Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 3017–3028
Bioorganic & Medicinal Chemistry

In silico ADME modelling: prediction models for blood–brain barrier permeation using a systematic variable selection method

Ramamurthi Narayanan* and Sitarama B. Gunturi

Bioinformatics Division, Advanced Technology Center, Tata Consultancy Services, 1, Software Units Layout, Madhapur, Hyderabad 500 081, India

Received 8 November 2004; revised 31 January 2005; accepted 31 January 2005

Abstract—Quantitative Structure–Property Relationship models (QSPR) based on in vivo blood–brain permeation data (logBB) of 88 diverse compounds, 324 descriptors and a systematic variable selection method, namely 'Variable Selection and Modeling method based on the prediction (VSMP)', are reported. Of all the models developed using VSMP, the best three-descriptors model is based on Atomic type E-state index (SsssN), AlogP98 and Van der Waal's surface area ($r = 0.8425$, $q = 0.8239$, $F = 68.49$ and SE = 0.4165); the best four-descriptors model is based on Kappa shape index of order 1, Atomic type E-state index (SsssN), Atomic level based AI topological descriptor (AIssssC) and AlogP98 ($r = 0.8638$, $q = 0.8472$, $F = 60.982$ and SE = 0.3919). The performance of the models on three test sets taken from the literature is illustrated and compared with the results from other reported computational approaches. Test set III constitutes 91 compounds from the literature with known qualitative BBB indication and is used for virtual screening studies. The success rate of the reported models is 82% in the case of BBB+ compounds and a similar success rate is observed with BBB− compounds. Finally, as the models reported herein are based on computed properties, they appear as a valuable tool in virtual screening, where selection and prioritization of candidates is required.

1. Introduction

The escalating cost of drug discovery and development[1] is due to the liabilities associated with the drugs that are often not identified until a compound reaches the clinic, the most expensive phase of pharma R&D. The liabilities in question include non-optimum values of absorption, distribution, metabolism and excretion, as well as toxicity, usually referred to as ADMET.[2–10] Blood–Brain Barrier (BBB) Permeation of new chemical entities (NCEs) is one of the most important ADMET properties considered in drug discovery and development. The BBB, a complex cellular system consisting of endothelial cells of the brain capillaries, plays the role of maintaining the homeostasis of the central nervous system (CNS) by separating the brain from the systemic blood circulation.[11] The distribution of potential drugs between the blood and the brain depends on the ability of compounds to penetrate the BBB. Lipophilic drugs can easily cross the BBB by passive diffusion; however, polar molecules normally do not cross the BBB, but sometimes active transport process facilitates their permeation.

In the search of new drugs targeted at CNS disease, the ideal drug candidates must be able to penetrate BBB effectively. On the other hand, peripherally acting drugs must have limited ability to cross BBB to avoid adverse CNS effects. In experiments, the relative affinity of a drug to the blood or brain tissue can be expressed in terms of the blood–brain partition coefficient, logBB = $\log(C_{brain}/C_{blood})$, where C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively. Experimental determination of BBB permeation is time consuming, expensive and requires a sufficient quantity of the pure compounds, often in radiolabelled form. This stringent criterion makes it not suitable for high-throughput screening of large compound libraries.[3,4] A reliable and easily applicable computational model[12,13] for predicting BBB permeation of drug candidates can help in early identification of compounds with poor BBB penetration profile, prior even to chemical synthesis and will therefore have a significant impact on drug discovery and development.

Keywords: Blood–brain barrier; ADMET; QSPR; Systematic descriptor selection; Virtual screening.
* Corresponding author. Tel.: +91 40 55673581; fax: +91 40 55672222; e-mail: narayananr@atc.tcs.co.in

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.01.061

Various authors[14–28] have attempted to predict BBB permeation with molecular properties such as lipophilicity (logP), topological indices, polar surface area, quantum chemical descriptors, etc., some of the reported models suffer from the use of computationally intensive calculations, making them not amenable for virtual screening of large libraries of compounds.[15,18] The most common statistical methods applied in these studies are multiple linear regression[14–16,18,19,22–27] (MLR), principal component analysis[20] (PCA) and partial least squares regression[17] (PLS). While the above statistical methods are successfully employed to develop predictive ADMET models, surprisingly, automated variable selection methods have received less attention, except for few scattered reports.[5,25,29] We believe that automated variable selection approaches[30–33] provide the following advantages over the variable reduction methods such as PLS, PCA, etc., (a) they can provide multiple models, based on different combinations of properties, thus providing the end user with multiple solutions and (b) they help in better understanding the mechanism of the modelled phenomenon.[5] These advantages prompted us to study the applications of the variable selection methods to generate predictive ADMET models and also to understand the molecular properties that influence the various ADMET properties.

In this paper, we describe the derivation of novel QSPR models for BBB permeation using VSMP,[32] an efficient systematic variable selection method along with MLR. Significantly, this is the first report as of date describing the application of VSMP for predictive ADME model generation to the best of our knowledge. Further, we report the performance of the QSPR models based on internal and external validations using three datasets taken from the literature and compared them with other published computational approaches. The application of the models for the whole medicinal chemical space is demonstrated by performing virtual screening experiments, on structurally new and diverse datasets. As the models reported herein are based on computed properties, they appear as valuable tools for virtual screening, where selection and prioritization of candidates is required.

