Journal of Computer-Aided Molecular Design **17:** 415–433, 2003.

Prediction of blood–brain barrier permeation using quantum chemically derived information

Michael C. Hutter*
Max-Planck-Institute of Biophysics, Theoretical Biophysics Group, Kennedyallee 70, D-60596 Frankfurt, Germany; Present address: Saarland University, Center for Bioinformatics, Gebäude 17.1 1.OG, P.O. Box 15 11 50, D-66041 Saarbrücken, Germany

Received 19 November 2002; accepted 19 June 2003

Key words: blood–brain partition coefficient, brain penetration, drug transport, hydrogen-bonding, QSPR, semi-empirical calculations, structure–activity relationship

Summary

A model for the prediction of the blood–brain distribution (logBB) is obtained by multiple regression analysis of molecular descriptors for a training set of 90 compounds. The majority of the descriptors are derived from quantum chemical information using semi-empirical AM1 calculations to compute fundamental properties of the molecules investigated. The polar surface area of the compounds can be described appropriately by six descriptors derived from the molecular electrostatic potential. This set shows a strong correlation with the observed logBB. Additional quantum chemically computed properties that contribute to the final model comprise the ionization potential and the covalent hydrogen-bond basicity. Complementary descriptors account for the presence of certain chemical groups, the number of hydrogen-bond donors, and the number of rotatable bonds of the compounds. The quality of the fit is further improved by including variables derived from principal component analysis of the molecular geometry.

Introduction

Among the pharmacokinetic issues during the design of new drugs, prediction of the blood–brain barrier (BBB) permeability is a crucial factor. Only substances targeting the central nervous system should readily pass the endothelial cells that separate the brain and the cerebrospinal fluid from the systemic blood circulation. Other peripherally acting agents must exhibit limited ability to cross the BBB. The ratio of the steady-state balance of the concentrations of a substance in the brain and in the blood is commonly expressed as $\log(C_{brain}/C_{blood})$, similar to the expression for logP [1], the water/*n*-octanol partition coefficient. In contrast to the values of logP that range from -4 to $+8$, most published values of logBB cover only a narrow range between -2.00 and $+1.00$. Compounds with logBB > 0.3 cross the BBB readily, while values < -1.0 indicate a low concentration in the brain. Experimental determination of the logBB is both demanding and costly [2, 3]. Moreover, a high percentage of potential drug candidates are withdrawn due to inadequate pharmacokinetic properties, often after they entered into clinical trials [4]. Thus *in silico* screening of ADME properties (absorption, distribution, metabolism, elimination) in preclinical development is indispensable and accurate predictions of logBB are highly desirable.

Early models for the predictions of logBB were based on quantitative structure–activity relationship (QSAR) equations using a rather limited number of variables that were directly related to experimentally accessible molecular properties, such as logP [1] and molecular size descriptors. In recent years sophisticated statistical approaches have been developed that can handle a vast number of descriptors, in particular those that stem from the three dimensional structure

*To whom correspondence should be addressed. E-mail: michael.hutter@bioinformatik.uni-saarland.de

and fundamental properties of the molecules [5–11]. On the other hand, the studies of Hall and Kier [12] used primarily topological 2D descriptors like the electrotopological states [13] where the generation of a three dimensional molecular structure is not needed. Further work has been reviewed by Norinder and Haeberlein [14], by Kramer [15], as well as by Clark [16]. Nevertheless, QSAR equations for the logBB are still of interest, i.e. Clark [17] and Feher et al. [8] combined logP and the polar surface area (PSA), see below.

Earlier, Kansy and van de Waterbeemd [18] employed the PSA together with the molecular volume, while Abraham and coworkers [19] introduced parameters that express the polarizability and the hydrogen bond donor/acceptor properties of the solute. A similar correlation was found for Δ logP(octanol-cyclohexane) by Young et al. [20]. Alternatively logP can be derived from the ratio of the free energies of solvation of the solute in water and in octanol. Lombardo et al. [21] and Keserü and Molnár [22] showed that the free energy of solvation computed in water is directly correlated to the logBB. A computationally more demanding approach was presented by Kaznessis et al. [23] using Monte Carlo simulations of compounds in water to derive the partitioning between water and octanol. Moreover, Kelder et al. [3] found that the PSA is directly correlated to the logBB.

The polar surface area (PSA) of the molecule is the part of its surface formed by exposed nitrogen or oxygen atoms, or by hydrogen atoms that are bonded to nitrogen or oxygen atoms. In other words, the PSA is the part of the molecular surface that is involved in the formation of hydrogen bonds. However, the PSA alone is not a sufficient criterion to predict logBB, as shown by the following example: Benzene and 3-methylpentane both do not have polar surface areas but their logBB values differ by more than one unit. In principle, the polar surface can be regarded as a simplified approach to describe the electrostatic potential on the surface of the molecule. To explicitly calculate the molecular electrostatic potential (MEP), a charge distribution is required either in the form of atom centered charges [24] in the most simple case or by the use of distributed point charges [25]. However, any charge distribution must be individually fitted on each molecule to reproduce electric moments, like the dipole and quadrupole moments. While atom centered charges can be computed from fragment based approaches, like the Gasteiger-Marsili charges [26], among the latter the VESPA point charges [27] were derived from the underlying semi-empirical method itself and have been shown to agree very well with the results from *ab initio* calculations employing the 6-31G* basis set.

A prerequisite for both PSA and MEP calculations is the availability of a 3D geometry which, in the case of the subsequent calculation of the according MEP, requires the energetic optimization of the respective molecular structure by means of quantum mechanics. That this is feasible even for the size of a combinatorial library, was shown by Clark et al. [28] who applied the semiempirical AM1 Hamiltonian to the whole set of molecules in the World Drug Index (WDI) [29]. The energetic optimization of the molecular structure moreover offers the benefits of yielding a number of fundamental molecular properties, such as the dipole moment, the ionization potential and the polarizability in addition to the conventional 1D and 2D topological descriptors.

A similar approach of deriving 3D molecular descriptors was reported by Crivori et al. and Ooms et al. for a BBB permeation model by principal component analysis of a large number of descriptors obtained by the VolSurf program [30] that computes these from 3D molecular interaction fields, which include electrostatic and lipophilicity potentials [7, 31]. Earlier, Norinder et al. employed the MolSurf program [32] to gain descriptors from the molecular wave function [5]. Alternatively, Liu et al. combined the topological polar surface area (TPSA) [33] and a lipoaffinity descriptor derived from atom connectivity information [9]. However, these potentials arise from the intrinsic electronic properties of the molecule and therefore in this study emphasis was placed on the direct computation of descriptors by means of quantum mechanics, while establishing a relationship between the descriptors and fundamental molecular properties.

Materials and methods

Drug data set

In this study, 174 compounds are used for which either the experimental logBB or the observed central nervous system activity is known. They were collected from the works of Abraham et al. [19, 34], Young et al. [20], Lin et al. [35], Kelder et al. [3], Salmien et al. [36], Calder and Ganellin [37], Luco [6], Crivori et al. [7], Lombardo et al. [21], Rose et al. [12] and Clark [17], representing a diverse range of agents. Not

Table 1. Comparison of experimental and predicted values of logBB for the training set.

No.	Compound	Observed logBB	Predicted logBB[a]	Difference
1	acetone	−0.15	−0.19	0.04
2	acetylsalicylic acid[b]	−0.50	−0.68	0.18
3	BBcpd14[b]	−0.12	−0.46	0.34
4	benzene	0.37	0.70	−0.33
5	BCNU[b]	−0.52	−0.73	0.21
6	1-butanone	−0.08	−0.10	−0.02
7	chloroform	0.29	0.41	−0.12
8	$F_3C–CH_2Cl$	0.08	0.36	−0.28
9	$Cl_2C=CHCl$	0.34	0.81	−0.47
10	clonidine[b]	0.11	−0.42	0.53
11	cimetidine[b]	−1.42	−0.90	−0.52
12	compound2[c]	−0.04	−0.68	0.64
13	icotidine[b]	−2.00	−1.45	−0.55
14	compound4[c]	−1.30	−1.27	−0.03
15	lupitidine[b]	−1.06	−1.43	0.37
16	compound6[c]	0.11	0.12	−0.01
17	mepyramine[b]	0.49	0.66	−0.17
18	compound8[c]	0.83	1.00	−0.17
19	ranitidine[b]	−1.23	−0.99	−0.24
20	compound10[c]	−0.82	−0.91	0.09
21	compound11[c]	−1.17	−1.20	0.03
22	tiotidine[b]	−0.82	−0.78	−0.04
23	compound13[c]	−0.67	−0.35	−0.32
24	compound14[c]	−0.66	−0.64	−0.02
25	compound15[c]	−0.12	−0.52	0.40
26	compound16[c]	−0.18	−0.42	0.24
27	compound17[c]	−1.15	−1.13	−0.02
28	compound18[c]	−1.57	−1.08	−0.49
29	compound19[c]	−1.54	−1.44	−0.10
30	compound20[c]	−1.12	−0.61	−0.51
31	compound21[c]	−0.73	−0.57	−0.16
32	compound22[c]	−0.27	−0.60	0.33
33	compound23[c]	−0.28	−0.46	0.18
34	compound24[c]	−0.46	−0.34	−0.12
35	compound25[c]	−0.24	−0.15	−0.09
36	compound26[c]	−0.02	−0.02	0.00
37	compound27[c]	0.69	0.11	0.58
38	compound28[c]	0.44	0.29	0.15
39	zolantidine[b]	0.14	0.26	−0.12
40	compound30[c]	0.22	0.05	0.17
41	carbamazepine[d]	0.00	−0.01	0.01
42	compound32[c]	−0.34	−0.22	−0.12
43	compound33[c]	−0.30	−0.54	0.24
44	compound34[c]	−1.34	−1.35	−0.01
45	compound35[c]	−1.82	−1.49	−0.33
46	amitriptyline[d]	0.98	1.11	−0.13
47	domperidone[d,e]	−0.78	−0.92	0.14
48	enflurane[b]	0.24	0.31	−0.07
49	diethyl ether	0.00	0.46	−0.46
50	ethanol	−0.16	−0.58	0.42

