
Tamitinol	+	–	–0.3	–	–0.74
Testosterone	+	+	0.4	+	0.75
Thioridazine	+	+	1.01	+	0.71
Flupentixol cis	+	+	0.47	+	0.85
Flupentixol trans	+	+	0.48	+	0.9
Aldosterone	–	–	–0.79	+	0.05
Astemizole	–	+	0.49	–	–0.21
Carbidopa	–	–	–1.59	–	–0.89
Carebastine	–	–	–0.73	+	0.59
Carmoxirol	–	–	–0.5	+	0.33
Ciprofloxacin	–	–	–1.08	–	–0.56
Cortisol	–	–	–0.71	–	–0.03
Difloxacin	–	–	–0.68	+	0.09
Dopamine	–	–	–0.61	–	–0.46
Ebastine	–	+	0.46	+	1.07
Enoxacin	–	–	–1.7	–	–0.79
Fleroxacin	–	–	–1.04	–	–0.12
Furosemide	–	–	–1.02	–	–1.33
Isoxicam	–	–	–0.69	–	–1.45
Levodopa	–	–	–1.97	–	–0.91
Lomefloxacin	–	–	–1.02	–	–0.5
Loperamide	–	+	0.36	+	1.03
Loratadine	–	+	0.355	+	0.73
Meloxicam	–	–	–0.58	–	–0.95
Mequitazine	–	+	1.05	+	0.93
Corticosterone	–	–	–0.09	+	0.31
Norfloxacin	–	–	–1.13	–	–0.47
Ofloxacin	–	–	–0.9	–	–0.61
Pefloxacin	–	–	–0.95	–	–0.16
Pirenzepine	–	–	–1.04	–	0.03
Piroxicam	–	–	–0.74	–	–1.05
Tenoxicam	–	–	–1.11	–	–1.05
Rufloxacin	–	–	–0.85	–	–0.18
Sparfloxacin	–	–	–1.4	–	–0.93

a ‘–’ value was assigned if $\log BB$ (predicted) < 0. The predicted CNS activity data are listed in Table 9. The CODESSA-PRO model Table 5 correctly predicts 34 among 40 of the observations, whereas the ISIDA ‘consensus model’ predicts correctly 30 among 40. Predicted values of two methods correlate: $\log BB_{ISIDA} = 0.27 + 0.75 \log BB_{CODESSA}$, $n = 40$, $R^2 = 0.619$, $F = 62$, and $s = 0.46$. Thus, the both approaches not only reasonably classify (‘active’–‘non-active’) the compounds from the external set, but the calculated activities are also in quantitative agreement. As may be seen from the Table 9, the predictive ability of both our models is good. Among the compounds incorrectly classified by both approaches the two CNS[–] compounds (loperamide and mequitazine) are common outliers as shown by Luco.³⁷ In the case of mequitazine, CNS activity is dose limiting, whereas, loperamide is a P-glycoprotein substrate drug. The overall results on validations confirm the robustness of the models to be used for the QSAR prediction of blood–brain distribution ratios.

3.3. Combined approach

We combined the fragment descriptors together with molecular descriptors calculated using CODESSA-PRO. A combined pool of 1751 descriptors was used to build the regression model for 127 compounds using the BMLR method implemented in the CODESSA-PRO. The statistical characteristics of the models showed no significant improvement as those of compared to the previous models using CODESSA-PRO and ISIDA approach.

4. Conclusions

This study has shown that the blood–brain barrier penetration of drugs ($\log BB$) can be modeled in terms of structure-based descriptors solely calculated from the structure of a molecule. The validation and cross-validation of the QSAR models (Table 5 of CODESSA-PRO and ‘consensus model’ of ISIDA) suggest that the models can be used to make predictions for compounds not in the original data set. The structural information encoded in the descriptors in the discussed models indicated the significant and specific structure information that may be useful for new compound design.

Thus, the proposed QSAR models could become a guiding tool for analog design to improve CNS penetration and for de novo modeling of CNS^{+/–} libraries.

Acknowledgement

The Estonian Science Foundation Grant No. 4548 is acknowledged for the partial support of this work.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bmc.2006.03.012](https://doi.org/10.1016/j.bmc.2006.03.012).