2. Results and discussion

2.1. QSPR models for BBB Permeation using VSMP

QSPR models are typically generated from manually selected compounds and molecular properties, often chosen by intuition and experience. The recent advances in the field of computational chemistry have resulted in the easy calculation of many molecular descriptors[34–37] with potential applications in QSPR studies. Consequently, the process of selecting the best combination of descriptors, having high significance to a biological property becomes extremely difficult, particularly from a large pool of descriptors without automation. Feature selection methods[30–33] can explore various descriptor combinations based on a fitness criterion and can provide efficient models; however, they are scarcely employed to build predictive BBB models, except for the report of Hou and Xu.[25] In their study, they applied genetic algorithm, which is evolutionary in nature and consequently does not explore all possible combinations of descriptors. In the present study, we applied VSMP[32] that efficiently explores all possible combinations of descriptors and hence provides the best possible descriptor combination within the given pool of input descriptors. The VSMP methodology[32] adopted in this study selects the descriptor combinations based on two parameters, namely, the inter correlation coefficient between pairs of descriptors, r_{int}, and the correlation coefficient of prediction, q^2. The number of descriptors to be selected is fixed at the beginning of the selection process and in the present study it is fixed as 3 and 4. The subsequent steps involved in the selection of the descriptor combinations are executed as explained in the original report of Liu et al.[32]

The QSPR models for BBB permeation were generated using a training set of 88 compounds, keeping the blood–brain partition coefficient, logBB, as the dependent variable. The logBB values used in the present study are based on in vivo measurements obtained from rat studies and the permeation of the compounds across BBB was assumed to be predominantly by passive transport. Three hundred and twenty four molecular descriptors are calculated for the training set of compounds using the in-house software, 'Bio-suite' and are stored as a data table. The data table is given as an input to VSMP program, to build models based on three and four descriptors, keeping the inter descriptor correlation below 0.75.

The best three of the three and four descriptor models selected by VSMP are shown in Table 1. The best three-descriptors model, V1, is based on descriptors 254, 311 and 320 with a correlation coefficient, R, of 0.8425, and the cross-validated correlation coefficient, Q, of 0.8239. The correlation coefficients of the other two VSMP models, V2 and V3, are 0.8411 and 0.8329, respectively. Significantly, the descriptors 254 and 311 are selected in all the best three-descriptor models of VSMP. The three descriptors, in the models V1, V2 and V3 are 320, 144 and 30, respectively. The inter correlation coefficient between the selected descriptors of the three models V1, V2 and V3, range from 0.4194 to 0.5994 (Table 1).

The best four-descriptors model, V4, is based on the descriptors 144, 254, 291 and 311 with a correlation coefficient, R, of 0.8638, and the cross-validated correlation coefficient, Q, of 0.8472. Descriptors 254 and 311 are again selected in all the best four-descriptor models of VSMP, V4, V5 and V6; descriptor 291 is common in V4 and V5 and 320 is common in V5 and V6. The inter correlation coefficient between the selected descriptors of the three models, V4, V5 and V6, ranges from 0.121 to 0.5994 (Table 1).

The regression equations derived by performing MLR of the three-descriptor combinations are given:

Table 1. Descriptors selected by VSMP

Sample size	No. of input descriptors	Model no.	Selected descriptors	*R*	*Q*	*F*	SE
88	324	V1	254, 311, 320	0.8425	0.8239	68.49	0.4165
		V2	144, 254, 311	0.8411	0.8221	67.73	0.4181
		V3	30, 254, 311	0.8329	0.8147	63.45	0.4278
		V4	144, 254, 291, 311	0.8638	0.8472	60.98	0.3918
		V5	254, 291, 311, 320	0.8588	0.8412	58.30	0.3985
		V6	201, 254, 311, 320	0.8552	0.8366	56.51	0.4031

Where 30—topological Xu index, 144—Kappa shape index of order 1, 201—autocorrelation descriptor (Moran) weighted by atomic Van der Waals radius—order 3, 254—atomic type E-state index (SsssN), 291—atomic level based AI topological descriptor (AIssssC), 311—AlogP98, 320—2D Van der Waals surface area.

$$\log BB = 0.378578(\pm 0.106952) + 0.230139(\pm 0.031888) * \text{Desc254} + 0.367865(\pm 0.035074) * \text{Desc311} - 0.00652(\pm 0.000514) * \text{Desc320} \quad \text{(V1)}$$

$$\log BB = 0.273546(\pm 0.102171) - 0.10856(\pm 0.008616) * \text{Desc144} + 0.192249(\pm 0.030690) * \text{Desc254} + 0.358259(\pm 0.034898) * \text{Desc311} \quad \text{(V2)}$$

$$\log BB = 0.165811(\pm 0.099806) - 0.19247(\pm 0.015826) * \text{Desc30} + 0.209871(\pm 0.032088) * \text{Desc254} + 0.347238(\pm 0.035406) * \text{Desc311} \quad \text{(V3)}$$

The regression equations derived by performing MLR of the four-descriptor combinations are given below:

$$\log BB = 0.320182(\pm 0.096654) - 0.11313(\pm 0.008177) * \text{Desc144} + 0.17469(\pm 0.029186) * \text{Desc254} + 0.046464(\pm 0.0130768) * \text{Desc291} + 0.347461(\pm 0.032848) * \text{Desc311} \quad \text{(V4)}$$

$$\log BB = 0.413638(\pm 0.103019) + 0.214969(\pm 0.030939) * \text{Desc254} + 0.039116(\pm 0.013224) * \text{Desc291} + 0.357370(\pm 0.033747) * \text{Desc311} - 0.006690(\pm 0.000495) * \text{Desc320} \quad \text{(V5)}$$

$$\log BB = 0.355632(\pm 0.103893) + 0.286818(\pm 0.111011) * \text{Desc201} + 0.235403(\pm 0.030929) * \text{Desc254} + 0.334636(\pm 0.036301) * \text{Desc311} - 0.006140(\pm 0.000518) * \text{Desc320} \quad \text{(V6)}$$

In the equations, quantities given in parentheses are the standard deviations of the coefficients.