51	Cl_3C-CH_3	0.40	0.39	0.01
52	fluroxene[b]	0.13	0.56	−0.43
53	halothane[b]	0.35	0.36	0.01
54	heptane	0.81	0.71	0.10
55	hexane	0.80	0.57	0.23
56	imipramine[b]	1.05	1.07	−0.02
57	indinavir[b]	−0.74	−0.98	0.24
58	isoflurane[b]	0.42	0.51	−0.09
59	2,2-dimethylbutane	1.04	0.86	0.18
60	2-methylpentane	0.97	0.65	0.32
61	2-methyl-1-propanol	−0.17	−0.07	−0.10
62	methylcyclopentane	0.93	0.81	0.12
63	3-methylhexane	0.90	0.81	0.09
64	3-methylpentane	1.01	0.64	0.37
65	mianserin[d]	0.99	0.99	0.00
66	mirtazapine[d]	0.53	0.84	−0.31
67	Org 12962[d]	1.64	0.55	1.09
68	Org 13011[d]	0.16	−0.09	0.25
69	Org 30526[d]	0.39	0.83	−0.44
70	Org 32104[d]	0.52	0.62	−0.10
71	Org 34167[d]	0.00	0.21	−0.21
72	Org 4428[d]	0.82	0.69	0.13
73	Org 5222[d]	1.03	0.74	0.29
74	9-OH-risperidone[d]	−0.67	−0.43	−0.24
75	*p*-acetamidophenol[b]	−0.31	−0.73	0.42
76	pentane	0.76	0.84	−0.08
77	2-propanol	−0.15	−0.20	0.05
78	1-propanone	−0.15	−0.12	−0.03
79	1-propanol	−0.16	−0.19	0.03
80	risperidone[d]	−0.02	0.13	−0.15
81	salicylic acid[b]	−1.10	−0.76	−0.34
82	SKF 101468[b]	0.24	0.12	0.12
83	SKF 89124[b]	−0.44	−0.36	−0.08
84	teflurane[b]	0.27	0.38	−0.11
85	temelastine[b]	−1.88	−1.46	−0.42
86	thioperamide[b]	−0.16	−0.36	0.20
87	tibolone[d]	0.40	0.73	−0.33
88	toluene	0.37	0.64	−0.27
89	trifluoroperazine[b]	1.44	1.31	0.13
90	valproic acid[b]	−0.22	−0.28	0.06

[a] Predicted using eq. (2) (see text).
[b] Refers to the name used in ref. [12].
[c] Refers to the numbering used in ref. [17, 21].
[d] Refers to the name used in ref. [3].
[e] Active removal from the brain by P-glycoprotein [46].

Table 2. Comparison of experimental and predicted values for the test set[a].

Compound	Observed logBB	Predicted logBB[b]	Difference
alprazolam	0.04	0.59	−0.55
antipyrine	−0.10	0.08	−0.18
caffeine[c]	−0.06	−0.75	0.69
chlorpromazine	1.06	1.64	−0.58
codeine	0.55	0.43	0.12
desipramine	1.20	0.69	0.51
didanosine	−1.30	−0.77	−0.53
hydroxyzine	0.39	0.12	0.27
ibuprofen	−0.18	0.09	−0.27
indomethacin	−1.26	−0.07	−1.19
midazolam	0.61	0.86	−0.25
nevirapine	0.00	0.64	−0.64
oxazepam	0.61	−0.30	0.91
pentobarbital	0.12	−0.62	0.74
phenserine	1.00	0.17	0.83
physostigmine	0.08	0.11	−0.03
promazine	1.23	1.59	−0.36
SB 222200	0.30	−0.04	0.34
tert-butylchlorambucil	1.00	0.43	0.57
theophyline	−0.29	−1.68	1.39
thioridazine	0.24	1.28	−1.04
verapamil	−0.70	−0.10	−0.60
zidovudine	−0.72	−1.11	0.39

[a]Names refer to those used in ref. [12].
[b]Predicted using eq. (2) (see text).
[c]Carrier-mediated transport into the brain [45].

considered are compounds for which strongly different values of logBB were reported. Table 1 shows the 90 molecules used as training set to derive the logBB prediction equations. A further 23 compounds given in Table 2 were used as an external validation set. This test set is similar to that used by Luco [6]. Another 61 compounds comprise the test set for the qualitative brain penetration data, given in Table 3. This set includes agents also used in the work of Crivori et al. [7], Luco [6] and by Rose et al. [12] based on the assumption of passive permeation. Cases where side effects are known have been annotated accordingly. The chemical structures of the compounds from Tables 1 and 2 can be found in ref. [3, 5, 6, 9, 12, 17, 21, 22]. Additional structural drawings of less common agents are shown in Figure 1.

Calculation of molecular properties and descriptors

The descriptors used for deriving the blood–brain barrier prediction model are listed in Table 4. They were computed employing a modified version of the semi-empirical program package VAMP [38] and were extracted from the respective output files using a PERL script to generate appropriate data sheets. Compounds were energetically optimized to a gradient norm below 0.1 kcal mol^{-1} $Å^{-1}$ using the default Eigenvector Following algorithm [39]. Titratable groups were modeled uncharged, yielding a net neutral charge for each molecule. Quantum chemical calculations and generation of the descriptors for all compounds took less than 3 h on a single CPU of a Compaq Alpha Server ES40 (667 Mhz), which corresponds to an average of 1 molecule per minute including structural optimization. Statistical analysis (correlation and multiple linear regression) was carried out using the OpenStat2 program using default settings [40].

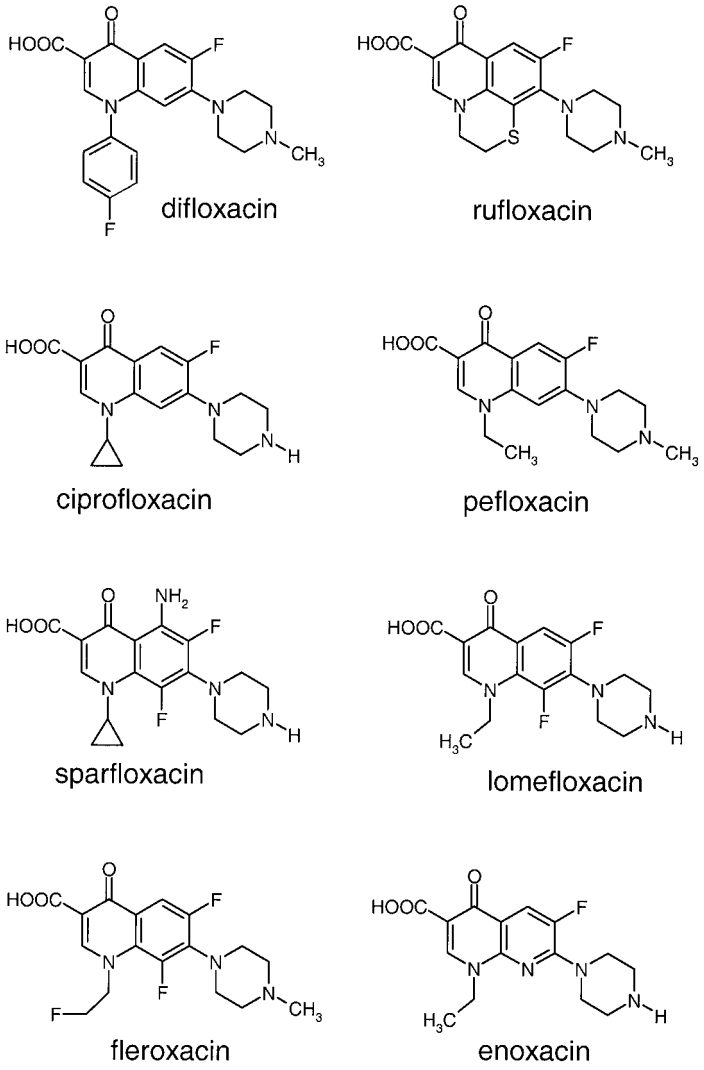

Figure 1. Chemical structures of the compounds used in Table 3. The respective isomer shown was used.