References and notes

- Bodor, N.; Buchwald, P. *Adv. Drug Delivery Rev.* **1999**, *36*, 229.
- Schinkel, A. H. *Adv. Drug Delivery Rev.* **1999**, *36*, 179.
- Terasaki, T.; Hosoya, K. *Adv. Drug Delivery Rev.* **1999**, *36*, 195.
- Li, A. P. *Drug Discovery Today* **2004**, *9*, 204.
- Gratton, J. A.; Abraham, M. H.; Bradbury, M. W.; Chadha, H. S. *J. Pharm. Pharmacol.* **1997**, *49*, 1211.
- Clark, D. E. *Drug Discovery Today* **2003**, *8*, 927.
- Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. *Eur. J. Med. Chem.* **2001**, *36*, 719.
- Salminen, T.; Pulli, A.; Taskinen, J. *J. Pharm. Biomed. Anal.* **1997**, *15*, 469.
- Kunsmann, G. W.; Rohrig, T. P. *Am. J. Forensic Med. Pathol.* **1993**, *14*, 48.
- Clark, D. E. *Comb. Chem. High Throughput Screening* **2001**, *4*, 477.
- Atkinson, F.; Cole, S.; Green, C.; Van de Waterbeemd, H. *Curr. Med. Chem.* **2002**, *2*, 229.
- Norinder, U.; Haeberlein, M. *Adv. Drug Delivery Rev.* **2002**, *54*, 291.
- Basak, S. C.; Gute, B. D.; Drewes, L. R. *Pharm. Res.* **1996**, *13*, 775.
- Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. *J. Drug Target.* **1998**, *6*, 151–165.
- Bemis, A. G. W.; Murcko, M. A. *J. Med. Chem.* **1999**, *42*, 4942.
- Kelder, J.; Grootenhuys, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. *Pharm. Res.* **1999**, *16*, 1514.
- Crivori, P.; Cruciani, G.; Caruut, P.-A.; Testa, B. *J. Med. Chem.* **2000**, *43*, 2204.
- Cruciani, G.; Pastor, M.; Guba, W. *Eur. J. Pharm. Sci.* **2000**, *11*, S29.
- (a) Keseru, G. M.; Molnar, L.; Greiner, I. *Comb. Chem. High Throughput Screening* **2000**, *3*, 535; (b) Lobell, M.; Molnar, L.; Keseru, G. M. *J. Pharm. Sci.* **2003**, *92*, 360.
- Subramanian, G.; Kitchen, D. B. *J. Comput. Aided Mol. Des.* **2003**, *17*, 643.
- Mahar, D. K. M.; Humphreys, J. E.; Webster, L. O.; Wring, S. A.; Shampine Larry, J.; Serabjit-Singh, C. J.; Adkison, K. K.; Polli, J. W. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 1029.
- Adenot, R.; Lahana, N. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239.
- (a) Ramamurthi, N.; Gunturi, S. B. *ARKIVOC* **2004**, *11*, 102; (b) Beteringhe, A.; Filip, P.; Tarko, L. *ARKIVOC* **2005**, *10*, 45.
- Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. *J. Med. Chem.* **1988**, *31*, 656.
- Seiler, P. *Eur. J. Med. Chem.* **1974**, *9*, 473.
- Van de Waterbeemd, H.; Kansy, M. *Chimia* **1992**, *46*, 299.
- Calder, J. A. D.; Ganellin, C. R. *Drug Design Discovery* **1994**, *11*, 259.
- Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. *J. Pharm. Sci.* **1994**, *83*, 1257.
- Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. *Drug Design Discovery* **1995**, *13*, 123.
- Abraham, M. H.; Takacs-Novak, K.; Mitchell, R. C. *J. Pharm. Sci.* **1997**, *86*, 310.
- Lombardo, F.; Blake, J. F.; Curatolo, W. J. *J. Med. Chem.* **1996**, *39*, 4750.
- Brewster, M. E.; Pop, E.; Huang, M.-. Ju.; Bodor, N. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1996**, *23*, 1775.