It is worth mentioning that the *R* value of the best four variables model, V4, has improved marginally, by a value of 0.0213 from that of the best three variables model, V1, and a similar trend is observed in the case of the models, V5 and V6. Consequently, the performance of the four variable models, V4 to V6, has shown little improvement over the models, V1 to V3, thus implying that the likelihood of improvement of the models with five descriptors may not be much. In order to verify this inference, we calculated the frequencies of the descriptors appeared in the four variable combinations, selected by VSMP [Supplementary information] and presented below some of the significant observations: (1) There are 3050 combinations of four descriptors having the correlation with logBB greater than or equal to 0.83. Of these combinations, 2316 combinations have AlogP98 and 2294 combinations have SsssN. The other significant descriptors are number of hydrogen bond donors and number of hydrogen bond acceptors with 495 and 391 occurrences, respectively (Fig. 1). (2) Similarly, there are 1393 combinations with *R* greater than or equal to 0.835. Out of these combinations, 1195 combinations have AlogP98 and 1185 combinations have SsssN. Interestingly, Kappa shape index of order 1 and 2D Van der Waals surface area are the other significant descriptors with 202 and 199 occurrences, respectively. Significantly, as the *R*-value is increased, VSMP has selected combinations without hydrogen bond donors and acceptors (Figs. 2 and 3). In the case of models with *R*-value greater than or equal to 0.84, there are 680 combinations of four descriptors. Out of these 680 combinations, 639 have AlogP98 and 632 have SsssN. Besides AlogP98 and SsssN, the other most significant descriptors are Kappa

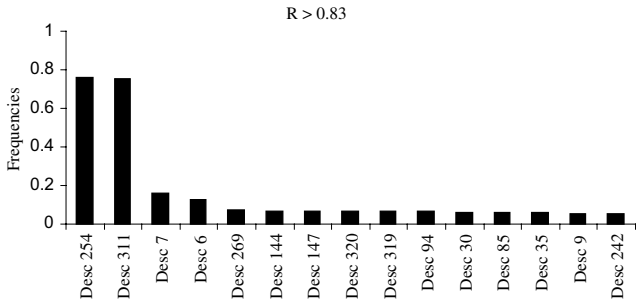

Figure 1. Frequencies of the descriptors in the four-descriptor combinations with *R*-values greater than or equal to 0.83.

shape index of order 1 (198 occurrences) and 2D Van der Waals surface area (197 occurrences). Of the 680 models selected by VSMP, the above four descriptors are the most significant ones and the rest of the descriptors have little significance (Fig. 3).

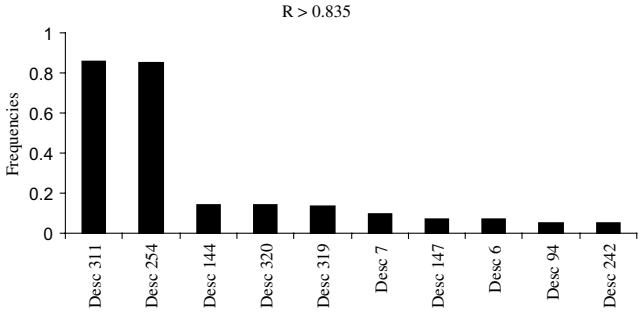

Figure 2. Frequencies of the descriptors in the four-descriptor combinations with *R*-values greater than or equal to 0.835.

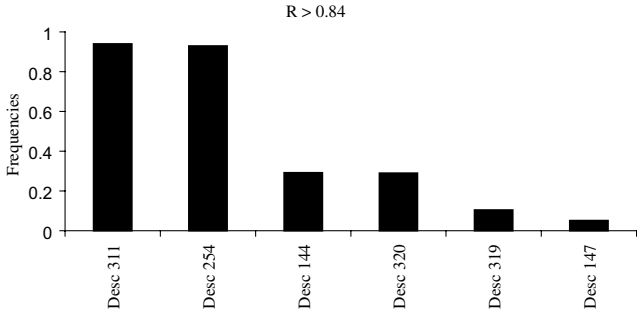

Figure 3. Frequencies of the descriptors in the four-descriptor combinations with *R*-values greater than or equal to 0.84. Where, Desc 6—no. of hydrogen bond acceptors, Desc 7—no. of hydrogen bond donors, Desc 9—no. of double bonds, Desc 30—Xu index, Desc 35—mean square distance index, Desc 85–Kier and Hall valance connectivity index of order 0, Desc 94—solvation molecular connectivity index of order 0, Desc 144—Kappa shape index of order 1, Desc 147—Kappa shape index of order 4, Desc 242—atom type E-state index—StsC, Desc 269—atom tppe E-state index—S-hydrogen bond donor, Desc 311—AlogP98, Desc 319—2D Van der Waals surface area, Desc 320—2D Van der Waals surface volume.