The non-standard use and generation of newly introduced descriptors are explained below. The molecular weight (MW) is simply the sum of the atomic masses. The occurrence of certain elements and chemical groups in the molecule is expressed by counting the number of hydrogen atoms (HY), the number of sulfur atoms (SU), the number of fluorine atoms (FL), the total number of all halogen atoms (HALO) incl. fluorine, chlorine, bromine, and iodine, as well as the number of NO_2-groups (NO2). The number of carboxylic acids (COOH), the number of amino groups (NH_2) and the number of guanidinium-like fragments (GUANID) is aimed to express the ambivalence of titratable groups that are subject to pK_a changes. The number of aromatic six-membered rings (AR6) is counted separately from the number of heterocyclic aromatic five-membered rings (AR5) that include N, O, and S. Contrary to the definition of Lipinski et al. [41] here, only those atoms are considered as hydrogen-bond acceptors (HBACC) that possess formal lone electron pairs (N, O, F, and S) and are not simultaneously hydrogen-bond donors. E.g. nitrogen atoms in peptide linkages (−NH−C(=O)−), as well as hydroxylic oxygen atoms (−OH) are considered as a hydrogen-bond donor. The number of hydrogen-bond donors (HBDON) also includes sulfur (−SH). Thus, depending on their substituents nitrogen, oxygen, and sulfur are either hydrogen-bond donors (if bonded to a hydrogen) or hydrogen-bond acceptors.

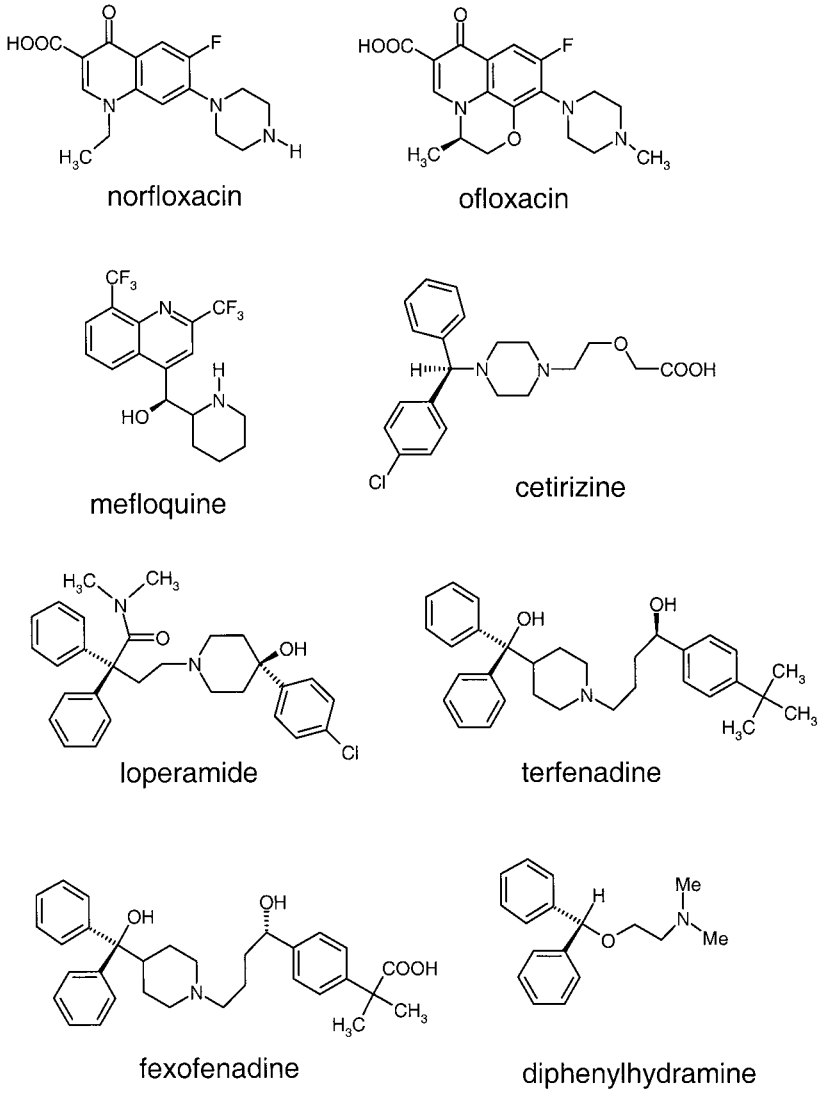

Figure 1. Continued.

To account for at least some of the entropic contributions, the number of rotatable σ-bonds (ROTB) is considered. Here, however, only those σ-bonds are counted that are terminated by non-hydrogen atoms and that are not in a ring.

Newly introduced in this study is a set of descriptors (PCGA, PCGB, and PCGC) that accounts for the molecular shape and expansion of the respective compound. These three descriptors are derived from a principal component analysis of the molecular geometry and are computed as the square root of the principal moments of the cartesian co-ordinates. The first principal component (PCGC) (the square root of the largest eigenvalue) corresponds to the largest extension of the molecule (e.g. the longest possible distance between two atoms). The second principal component (PCGB) describes the largest extension of the molecule perpendicular to the first component, and the third principal component (PCGA) in turn is the largest extension that is perpendicular to both the first and the second direction. In other words, if we had to fit the molecule into a tube with an ellipsoidal cross-section, these three descriptors would correspond to the dimensions (length and radii) of that tube. The magnitude of these dimensions is assumed to be related to the likelihood of penetrating the membrane, e.g. the probability to create an opening in between the phospholipids.

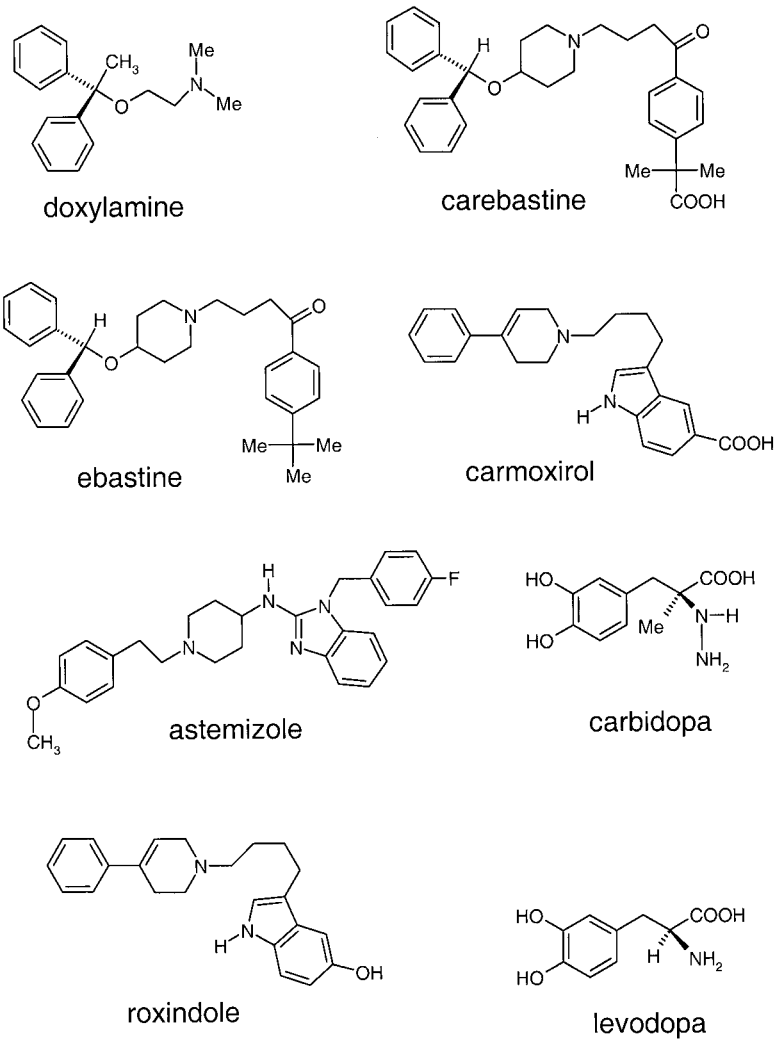

Figure 1. Continued.

Results and discussion

For some compounds used in the training sets of similar studies [9, 12, 17] strongly different experimental values of logBB were reported, e.g. chlorambucil. Preliminary studies identified those compounds causing large errors during the fitting procedure and they were therefore excluded from the training set. Furthermore methane, N_2 and N_2O were dropped. They show too few features, causing the majority of descriptors to be zero and thus producing large errors. This is in line with the findings of related studies that excluded compounds which either produced outliers or showed too few features.