33. Kaliszan, R.; Markuszewski, M. *Int. J. Pharm.* **1996**, *145*, 9.
34. Norinder, U.; Sjöberg, P.; Oesterberg, T. *J. Pharm. Sci.* **1998**, *87*, 952.
35. Oesterberg, T.; Norinder, U. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1408.
36. Segarra, V.; Lopez, M.; Ryder, H.; Palacios, J. M. *Quant. Struct.-Act. Relat.* **1999**, *18*, 474.
37. Luco, J. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396.
38. Clark, D. E. *J. Pharm. Sci.* **1999**, *88*, 815.
39. Feher, M.; Sourial, E.; Schmidt, J. M. *Int. J. Pharm.* **2000**, *201*, 239.
40. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
41. Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. *J. Comput. Aided Mol. Des.* **2001**, *15*, 697.
42. Keseru, G. M.; Molnar, L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 120.
43. Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651.
44. Hou, T.; Xu, X. *J. Mol. Model.* **2002**, *8*, 337.
45. Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. *J. Pharm. Res.* **2002**, *19*, 1611.
46. Sun, H. A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748.
47. Cabrera, M. A.; Bermejo, M.; Perez, M.; Ramos, R. *J. Pharm. Sci.* **2004**, *93*, 1701.
48. Stanton, D. T.; Mattioni, B. E.; Knittel, J. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1010.
49. Fu, X. C.; Wang, G. P.; Liang, W. Q.; Yu, Q. S. *Pharamazie* **2004**, *59*, 126.
50. Pan, D.; Iyer, M.; Jianzhong, L.; Li, Y.; Hopfinger, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083.
51. Narayanan, R.; Gunturi, S. B. *Bioorg. Med. Chem.* **2005**, *13*, 3017.
52. CODESSA-PRO User's Manual, <http://www.codessa-pro.com/manual/manual.htm>.
53. ISIDA (In Silico Design and Data Analysis) Project: <<http://infochim.u-strasbg.fr/recherche/isida/index.php/>>, 2004.
54. STATISTICA 6.0.
55. Dutta, S.; Matsumoto, Y.; Muramatsu, A.; Matsumoto, M.; Fukuoka, M.; Ebling, W. F. *Br. J. Anaesth.* **1998**, *81*, 422.
56. Cheng, F. C.; Ho, Y. F.; Hung, L. C.; Chen, C. F.; Tsai, T. H. *J. Chromatogr., A* **2002**, *949*, 35.
57. Cheng, H. Y.; Liu, T.; Feuerstein, G.; Barone, F. C. *Free Radical Biol. Med.* **1993**, *14*, 243.
58. Brown, E. A.; Griffiths, R.; Harvey Carol, A.; Owen, D. A. *Br. J. Pharmacol.* **1986**, *87*, 569.
59. Gabrielsson, J.; Bondesson, U. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 583.
60. Goldman, S. S.; Hass, W. K.; Ransohoff, J. *Am. J. Physiol.* **1980**, *238*, H776.
61. Boehlen, P.; Schlunegger, Urs. P.; Laeuppi, E. *Toxicol. Appl. Pharm.* **1973**, *25*, 242.
62. Katritzky, A. R.; Mu, L.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162.
63. Katritzky, A. R.; Taemm, K.; Kuanar, M.; Fara, D. C.; Oliferenko, A.; Oliferenko, P.; Huddleston, J. G.; Rogers, R. D. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 136.
64. Katritzky, A. R.; Wang, Y.; Slid, S.; Tamm, T.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720.
65. Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr. *Bioorg. Med. Chem.* **2004**, *12*, 4735.
66. Maran, U.; Karelson, M.; Katritzky, A. R. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3.
67. Katritzky, A. R.; Tatham, D. B.; Maran, U. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162.
68. Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529.
69. Katritzky, A. R.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2004**, *12*, 3027.
70. Katritzky, A. R.; Oliferenko, A.; Lomaka, A.; Karelson, M. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3453.
71. Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E., Jr.; Solov'ev, V. P.; Varnek, A. *Bioorg. Med. Chem.* **2005**, *13*, 6450.
72. Katritzky, A.; Dobchev, D.; Hur, E.; Fara, D.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 1623.
73. Katritzky, A. R.; Kulshyn, O. V.; Stoyanova-Slavova, I.; Dobchev, D. A.; Kuanar, M.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2006**, *14*, 2333.
74. Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. *Curr. Top. Med. Chem.* **2002**, *12*, 1333.
75. ISIS draw.
76. Hyperchem, v. 7.5; Hypercube Inc.; Gainesville, FL, 2003.
77. Stewart, J.J.P. MOPAC 6.0; QCPE No 455, 1989.
78. Dewar, J. S. M.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
79. Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem.* **1996**, *100*, 10400.
80. (a) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000; (b) Devillers, J.; Balaban, A. T. Eds.; *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999; (c) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: New York, 2000.
81. Solov'ev, V. P.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847.
82. Varnek, A.; Wipff, G.; Solov'ev, V. P. *Solvent Extr. Ion Exch.* **2001**, *19*, 791.
83. Varnek, A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 812.
84. Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1703.
85. Varnek, A.; Fourches, D.; Solov'ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D. C.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1365.
86. Solov'ev, V. P.; Varnek, A. A. *Russ. Chem. Bull.* **2004**, *53*, 1434.
87. Varnek, A.; Solov'ev, V. P. *Comb. Chem. High Throughput Screening* **2005**, *8*, 403.
88. (a) Grubbs, F. *Technometrics* **1969**, *11*, 1–21; (b) Muller, P. H.; Neumann, P.; Storm, R. *Tafeln der mathematischen Statistik*; VEB Fachbuchverlag: Leipzig, 1979.
89. Barai, S. V.; Reich, Y. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **1999**, *13*, 377.
90. Bianconi, R.; Galmarini, S.; Bellasio, R. *Environmental Modelling & Software* **2004**, *19*, 401.
91. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244.
92. [ClogP] calculated by using Chem draw version 8.0.3.
93. Mannhold, R.; Van de Waterbeemd, H. *J. Comput. Aided Mol. Des.* **2001**, *15*, 337.
94. *Blood-Brain Barrier and Drug Delivery to the CNS*; Begely, D. J., Regina, A., Khan, E. U., Rollinson, C., Abbott, J., Anthony, R., Begely, M., Roux, F., Bradbury, M. W., Kreuter, J., Eds.; Marcel Dekker: New York, 2000; pp 93–120.

95. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A.; Li, Y. *J. Comput. Aided Mol. Des.* **2003**, *17*, 83.
96. Kier, L. B. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science: New York, 1990; p 329.
97. Wold, S. In *Chemometric Methods in Molecular Design*. van de Waterbeemd, H. Eds.; 1995; Vol. 2, Wiley-VCH: Weinheim, Chapter 4.4, pp 195–218.
98. Abraham, M. H.; Weathersby, P. K. *J. Pharm. Sci.* **1994**, *83*, 1450.