To summarize, we believe that (1) AlogP98 and SsssN are the most important descriptors that determine the blood–barrier permeation of a compound in our models. The role of lipophilicity in BBB permeation is well known[14,16–18,20,23,25,26,49] and the significant role of SsssN in BBB permeation is a new observation. (2) Even though descriptors such as hydrogen bond donors, acceptors and PSA have high individual correlation to BBB permeation, VSMP has not selected any combinations based on them. It is because, in the present study, the criterion of selection or omission of descriptors is based on the inter-descriptor correlation coefficient between the descriptors, which is fixed as equal to (or) less than 0.75 and hence, these descriptors are not selected as a combination by the VSMP program. (3) All the models have descriptors relating to (a) lipophilicity (descriptor number 311), (b) shape or size (descriptors 30, 144 and 320) and (c) the property SsssN (descriptor number 254), an electro topological descriptor that arises from the electronic environment of a tri substituted (with single bonds) nitrogen due to its intrinsic electronic properties and the influence of other atoms on it.

2.2. Model validation

The models V1 to V6 are validated using Leave-One-Out (LOO) method.[38] In this approach, the prediction of the property of a compound in a given set is based on the regression equation derived from the rest of the compounds of the set. The results of LOO cross-validations are given in Table 2 for models V1 and V4; and similar results for the models V2, V3, V5 and V6 are given in the Supplementary material. Similarly, a plot of the predicted logBB values versus the observed logBB values for the training set using V1 and V4 is shown in Figures 4 and 5, respectively, and the plots of the models V2, V3, V5 and V6 are given in the Supplementary material. Based on the cross-validated results given in Table 2, we performed an analysis of the percentage of confidence of prediction of the training set compounds. The confidence level of prediction of these models is 100% for BBB− compounds with experimental logBB values in the range of −2.15 to −0.50 and 66% for compounds in the range of −0.50 to 0. Similarly, the confidence level of prediction is 76% in the case of

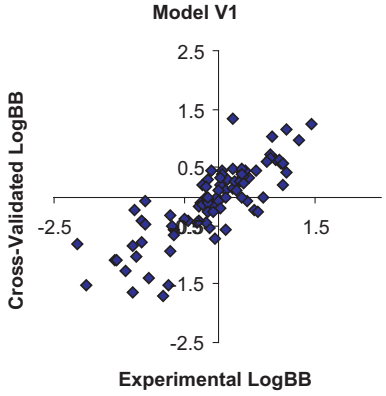

Figure 4. Plot of cross-validated logBB values using model V1 versus experimental logBB values.

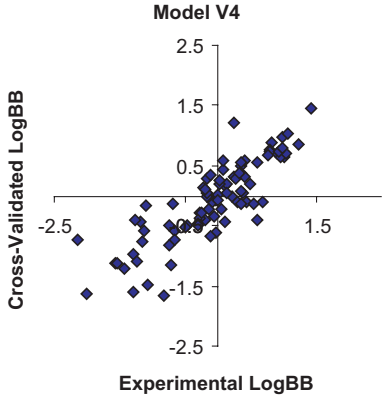

Figure 5. Plot of cross-validated logBB values using model V4 versus experimental logBB values.

Table 2. Leave-One-Out cross-validation reports of VSMP models

Compd no.	Compound name	Expt logBB	V1	V4
1	Cimetidine	−1.42	−1.27	−1.20
2	ICI17148	−0.04	−0.72	−0.61
3	Icotidine	−2.00	−1.52	−1.62
4	SK&F93319	−1.30	−0.83	−0.96
5	Lupitidine	−1.06	−1.39	−1.47
6	Clonidine	0.11	0.11	0.15
7	Mepyramine	0.49	0.31	0.19
8	Imipramine	0.83	1.04	0.89
9	Ranitidine	−1.23	−1.03	−1.10
10	Tiotidine	−0.82	−1.71	−1.66
11	BBCPD10	−1.17	−0.40	−0.42
12	BBCPD11	−2.15	−0.81	−0.72
13	BBCPD12	−0.67	−0.58	−0.57
14	BBCPD13	−0.66	−0.67	−0.72
15	BBCPD14	−0.12	−0.54	−0.68
16	BBCPD15	−0.18	−0.43	−0.40
17	BBCPD57	−1.15	−0.78	−0.76
18	BBCPD16	−1.57	−1.10	−1.12
19	BBCPD58	−1.54	−1.08	−1.13
20	BBCPD17	−1.12	−0.49	−0.58
21	BBCPD60	−0.73	−0.33	−0.50
22	BBCPD18	−0.27	−0.10	−0.28
23	BBCPD19	−0.28	−0.17	−0.38
24	BBCPD20	−0.46	−0.41	−0.49
25	BBCPD21	−0.24	−0.14	−0.29
26	BBCPD22	−0.02	0.00	−0.06
27	BBCPD23	0.69	0.00	−0.09
28	BBCPD24	0.44	−0.08	−0.11
29	Zolantidine	0.14	0.30	0.19
30	BBCPD26	0.22	0.14	−0.02
31	Butanone	−0.08	−0.13	−0.10
32	Benzene	0.37	0.41	0.49
33	3-Methylpentane	1.01	0.56	0.64
34	3-Methylhexane	0.90	0.62	0.69
35	2-Propanol	−0.15	−0.07	0.00
36	2-Methylpropanol	−0.17	−0.02	0.05
37	2-Methylpentane	0.97	0.56	0.64
38	2,2-Dimethylbutane	1.04	0.40	0.69
39	1,1,1,Trifluoro-2-chloro ethane	0.08	0.46	0.44
40	1,1,1,Trifluoro ethane	0.40	0.34	0.58
41	Diethyl ether	0.00	−0.08	0.01
42	Enflurane	0.24	0.49	0.31
43	Ethanol	−0.16	−0.10	−0.02
44	Fluroxene	0.13	−0.02	0.03
45	Halothane	0.35	0.28	0.38
46	Heptane	0.81	0.71	0.76
47	Hexane	0.80	0.66	0.72
48	Isoflurane	0.42	0.43	0.31
49	Methylchloro pentane	0.93	0.62	0.72
50	Nitrogen	0.03	0.21	0.20
51	Pentane	0.76	0.60	0.67
52	*n*-Propanol	−0.16	−0.02	0.05
53	Propanone	−0.15	−0.27	−0.22
54	Teflurane	0.27	0.26	0.28
55	Toluene	0.37	0.48	0.55
56	Trichloro ethane	0.34	0.27	0.36
57	Acetylsalicylic acid	−0.50	−0.39	−0.52
58	Pentobarbital	0.12	−0.58	−0.43
59	Physostigmine	0.079	0.38	0.59
60	Salicylic acid	−1.10	−0.09	−0.17
61	Trifluoro perazine	1.44	1.25	1.46
62	Valproic acid	−0.22	0.20	0.14
63	Verapamil	−0.70	−0.52	−0.14
64	Zidovudine	−0.72	−0.94	−1.15
65	SB222200	0.30	0.11	−0.11