Since the total number of the descriptors computed (given in Table 4) for each compound is quite large, we first identified the most relevant descriptors. This was done by inspection of the inter-correlation matrix, which is provided in the supporting information. Significant linear correlation between the observed logBB of the compounds in the training set (see Table 1) and single descriptors was found for M–ESP (0.621), QSUMN (0.553), QSUMO (0.503), and GLOBUL (0.406), respectively. Likewise anti-correlated are HBDON (−0.801), VTOT (−0.699), VXBAL (−0.697), and ESPMAX (−0.686). Some of these are, however, strongly inter-correlated with other descriptors and these were therefore dropped from the further statistical analysis as described below.

The molecular weight (MW), the molecular volume (MOLVOL), and the molecular polarizability

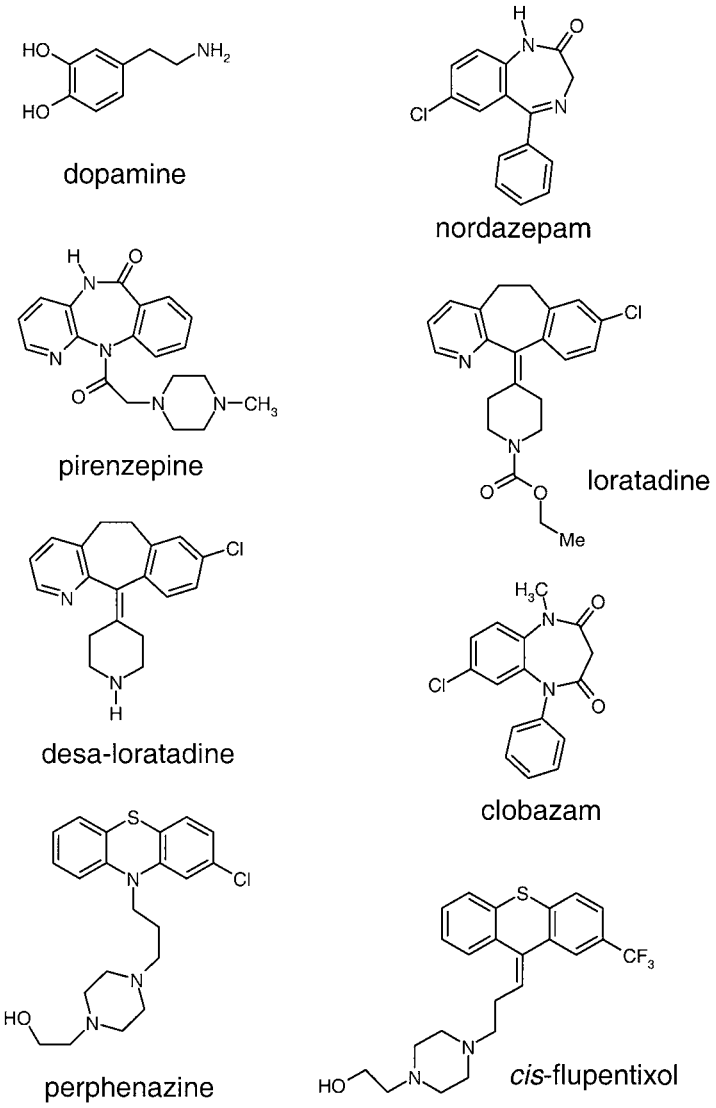

Figure 1. Continued.

(MPOLAR) form a group of descriptors that are related to the size of the respective compound. It is easy to see that the molecular polarizability scales almost linearly with the molecular volume and the number of aromatic rings due to the increased size of the (polarizable) electron cloud. The electronic multipole moments of the respective agent are much better described by the components of the electrostatic potential (atom centered point charges) than by the dipole moment (DIPOLM), which is thus obsolete. Furthermore, the relationship between the molecular volume and the dipole moment is described by the dipolar density. Interestingly, the globularity (GLOBUL) is highly anti-correlated with the number of aromatic five- and six-membered rings, the number of hydrogen bond donors and acceptors, and with a series of descriptors related to the weight, the three dimensional size and shape of the molecule, including the newly introduced shape descriptors PCGA, PCGB, and PCGC (see Methods section for details). The latter three, however, describe the extension and shape of a molecule in greater detail than the globularity which expresses the ratio between the surface and the volume [42]. They themselves exhibit a strong correlation to the total number of hydrogen atoms and the number of aromatic six-membered rings, particu-

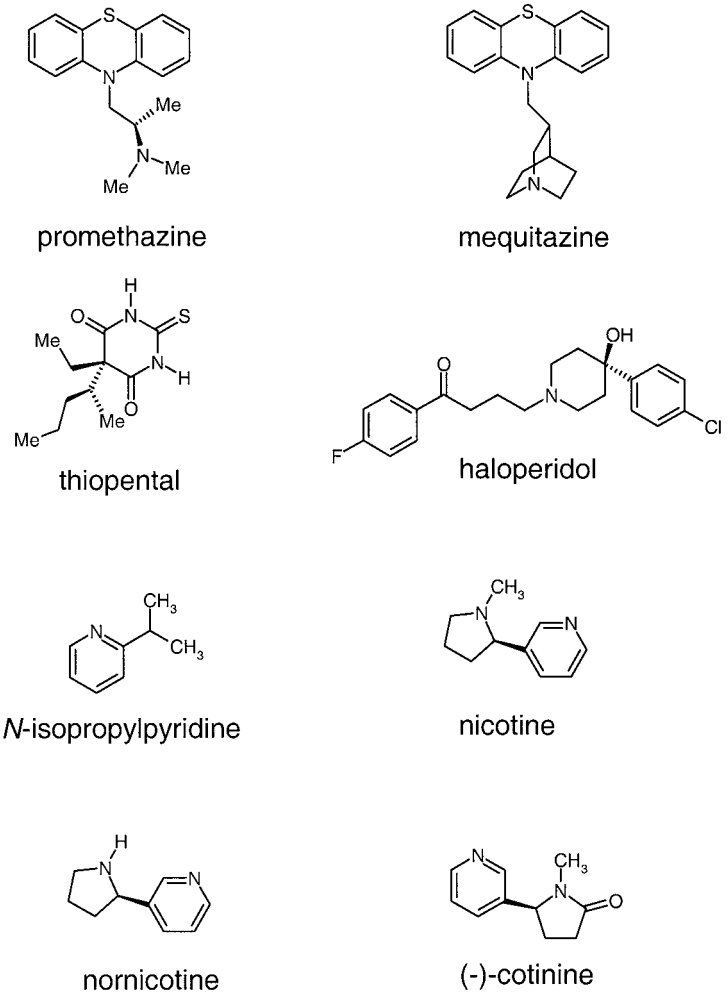

Figure 1. Continued.

larly PCGB (0.817). This is not surprising, as a phenyl ring can be regarded as the largest molecular fragment that in turn determines the minimum diameter of the cross-section. Likewise the total number of hydrogen atoms increases linearly with the size of the molecule, shown by strong correlation with the molecular weight (0.782) and the molecular volume (0.909). Among the agents in the training set, sulfur is found predominantly in heterocyclic rings systems, which explains the correlation with the number of aromatic five-membered rings. Similarly, fluorine is the most frequent element among the halogens, but interestingly, these are not responsible for the minimum of the ESP-derived charges. Likewise, the number of NO_2 groups in the molecule was found to be uncorrelated with most other variables.

To express the hydrogen-bond donor and acceptor capabilities of the compounds, a series of descriptors is available, like the number of hydrogen-bond donors and acceptors (used here is a slightly different definition compared to that of Lipinski et al. [41], see Method section for details), the covalent hydrogen-bond acidity and basicity, as well as the electrostatic hydrogen-bond acidity and basicity [43]. Despite only moderate pair-wise correlation among them, the number of hydrogen-bond donors (HBDON) and the covalent hydrogen-bond basicity (CHBBA) turned out to be most significant (see below).

From the variables representing the ESP, ESPMAX was found to be highly correlated with a large number of other descriptors. The variance of the ESP can be better described by the positive and negative variance V+ and V− than by the total variance VTOT alone.