(*continued on next page*)

Table 2 (*continued*)

Compd no.	Compound name	Expt logBB	V1	V4
66	Hydroxyzine	0.39	0.24	0.03
67	Thioridazine	0.24	1.34	1.22
68	Alprazolam	0.044	−0.19	−0.23
69	Phenserine	1.00	0.57	0.79
70	Midazolam	0.36	−0.02	−0.13
71	*tert*-Butyl chlorambucil	1.00	0.21	0.98
72	Codeine	0.55	−0.23	−0.12
73	Chlorpromazine	1.06	1.14	1.02
74	Promazine	1.23	0.97	0.85
75	Nevirapine	0.00	0.11	−0.06
76	Thioperamide	−0.16	0.29	0.27
77	Didanosine	−1.301	−1.63	−1.60
78	Ibuprofen	−0.18	0.16	0.10
79	Antipyrine	−0.097	0.46	0.33
80	Thiophyline	−0.29	−0.42	−0.47
81	*p*-Acetamido phenol	−0.31	−0.46	−0.48
82	Sulfahexa fluoride	0.36	0.37	0.08
83	Nitrous oxide	0.03	0.31	0.24
84	Carbon disulfide	0.60	0.44	0.55
85	Caffeine	−0.055	−0.25	−0.34
86	Indomethacin	−1.26	−0.23	−0.40
87	Indinavir	−0.745	−1.53	−0.81
88	Oxazepam	0.61	−0.27	−0.41

BBB+ compounds with experimental logBB in the range +0.01 to +1.0 and 100% for compounds in the range of +1 to +1.5. Based on these analyses, we believe that the models V1 and V4 are the best three- and four-descriptor models, generated by VSMP.

2.3. External validation of test sets I and II

The quality of a predictive model is best-tested using compounds that are not used in the training set. As we believe that the models V1 and V4 are the best models, we use them for external validation using test sets I and II. The predicted logBB values of the compounds in test sets I and II using the models V1 and V4 are given in Table 3. Analysis of the predicted values of the compounds in the test sets suggests that the actual predictive power of the models reported herein is excellent. Further, we observed the following interesting findings: (a) Of the 28 compounds used for external validation, the models V1 and V4 are able to predict the correct logBB sign in 82% of the cases and 93% of the cases, respectively. (b) Compounds Y-G19 and Y-G20 are either considered as outliers by some researchers or predicted very poorly by some of the published models and significantly, our models V1 and V4 performed very well in these cases. This observation augments well for the excellent predictive power of the models, V1 and V4. (c) SKF101468 is predicted to have logBB value of 0.14 by the model, V1, and is the best predicted value for it as against the models of Hou and Xu[26] and Liu et al.[23]