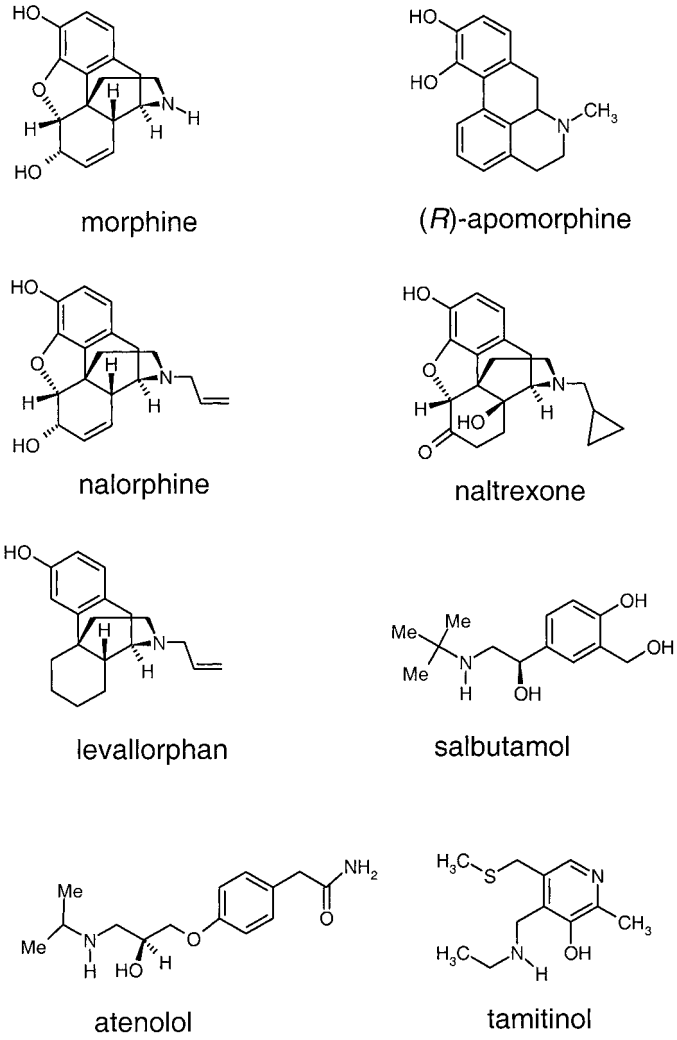

Figure 1. Continued.

Likewise VXBAL replaces VTOT and BALESP as it combines both (see the definition in Table 4). Further selection among the descriptors related to the ESP proved to be less obvious in terms of dismissing dependent variables. Therefore a best combination regression analysis was performed to elucidate how the remaining variables correlate with the observed logBB. It turned out that the majority of descriptors related to positive signed potential values can be omitted without significantly lowering the correlation. The remaining term is QSUMH, which can be regarded as a measure of the positive signed surface areas. Likewise QSUMN and QSUMO express the negatively charged surface areas arising from nitrogen and oxygen atoms. The regression eq. (1) between logBB and the employed descriptors describing the ESP found is:

$$\begin{aligned}\log \mathrm{BB} &= 0.669 - 0.012(\pm 0.004)\mathrm{ESPMIN} \\ &+0.038(\pm 0.014)\mathrm{M\text{–}ESP} - 0.010(\pm 0.002)\mathrm{VXBAL} \\ &+0.093(\pm 0.077)\mathrm{QSUMH} + 0.292(\pm 0.044)\mathrm{QSUMN} \\ &+0.673(\pm 0.115)\mathrm{QSUMO} \end{aligned} \quad (1)$$

with $r = 0.873$, $r^2 = 0.762$, adj. $r^2 = 0.745$, $F = 44.228$, $se = 0.395$, and $n = 90$ (training set).
r is the correlation coefficient, r^2 is the squared regression coefficient, adj. r^2 is the adjusted squared regression coefficient, F is the Fisher value, *se* is the standard error of estimate, and n is the number of compounds. This direct correlation (see Figure 2A) between terms related to the polarity of the molecular surface and logBB is in excellent agreement with the findings of Kelder et al. [3] and Clark [17].

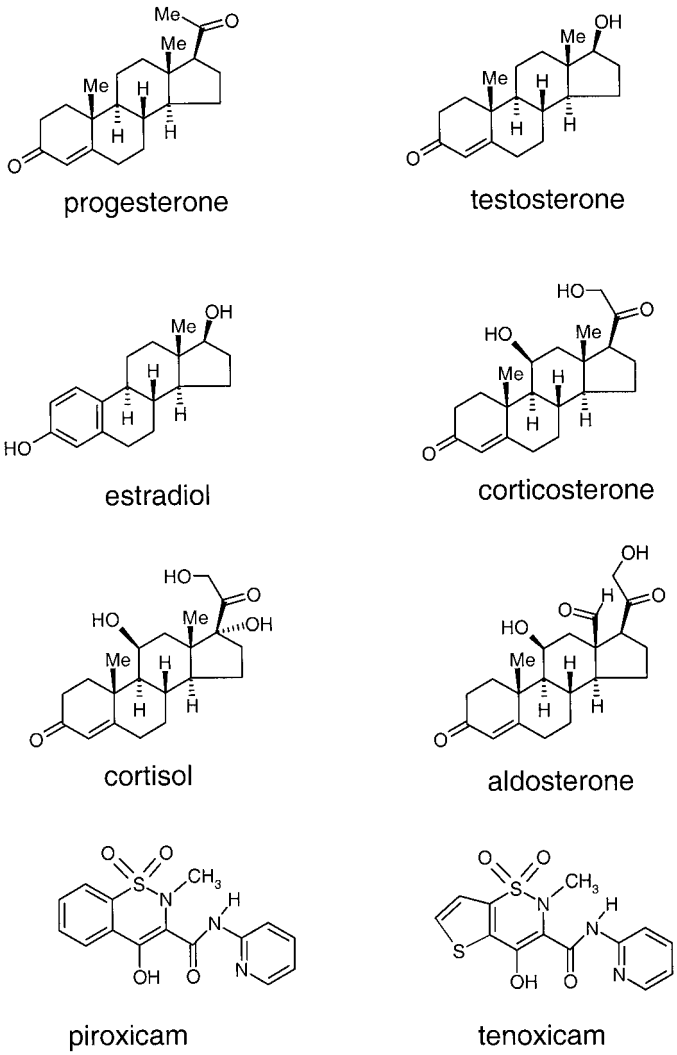

Figure 1. Continued.

Further reduction among the ESP-related variables yielded M–ESP and VXBAL as the most significant descriptors. Likewise, the remaining of the 28 variables given in Table 4 can be grouped into categories, each representing specific properties: the hydrogen-bond donor and acceptor capabilities, the presence of certain chemical groups, observable physical quantities, as well as shape and entropy descriptors. The most significant descriptors of each group were identified by a best combination regression successively reducing the number of descriptors. Thus, a total of 12 variables was retained for the final model. It turned out that omitting further descriptors affected the predictive power for both the test set (given in Table 2), and for the qualitative blood–brain barrier penetration set (given in Table 3) more than the quality of the training set. Among these 12 terms are M-ESP and VXBAL in the block expressing the ESP-related variables which are also used in eq. (1) and give rise to the polarity of the molecular surface (PSA). The second block holds the terms expressing the hydrogen-bond donor/acceptor capabilities of the compounds. Here, HBDON and CHBBA were found to be most significant. The third block summarizes those variables that indicate the presence of certain chemical groups in the agent. Among these HALO, NO2, SU, and AR6 turned out to be relevant. Particularly NO2 is necessary to allow an appropriate fit to the training set, as around 9% of the compounds possess an NO_2-group. AR6 showed better statistical properties than AR5.

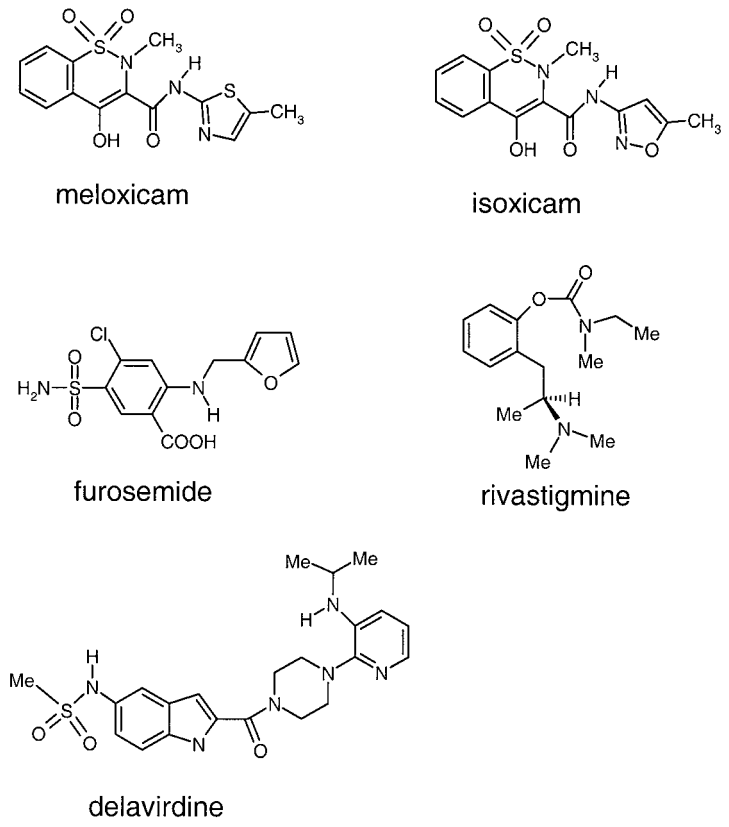

Figure 1. Continued.

Interestingly, the variables that account for titratable groups (COOH, NH2, and GUANID) were found to be less sound. Among the fourth block containing directly observable physical quantities (IP, DIPDENS, DIPOLM, and MPOLAR) only IP was retained. The other terms are obviously represented better by the included ESP-related variables. The last block finally includes the set of descriptors that are derived from the three-dimensional geometry of the molecule (PCGA, PCGB, PCGC, and GLOBUL) as well as the resulting entropic contributions (ROTB). Here, PCGC (the longest extension of the molecule), PCGA (the 3rd longest extension of the molecule), and ROTB were found to be most important. PCGB was not used, as it is highly intercorrelated with AR6 that is already present. Blockwise multiple regression showed that all of these five groups contribute significantly to the correlation, giving rise to the final regression eq. (2).

$$\begin{aligned}\log \mathrm{BB} &= 3.5075 + 0.0449(\pm 0.0089)\mathrm{M\text{–}ESP} \\ &\quad -0.0096(\pm 0.0027)\mathrm{VXBAL} - 0.3723(\pm 0.0516)\mathrm{HBDON} \\ &\quad +0.0181(\pm 0.0074)\mathrm{CHBBA} + 0.0948(\pm 0.0287)\mathrm{HALO} \\ &\quad +0.4456(\pm 0.1623)\mathrm{NO2} + 0.2148(\pm 0.1215)\mathrm{SU} \\ &\quad -0.0927(\pm 0.0626)\mathrm{AR6} - 0.2457(\pm 0.0552)\mathrm{IP} \\ &\quad +0.1170(\pm 0.0253)\mathrm{PCGA} - 0.0205(\pm 0.0066)\mathrm{PCGC} \\ &\quad -0.0395(\pm 0.0182)\mathrm{ROTB} \end{aligned} \tag{2}$$

with $r = 0.930$, $r^2 = 0.865$, adj. $r^2 = 0.844$, $F = 41.125$, $se = 0.309$, and $n = 90$ (training set).

A plot of the predicted versus the experimental values of logBB is shown in Figure 2B. Keeping all relevant raw variables in the regression equation can be considered as an alternative to the use of a partial least square algorithm to obtain a strongly reduced number of latent variables, that exhibit an apparently low descriptor : compound ratio. Here, however, all 12 retained variables are grouped into 5 categories, each representing specific properties: The PSA, the hydrogen-bond donor and acceptor capabilities, the presence of certain chemical groups, observable physical quantities, as well as shape and entropy

Table 3. Comparison of experimental and predicted blood–brain barrier penetration for the qualitative blood–brain barrier penetration set.

Compound	Observed[a]	Predicted[b]	Predicted logBB[b]
difloxacin	−	−	−0.14
rufloxacin	−	−	−0.27
ciprofloxacin	−	−	−0.26
pefloxacin	−	−	−0.32
sparfloxacin	−	−	−0.16
lomefloxacin	−	−	−0.15
fleroxacin	−	−	−0.24
enoxacin	−	−	−0.34
norfloxacin	−	−	−0.28
ofloxacin	−	−	−0.28
mefloquine	−	+	0.35
cetirizine	−	+	0.28
loperamide	−[c]	+	0.22
terfenadine	−[c]	−	−0.47
fexofenadine	−	−	−1.20
diphenylhydramine	+	+	0.57
doxylamine	+	+	0.77
carebastine	−	−	−0.63
ebastine	−	+	0.02
carmoxirol	−	−	−0.66
astemizole	−	−	−0.19
roxindole	+	−	−0.62
carbidopa	−	−	−1.58
levodopa	−	−	−1.33
dopamine	−	−	−0.87
nordazepam	+	+	0.31
pirenzepine	−	−	−0.43
loratadine	−	+	0.95
desa-loratadine	−	+	1.29
clobazam	+	+	0.09
perphenazine	+	+	1.01
cis-flupentixol	+	+	0.62
promethazine	+	+	1.27
mequitazine	−[d]	+	1.39
thiopental	+[e]	−	−0.04
haloperidol	+	−	−0.27
N-isopropylpyridine	+	+	0.80
nicotine	+	+	0.90
nor-nicotine	+	+	0.65
(−)-cotinine	+	−	−0.05
morphine	+	+	0.19
R-apomorphine	+	−	−0.54
nalorphine	+	+	0.09
naltrexone	+	−	−0.21
levallorphan	+	+	0.96
salbutamol	−	−	−1.43
atenolol	−	−	−1.20
tamitinol	+	−	−0.43
progesterone	+	+	0.57
testosterone	+	+	0.15
estradiole	+	+	0.06
corticosterone	−	−	−0.40
cortisol	−	−	−0.94
aldosterone	−	−	−0.65
piroxicam	−	−	−0.92
tenoxicam	−	−	−0.77
meloxicam	−	−	−0.52
isoxicam	−	−	−0.77
furosemide	−	−	−1.21
rivastigmine	+	+	0.31
delavirdine	−	−	−1.30

[a]Compounds with poor brain penetration are denoted as −, those that are readily absorbed as +.
[b]+/− assignment based on the predicted logBB using eq. (2). Compounds are assigned CNS– if the calculated logBB < 0, otherwise CNS+.
[c]Active removal from the brain by P-glycoprotein [46].
[d]CNS activity is dependent on dose.
[e]Anaesthetic that changes the permeability of the blood–brain barrier.

descriptors. Thus the actual descriptor : compound ratio in this approach is around the same magnitude as in related studies. Moreover, these categories allow an easier interpretation of the descriptor space compared to a recent study of Hou and Xu who employed genetic algorithms to select and fit a total of 27 descriptors to their training set [11].

When omitting the two descriptors PCGA and PCGC the quality of the statistical fit decreases notably ($r = 0.908$, $r^2 = 0.825$, adj. $r^2 = 0.803$, $F = 37.237$, $se = 0.347$), in particular the standard error of estimate (se) is affected. As mentioned above, these descriptors correspond to the longest and shortest radii of a cylindrical cavity that has to be created in between the phospholipids of the endothelial cell membrane upon permeation. This process is similar to the creation of a cavity in a solvent and the required energy is therefore related to the free energy of solvation which is in turn correlated with the logBB [21, 22]. From the thermodynamic point of view, this energy term is dependent on the actual size of the cavity (which would be expressed by the molecular volume), the electrostatic interaction energy between solute and solvent, as well as dispersive energy terms. Usually the former two terms are of similar magnitude while the dispersive contributions are much smaller. Here, the electrostatic and surface related terms are represented by M–ESP and VXBAL. Moreover, when replacing PCGA and PCGC by either MOLVOL or GLOBUL all statistic criteria worsen. It can be assumed that the

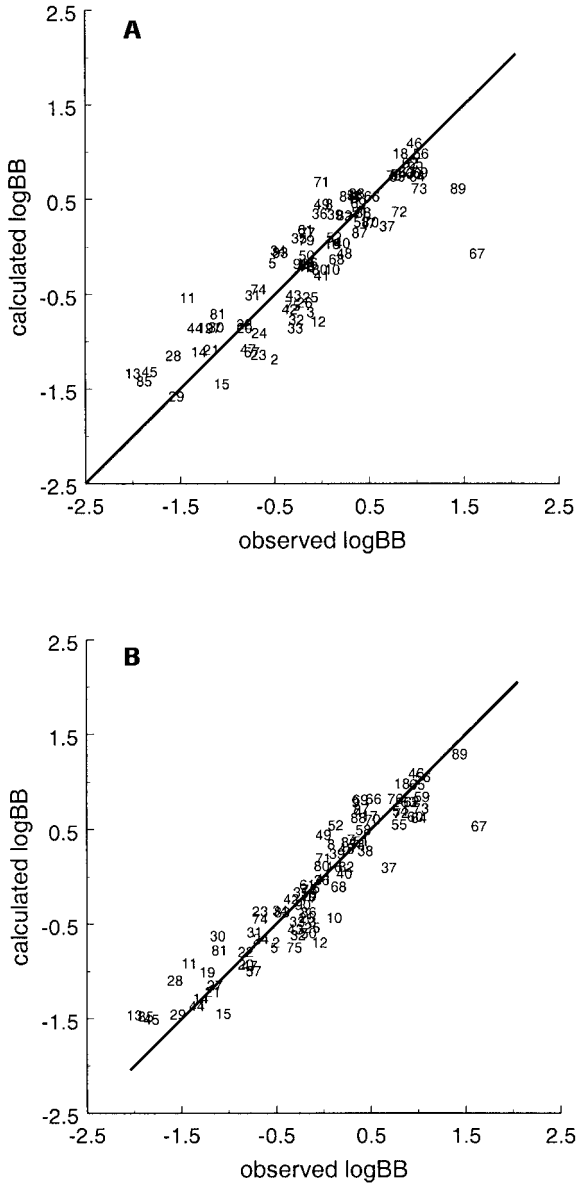

Figure 2. Plot of computed versus observed logBB values for the compounds in the training set given in Table 1. (A) Using eq. (1) that contains six descriptors derived from the molecular electrostatic potential. (B) Using the set of descriptors given in eq. (2).

shape of the molecule plays an important role during the transition from a solvent of high dielectric constant (water) to a medium or low dielectric constant (the inner part of the membrane). This is another reason why the logBB was found to be correlated with the logP (the partition coefficient between water and *n*-octanol), particularly in combination with the PSA [8, 17, 19]. Among descriptors which describe the polarity of the surface, regression eq. (2) also contains descriptors which were previously used in a neuronal net to predict logP, either expressing identical properties (M–ESP, HALO), or similar terms (GLOBUL instead of PCGA and PCGC) [44]. Thus this model includes descriptors of logP to some extent.