2.4. Analysis of the cross-validation and external validation results of models V1 and V4

The models V1 and V4 are based on the largest dataset, known as of date in the literature to the best of our knowledge (116 compounds and 324 descriptors) and significantly, the other reported models[14–23] in the literature are based on a relatively smaller dataset. A detailed investigation of the cross-validated values of the training set compounds (Table 2), the predicted values of the test compounds (Table 3) and the comparative analysis of the models, V1 and V4, with other reported models (Table 4) suggest that the predictive power of the models, V1 and V4, are excellent. We present below some of the interesting findings of our analysis: (1) Clonidine, **6**, a hypertensive agent is known to cause dry mouth and sedation as major adverse effects.[39] It has an experimental logBB value of 0.11 and our model, V1, predicted its value to be 0.11 and significantly all the other reported models predicted it incorrectly. (2) Diethyl ether, **41**, a highly volatile solvent with an experimental logBB value of '0' is predicted well by the model V4, which suggests, that this model can be applied to even volatile compounds like diethyl ether, as against the other reported models.[24,27] (3) Hydroxyzine,[40] **66**, is a first generation H_1 antagonist, with an experimental logBB value of 0.39 and the models V1 and V4 calculated the logBB as 0.24 and 0.03, respectively, and they stand as the best calculated values for Hydroxyzine among the reported models.[24,26] Interestingly, Citirizine,[41] **123**, a second generation H_1 antagonist is a metabolite of Hydroxyzine and is largely excluded from brain when given in therapeutic doses, because it does not cross the BBB appreciably. The model, V4, predicted the logBB of Citirizine as −0.08 (Table 5), an observation consistent with the experimental observation. (4) Pentobarbital,[42] **58**, a sedative-hypnotic drug has an experimental logBB value of 0.12 and the calculated logBB values based on the models V1 and V4 are −0.58 and −0.43, respectively. Significantly, all the reported models failed to predict the logBB value of Pentobarbital. It is well known that the carbonyl group at position 2 takes on acidic character because of

Table 3. Predicted logBB values of compounds in Test set I and II using VSMP

Compd no.	Compound name	Expt logBB	V1	V4
Test set I				
89	Y-G14	−0.30	−0.46	−0.36
90	Y-G15	−0.06	0.06	0.09
91	Y-G16	−0.42	−0.57	−0.43
92	Y-G19	−1.30	−0.32	−0.25
93	Y-G20	−1.40	−0.80	−0.67
94	SKF89124	−0.43	−0.04	−0.17
95	SKF101468	0.25	0.14	0.05
96	CBZ	0.00	0.17	0.05
97	CBZ-EPO	−0.34	−0.01	−0.16
98	L-663581	−0.30	−0.45	−0.55
99	M1L-663581	−1.34	−0.92	−0.32
100	M2L-663581	−1.82	−1.32	−0.80
101	Amitriptyline	0.88	0.66	0.57
Test set II				
102	Desipramine	1.00	0.48	0.41
103	Mianseren	0.99	1.14	0.95
104	ORG4428	0.82	0.06	0.21
105	ORG5222	1.03	0.66	0.55
106	ORG12962	1.64	0.30	0.48
107	ORG13011	0.16	0.37	0.70
108	ORG32104	0.52	−0.48	−0.28
109	ORG30526	0.39	0.08	0.04
110	Mirtazapine	0.53	0.89	0.71
111	Tibolone	0.40	−0.09	0.60
112	Domperidone	−0.78	0.10	−0.21
113	ORG34167	0.00	−0.07	−0.15
114	9OH-Resperidone	−0.02	0.09	−0.24
115	Resperodone	−0.67	−0.33	−0.67
116	Temelastine	−1.88	−1.20	−1.21

lactam-lactim tautomerization and this can probably affect the lipid-solubility property and consequently the transport of Pentobarbital. (5) Codeine,[43] **72**, an analgesic with an experimental logBB value of 0.55 is calculated to have logBB values of −0.23 and −0.12 by the models V1 and V4, respectively. Codeine has an exceptionally low affinity for opioid receptors and the analgesic effect of codeine is due to its conversion to morphine. While the calculated value is different from the experimental value, the metabolic behaviour of codeine accounts for the difference in the logBB values. Codeine is known to get metabolized to morphine by the cytochrome P450 enzyme CYP2D6 and the major pathway of metabolism of morphine is conjugation with glucuronic acid. The two major metabolites formed are morphine-6-glucuronide and morphine-3-glucuronide. Although the 3- and 6-glucuronides are quite polar, both can cross the blood–brain barrier to exert significant clinical effects.[43]

The results of the comparative study of the models V1 and V4 with other reported models are given in Table 4.

2.5. Analysis of the virtual screening results

In recent years, virtual screening methods[44–47] are increasingly used in drug discovery and development. The performance of a predictive model is best tested by virtual screening experiments using structurally new and diverse chemical compounds. Among the published reports of predictive models for BBB permeation, only two[20,27] describe their application for virtual screening. In our study, we chose the model V4 to screen the compounds of test set III. Many of these compounds are part of the test set used by Hutter[27] and Crivori et al.[20] for virtual screening studies. The results of virtual screening are given in Table 5 and are compared with those of Hutter.[27] Crivori et al.[20] reported a success rate of 90% in the case of BBB+ compounds (40 out of 44) and about 65% accuracy for the BBB− compounds (46 out of 71). Hutter[27] reported a success rate of 71% in the case of BBB+ compounds (17 out of 24) and 81% in the case of BBB− compounds (29 out of 36). From the analysis of the results in Table 5, it clearly appears that the prediction of BBB using V4 is very good, with a correct classification of 80% of the BBB+ compounds (32 out of 40) and about 81% accuracy for the BBB− compounds (42 out of 52).

A detailed investigation of the results of the external prediction shows interesting findings and some of them are listed below: (1) The correct prediction of the permeation of BBB of all the steroids **117**, **125**, **173**, **179** and **183** suggests a passive diffusion mechanism for the transport of steroids. Of significance is that the selective permeation properties of the major steroid hormones are well calculated and are coherent with the experimental data. (2) Thiopental,[48] **184**, a well-known anaesthetic is a very active CNS+ compound. It is predicted as BBB+ compound by the model, V4, with the predicted logBB

Table 4. Comparison of the calculated logBB values of the reported models

Compd no.	Expt logBB	V1	V4	Hou and Xu[26]	Rose et al.[24]	Hutter[27]	Clark[18]	LomBardo et al.[15]	Norinder et al.[17]	Kelder et al.[19]	Young et al.[14]	Keseru and Molnar[22]
1	−1.42	−1.28	−1.21	N/A	−1.01	−0.90	−1.17	−0.78	−1.21	−0.85	N/A	−0.73
2	−0.04	−0.69	−0.59	−0.82	−0.37	−0.68	−0.88	−0.82	−0.46	N/A	−1.07	0.12
3	−2.00	−1.56	−1.65	−0.90	−1.52	−1.45	−0.90	−0.68	−1.61	−1.36	−1.62	−0.63
4	−1.30	−0.86	−0.98	−1.25	−1.00	−1.27	−0.33	−0.66	−0.93	−0.73	−1.10	N/A
5	−1.06	−1.37	−1.44	−1.35	−0.89	−1.43	−0.86	−0.80	−1.05	−1.11	−0.96	−0.76
6	0.11	0.11	0.15	−0.03	−0.44	−0.42	−0.32	−0.11	−0.53	−0.01	−0.29	−0.31
7	0.49	0.32	0.21	N/A	0.52	0.66	0.17	−0.02	0.30	0.53	0.54	0.58
8	0.83	1.02	0.88	N/A	0.83	1.00	0.72	0.24	0.87	1.16	0.48	0.80
9	−1.23	−1.04	−1.11	N/A	−0.85	−0.99	−0.91	−0.49	−1.04	−0.99	N/A	−0.81
10	−0.82	−1.64	−1.60	−1.60	−1.45	−0.91	−1.79	−1.55	−1.46	N/A	N/A	−0.98
11	−1.17	−0.43	−0.45	−1.00	−0.62	−1.20	−1.03	−0.78	−0.77	−1.07	−0.68	−0.58
12	−2.15	−0.87	−0.77	N/A	N/A	N/A	N/A	−0.51	−0.61	N/A	N/A	−0.73
13	−0.67	−0.58	−0.57	−0.23	−0.53	−0.35	−0.68	−0.56	−1.00	−0.80	−0.61	−0.67
14	−0.66	−0.67	−0.72	−0.57	−0.51	−0.64	−0.83	−0.51	−1.05	−0.82	N/A	−1.02
15	−0.12	−0.51	−0.64	−0.32	−0.25	−0.52	−0.46	−0.27	−0.50	N/A	N/A	−0.70
16	−0.18	−0.42	−0.40	−0.43	−0.20	−0.42	−0.57	−0.93	−0.27	N/A	−1.01	0.21
17	−1.15	−0.79	−0.77	−1.01	−0.67	−1.13	−1.13	−1.16	−0.60	−1.57	−1.34	N/A
18	−1.57	−1.12	−1.14	−0.98	−0.96	−1.08	−1.17	−1.26	−0.92	−1.80	N/A	−1.12
19	−1.54	−1.10	−1.15	−1.21	−1.77	−1.44	−1.58	−1.51	−1.41	N/A	−1.18	N/A
20	−1.12	−0.50	−0.60	−0.62	−0.73	−0.61	−0.78	−0.54	−1.03	−0.83	−0.62	−0.90
21	−0.73	−0.34	−0.51	−1.49	−0.64	−0.57	−0.41	−0.43	−0.38	−0.72	−0.34	N/A
22	−0.27	−0.11	−0.28	−0.37	−0.77	−0.60	−0.50	−0.34	−0.87	−0.88	−0.51	−0.49
23	−0.28	−0.17	−0.38	−0.40	−0.54	−0.46	−0.64	−0.34	−0.70	−0.94	N/A	−0.72
24	−0.46	−0.41	−0.49	−0.16	−0.26	−0.34	−0.25	−0.08	−0.53	0.03	−0.05	−0.37
25	−0.24	−0.14	−0.29	0.24	0.04	−0.15	0.08	−0.07	−0.30	0.15	−0.02	−0.16
26	−0.02	0.00	−0.06	−0.05	−0.26	−0.02	−0.07	0.03	−0.30	0.37	0.37	0.53
27	0.69	0.02	−0.07	0.27	0.09	0.11	0.21	0.00	0.19	0.33	N/A	0.45
28	0.44	−0.06	−0.10	0.28	0.12	0.29	0.18	−0.38	0.16	0.36	N/A	0.54
29	0.14	0.29	0.19	0.31	0.34	0.26	0.43	−0.35	0.32	0.36	0.07	0.66
30	0.22	0.14	−0.01	0.37	−0.19	0.05	0.09	−0.13	0.22	0.10	N/A	0.72
31	−0.08	−0.13	−0.10	−0.11	0.33	−0.10	−0.07	0.26	−0.17	N/A	N/A	−0.28
32	0.37	0.41	0.48	0.47	0.57	0.70	0.46	0.39	0.49	N/A	N/A	0.58
33	1.01	0.58	0.65	0.77	0.33	0.64	0.71	0.54	0.81	N/A	N/A	0.71
34	0.90	0.63	0.70	0.87	0.33	0.81	0.79	0.55	0.82	N/A	N/A	0.75
35	−0.15	−0.07	0.00	−0.21	−0.15	−0.20	−0.20	0.21	−0.28	N/A	N/A	0.34
36	−0.17	−0.02	0.04	−0.16	−0.15	−0.07	−0.09	0.21	−0.17	N/A	N/A	0.38
37	0.97	0.58	0.65	0.77	0.32	0.65	0.71	0.54	0.81	N/A	N/A	0.72
38	1.04	0.42	0.70	0.76	0.20	0.86	0.69	0.53	0.80	N/A	N/A	0.73
39	0.08	0.45	0.43	0.44	0.29	0.36	0.40	0.46	0.17	N/A	N/A	−0.15
40	0.40	0.34	0.58	0.54	−0.13	0.39	0.52	0.38	0.55	N/A	N/A	0.40
41	0.00	−0.08	0.01	0.18	0.31	0.46	0.10	0.41	0.09	N/A	N/A	0.62
42	0.24	0.48	0.31	0.51	−0.01	0.31	0.34	0.47	0.12	N/A	N/A	−0.22
43	−0.16	−0.10	−0.02	−0.31	−0.14	−0.58	−0.26	0.16	−0.38	N/A	N/A	0.24
44	0.13	−0.01	0.03	0.28	0.12	0.56	0.25	0.42	0.00	N/A	N/A	0.34
45	0.35	0.28	0.38	0.62	0.34	0.36	0.51	0.45	0.39	N/A	N/A	N/A
46	0.81	0.72	0.77	0.87	0.34	0.71	0.81	0.55	0.83	N/A	N/A	0.74
47	0.80	0.66	0.72	0.77	0.34	0.57	0.73	0.54	0.82	N/A	N/A	0.71
48	0.42	0.43	0.32	0.49	0.00	0.51	0.43	0.46	0.15	N/A	N/A	0.07
49	0.93	0.63	0.73	0.70	0.29	0.81	0.64	0.52	0.80	N/A	N/A	0.64
50	0.03	0.20	0.19	0.05	0.34	N/A	N/A	0.17	0.51	N/A	N/A	N/A
51	0.76	0.61	0.68	0.66	0.34	0.84	0.65	0.53	0.80	N/A	N/A	0.67
52	−0.16	−0.02	0.05	−0.24	−0.14	−0.19	−0.18	0.18	−0.27	N/A	N/A	0.27
53	−0.15	−0.27	−0.22	−0.25	0.33	−0.12	−0.15	0.21	−0.29	N/A	N/A	−0.32
54	0.27	0.26	0.28	0.49	0.27	0.38	0.44	0.50	0.31	N/A	N/A	−0.35
55	0.37	0.47	0.54	0.57	0.51	0.64	0.54	0.40	0.58	N/A	N/A	0.64
56	0.34	0.27	0.36	0.52	0.34	0.81	0.54	0.38	0.60	N/A	N/A	−0.10
57	−0.50	−0.39	−0.52	−0.56	−0.59	−0.68	N/A	N/A	N/A	N/A	N/A	N/A
58	0.12	−0.56	−0.41	N/A	−0.76	N/A	N/A	N/A	N/A	N/A	N/A	N/A
59	0.08	0.36	0.55	−0.31	−0.12	N/A	N/A	N/A	N/A	N/A	N/A	N/A
60	−1.10	−0.11	−0.19	−0.26	−0.86	−0.76	N/A	N/A	N/A	N/A	N/A	N/A
61	1.44	1.28	1.46	N/A	0.26	1.31	N/A	N/A	N/A	N/A	N/A	N/A
62	−0.22	0.19	0.14	−0.11	−0.26	−0.28	N/A	N/A	N/A	N/A	N/A	N/A
63	−0.70	−0.54	−0.21	N/A	−0.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A
64	−0.72	−0.93	−1.13	−1.49	−1.38	N/A	N/A	N/A	N/A	N/A	N/A	N/A
65	0.30	0.13	−0.06	0.46	1.05	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 4 (*continued*)