Interpretation of the results has to be done with respect to the nature of the blood–brain distribution of the agents. The presented model is supposed to account for passive transport through the blood–brain barrier only, while *in vivo* a number of side reactions are simultaneously present. Active transport increases the concentration, e.g. of caffeine in the brain [45], while loperamide, domperidone, and terfenadine are removed by P-glycoprotein [46]. Some compounds may also involve interaction through their metabolites (e.g. carebastine from ebastine) which exhibit different behavior.

When using the model of eq. (1) which is based on ESP-derived descriptors (Figure 2A) the strongest deviations result from compounds **67** (Org12962), **2** (acetylsalicylic acid), **13** (icotidine), **11** (cimetidine), **89** (trifluoroperazine), **71** (Org34167), **15** (lupitidine), **12** (compound2), and **85** (temelastine). Comparison with the final model derived from eq. (2) (see Figure 2B) shows a significant improvement for most of them (**67, 2, 13, 11, 89**, and **71**). Org12962, icotidine, temelastine, and cimetidine were shown to exhibit large deviations from the linear relationship between the dynamic polar surface area and the logBB, while the static and dynamic PSA of these compounds do not differ significantly [3]. Trifluoroperazine and compound2 were not part of the data set in that work.

Similar studies that employ the PSA also reveal icotidine, cimetidine, and compound2 as the largest outliers, while Org12962 and temelastine were not among the data used [9, 17]. For a number of agents (e.g. tiotidine which shows a low logBB, compound22, compound23) significantly better results were obtained with the final model than in related studies. Especially tiotidine has been identified as a problematic compound throughout many approaches [5, 9, 12, 17, 21]. The training sets of Lombardo et al., Clark, Luco, Norinder et al., Keserü and Molnár, Liu et al., and Rose et al. are similar regarding the scope and number of compounds, while Platts et al. [10] employed a total of 148 molecules in their set. The training set of Crivori et al. differs from those mentioned above as only drug-like agents were included [7].

A comparison of the results for the test set (given in Table 2) with the results of other studies is hampered by mutual overlap between the various training and test sets [6, 12]. However, oxazepam and pentobarbital are also among the largest outliers in the studies of Rose et al. and Luco, which both differ from the present approach. In contrast, they reveal difficulties with trifluoroperazine, *tert*-butylchlorambucil, and desipramine, while conversely the outliers here comprise theophyline, thioridazine, and indomethazine. The logBB of caffeine was predicted too low in agreement with the observed active transport into the brain [45]. For other compounds among the outliers similar metabolic interactions might be present. Compared to the training set (data given in parentheses) the absolute mean error is 0.566 (0.217), the variance of the error is 0.450 (0.082), and the standard deviation of the error is 0.671 (0.287).

The most similar approach to this study using quantum chemical derived descriptors is the approach of Norinder et al. [5], employing the MolSurf program [32] that also requires AM1-optimized molecule geometries using a total of 14 descriptors on 56 compounds in their training set. They, however, derived the surface related descriptors from additional single point *ab initio* calculations at RHF/3-21G* level, while the ESP-related properties used here stem from the semi-empirical AM1 wave function and have proved to be in very good agreement with results from the larger and more balanced RHF/6-31G* basis set [27]. For the computation of descriptors Crivori et al. used the VolSurf program [30] that also requires semi-empirical optimized geometries, but including solvation effects [7]. Relevant differences of the molecular geometry compared to AM1 (gas phase) optimized structures can, however, be expected only in the case of a hydrophobic collapse-like back-folding of the molecule due to pronounced intra-molecular hydrogen-bonding and/or salt bridges. Moreover, the actual geometry during the membrane passage is uncertain due to fluctuating interactions with the phospholipids. Likewise the polar surface is almost independent of the molecular conformation [3].

The approach of Crivori et al. [7] differs furthermore from other works as no numerical prediction of the logBB is intended. Instead, the resulting principal components of the variables were used to assign the CNS activity of compounds in the test set to either low penetration (BBB−), medium permeability (BBB±), or ready absorbance (BBB+). Luco and Rose et al. separated their compounds into CNS− and CNS+ while employing a margin of −1 or 0 for logBB, respectively, while the choice of agents used differs significantly [6, 12]. The latter margin was also used for the set of compounds given in Table 3. Again a series of molecules is found that exhibit problems throughout all prediction approaches: mequitazine, loperamide and terfenadine, which are known to show side reactions. Further compounds where metabolic interaction was reported comprise naltrexone, thiopental, caffeine, cetirizine, morphine, loratadine, desa-loratadine, and mefloquine. Those were also identified in the comprehensive data set of Crivori et al. [7]. Apparent are the difficulties of the presented model for morphine-like agents, especially *R*-apomorphine and naltrexone. This could be due to the presence of bridged polycyclic ring systems which are known to cause misleading results when predicting electronic properties from AM1 optimized geometries [47]. On the other hand, the prediction is very good for the group of quinolones (difloxacin and related compounds), similar to the results of Crivori et al. and Rose et al. [12]. The CNS activity of the steroids shows a clear differentiation between those that are readily absorbed and CNS inactive compounds.

Conclusions

The presented model for the passive transport through the blood–brain barrier was applied successfully to the prediction of logBB data. Further improvements can be expected when more data become available and when the accuracy of the existing experimental data is extensively validated. Compounds that were found causing large deviations usually exhibit alternative mechanisms of uptake or efflux. This is in line with the tendency of quantum chemical property predictions to detect outliers that are caused by either inaccurate observations or alternative mechanisms of effect. Most of the employed variables stem from the AM1 semi-empirical optimized geometry of the respective compound or are quantum chemical descriptors derived from the molecular wave function. This allows the inclusion of fundamental electronic properties like the ionization potential. The set of variables which accounts for the polarity of the molecular surface arises from the molecular electrostatic potential and exhibits a high correlation to the observed logBB. These variables may therefore be regarded as a more general quantum chemical expression of the polar surface area. Some of these descriptors have been used

Table 4. Definition of the computed descriptors[a].

Variable	Descriptor definition	Reference
MW	molecular weight	this study
MOLVOL	molecular volume	[48]
GLOBUL	globularity	[42]
DIPOLM	dipole moment (from the density matrix)	[49]
MPOLAR	polarizability	[50]
IP	ionization potential (from Koopmans' theorem)	[49]
DIPDES	dipolar density (DIPOLM / MOLVOL)	[51]
PCGA	3^{rd} PC of molecular geometry	this study
PCGB	2^{nd} PC of molecular geometry	this study
PCGC	1^{st} PC of molecular geometry	this study
HBDON	number of hydrogen bond donors	this study
HBACC	number of hydrogen bond acceptors	this study
CHBAC	covalent hydrogen bond acidity	[43]
CHBBA	covalent hydrogen bond basicity	[43]
EHBAC	electrostatic hydrogen bond acidity	[43]
EHBBA	electrostatic hydrogen bond basicity	[43]
ESPMIN	minimum of the molecular electrostatic potential (ESP)	[27, 44, 52]
ESPMAX	maximum of ESP	[27, 44, 52]
MESP	mean ESP	[27, 44, 52]
M+ESP	mean + ESP	[27, 44, 52]
M−ESP	mean − ESP	[27, 44, 52]
VTOT	total variance of ESP	[27, 44, 52]
V+	positive variance of ESP	[52]
V−	negative variance of ESP	[52]
BALESP	balance parameter to the ESP	[52]
VXBAL	VTOT * BALESP	[53]
QSUMH	sum of ESP derived atomic charges on hydrogen atoms	[27]
QSUMN	sum of ESP derived atomic charges on nitrogen atoms	[27]
QSUMO	sum of ESP derived atomic charges on oxygen atoms	[27]
QSUMS	sum of ESP derived atomic charges on sulfur atoms	[27]
AR5	number of aromatic 5-membered rings	this study
AR6	number of aromatic 6-membered rings	this study
HALO	sum of halogen atoms (F, Cl, and Br)	this study
FL	number of fluorine atoms	this study
HY	number of hydrogen atoms	this study
SU	number of sulfur atoms	this study
NO2	number of NO_2 groups	this study
COOH	number of carboxylic acids	this study
NH_2	number of amino groups	this study
GUANID	number of guanidinium-like fragements	this study
ROTB	number of rotatable σ-bonds	this study

[a]Descriptors are explained in full detail in the references given, otherwise their generation and use is explained in the Methods section.

earlier to predict logP data emphasizing their suitability for quantitative structure–property relationship purposes. Further quantum chemical variables include terms expressing the electronic hydrogen-bond donor and acceptor capabilities in addition to the numerical count of hydrogen-bond donors. The group of one-dimensional descriptors is completed by variables indicating the occurrence of certain chemical groups and substituents, such as halogens, sulfur, NO_2 and aromatic fragments. Along entropic contributions which

are expressed by the number of rotatable σ-bonds, the presented model benefits from variables derived from the geometry of the molecule which are introduced in this study. They are aimed to account for the shape and extension of the molecule and were found to be superior to the globularity descriptor. The employed variables are all related to molecular properties and therefore, instead of performing a partial least square algorithm to obtain a reduced number of latent variables, the 12 retained terms were grouped into five categories, that were assigned to specific features. For the properties comprising the PSA, the hydrogen-bond donor and acceptor capabilities, as well as the molecular shape, the according quantum chemical terms could be assigned straightforward. In contrast to most related studies the presented model for logBB prediction is strongly based on quantum chemical variables. Further improvements can be expected through the inclusion of additional terms that enlarge the available descriptor space, such as complex topological and shape indices.

Acknowledgements

I would like to thank my colleagues Elena Herzog for assistance with the PERL scripting and Volkhard Helms for useful comments.

Supporting information

The inter-correlation matrix of all variables given in Table 4 for the set of compounds shown in Table 1.

References

1. Hansch, C., Bjorkroth, J.P., and Leo, A., J. Pharm. Sci., 76 (1987) 663.
2. Van Belle, K., Sarre, S., Ebinger, G., and Michotte, Y., J. Pharmacol. Exp. Ther., 272 (1995) 1217.
3. Kelder, J., Grootenhuis, P.D.J., Bayada, D.M., Delbressine, L.P.C., and Ploemen, J.- P., Pharm. Res., 16 (1999) 1514.
4. Lin, J.H., and Y, L.A., Pharmacol. Rev., 49 (1997) 403.
5. Norinder, U., Sjöberg, P., and Österberg, T., J. Pharm. Sci., 87 (1998) 952.
6. Luco, J.M., J. Chem. Inf. Comput. Sci., 39 (1999) 396.
7. Crivori, P., Cruciani, G., Carrupt, P.-A., and Testa, B., J. Med. Chem., 43 (2000) 2204.
8. Feher, M., Sourial, E., and Schmidt, J.M., Int. J. Pharm. Sci., 201 (2000) 239.
9. Liu, R., Sun, H., and So, S.-S., J. Chem. Inf. Comput. Sci., 41 (2001) 1623.
10. Platts, J.A., Abraham, M.H., Zhao, Y.H., Hersey, A., Ijaz, L., and Butina, D., Eur. J. Med. Chem., 36 (2001) 719.
11. Hou, T., and Xu, X., J. Mol. Model, 8 (2002) 337.
12. Rose, K., Hall, L.H., and Kier, L.B., J. Chem. Inf. Comput. Sci., 42 (2002) 651.
13. Kier, L.B., and Hall, L.H., Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, 1999.
14. Norinder, U., and Haeberlein, M., Adv. Drug Deliv. Rep., 54 (2002) 291.
15. Kramer, S.D., Pharm. Sci. Techn. Today, 2 (1999) 373.
16. Clark, D.E., Comb. Chem. H. T. S., 4 (2001) 477.
17. Clark, D.E., J. Pharm. Sci., 88 (1999) 815.
18. van de Waterbeemd, H., and Kansy, M., Chimia, 46 (1992) 299.
19. Abraham, M.H., Chandha, H.S., and Mitchell, R.C., Drug Des. Discov., 13 (1995) 123.
20. Young, R.C., Mitchell, R., Brown, T.H., Ganellin, C.R., Griffith, R., Jones, M., Rana, K.K., Saunders, D., Smith, I.R., Sore, N.E., and Wilks, T.K., J. Med. Chem., 31 (1988) 656.
21. Lombardo, F., Blake, J.F., and Curatolo, W.J., J. Med. Chem., 39 (1996) 4750.
22. Keserü, G.M., and Molnár, L., J. Chem. Inf. Comput. Sci., 41 (2001) 120.
23. Kaznessis, Y.N., Show, M.E., and Blankley, C.J., J. Comput.-Aided Mol. Des., 15 (2001) 697.
24. Cox, S.R., and Williams, D.E., J. Comput. Chem., 2 (1981) 304.
25. Stone, A.J., Chem. Phys. Lett., 83 (1981) 233.
26. Gasteiger, J., and Marsili, M., Tetrahedron, 36 (1980) 3219.
27. Beck, B., Glen, R.C., and Clark, T., J. Comput. Chem., 18 (1997) 744.
28. Beck, B., Horn, A., Carpenter, J.E., and Clark, T., J. Chem. Inf. Comput. Sci., 38 (1998) 1214.
29. Derwent World Drug Index (Derwent WDI), Derwent Information, 14 Great Queen Street, London WC2B 5DF, U.K., 1997.
30. Cruciani, G., Crivori, P., Carrupt, P.-A., and Testa, B., J. Mol. Struct. (Theochem), 503 (2000) 17.
31. Ooms, F., Weber, P., Carrupt, P.-A., and Testa, B., Biochim. Biophys. Acta Mol. Basis Dis., 1587 (2002) 118.
32. Sjöberg, P., In van de Waterbeemd, H., Testa, B., and Flokers, G. (Eds.), Computer-Assisted Lead Finding and Optimization, Verlag Helvetica Chimica Acta, Basel, Switzerland, 1997, pp. 81–92.
33. Ertl, P., Rohde, B., and Selzer, P., J. Med. Chem., 43 (2000) 3714.
34. Abraham, M.H., Chanda, H.S., and Mitchell, R.C., J. Pharm. Sci., 83 (1994) 1257.
35. Lin, J.H., Chen, I.-W., and Lin, T.-H., J. Pharmacol. Exp. Ther., 271 (1994) 1197.
36. Salmien, T., Pulli, A., and Taskinen, J., J. Pharmaceut. Biomed. Anal., 15 (1997) 469.
37. Calder, J.A.D., and Ganellin, R., Drug Des. Discov., 11 (1994) 259.
38. Rauhut, G., Alex, A., Chandrasekhar, J., Steinke, T., Sauer, W., Beck, B., Hutter, M., Gedeck, P., and Clark, T., VAMP Version 6.5, Oxford Molecular, Erlangen, Germany, 1997.
39. Baker, J., J. Comput. Chem., 7 (1986) 385.
40. Miller, W.G., OpenStat2, Version 6.1, West DeMoines, Iowa.
41. Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J., Adv. Drug Deliv. Rev., 23 (1997) 3.
42. Meyer, A.Y., Chem. Soc. Rev., 15 (1986) 449.

43. Cronce, D.T., Famini, G.R., De Soto, J.A., and Wilson, L.Y., J. Chem. Soc., Perkin Trans. 2 (1998) 1293.
44. Breindl, A., Beck, B., Clark, T., and Glen, R.C., J. Mol. Model., 3 (1997) 142.
45. McCall, A.L., Mellington, W.R., and Wurtman, R., Life Sci., 31 (1982) 2709.
46. Schinkel, A.H., Wagenaar, E., Carla, A.A.M., and van Deemter, L., J. Clin. Invest., 97 (1996) 2517.
47. Clark, T., Rauhut, G., and Breindl, A., J. Mol. Model., 1 (1995) 22.
48. Pascual-Ahuir, J.L., Silla, E., and Tuñon, I., J. Comput. Chem., 15 (1994) 1127.
49. Leach, A.R., Molecular Modelling: Principles and Applications, Addison Wesley Longman Ltd, Harlow, Essex, U.K., 1996.
50. Rinaldi, D., and Rivail, J.L., Theor. Chim. Acta, 32 (1974) 243.
51. Mu, L., Drago, R.S., and Richardson, D.E., J. Chem. Soc., Perkin Trans. 2 (1998) 159.
52. Murray, J.S., Lane, P., Brinck, T., Paulsen, K., Grice, M.E., and Politzer, P., J. Phys. Chem., 97 (1993) 9369.
53. Brüstle, M., Beck, B., Schindler, T., King, W., Mitchell, T., and Clark, T., J. Med. Chem., 45 (2002) 3345.