Compd no.	Expt logBB	V1	V4	Hou and Xu[26]	Rose et al.[24]	Hutter[27]	Clark[18]	LomBardo et al.[15]	Norinder et al.[17]	Kelder et al.[19]	Young et al.[14]	Keseru and Molnar[22]
66	0.39	0.25	0.06	−0.04	0.02	N/A	N/A	N/A	N/A	N/A	N/A	N/A
67	0.24	1.22	1.10	N/A	0.89	N/A	N/A	N/A	N/A	N/A	N/A	N/A
68	0.04	−0.18	−0.22	N/A	0.64	N/A	N/A	N/A	N/A	N/A	N/A	N/A
69	1.00	0.60	0.81	0.24	0.24	N/A	N/A	N/A	N/A	N/A	N/A	N/A
70	0.36	−0.01	−0.11	N/A	0.29	N/A	N/A	N/A	N/A	N/A	N/A	N/A
71	1.00	0.25	0.99	0.62	0.36	N/A	N/A	N/A	N/A	N/A	N/A	N/A
72	0.55	−0.21	−0.10	N/A	−0.29	0.43	N/A	N/A	N/A	N/A	N/A	N/A
73	1.06	1.13	1.02	N/A	0.88	N/A	N/A	N/A	N/A	N/A	N/A	N/A
74	1.23	0.99	0.89	N/A	0.97	N/A	N/A	N/A	N/A	N/A	N/A	N/A
75	0.00	0.11	−0.06	0.02	−0.16	N/A	N/A	N/A	N/A	N/A	N/A	N/A
76	−0.16	0.28	0.25	N/A	−0.37	−0.36	N/A	N/A	N/A	N/A	N/A	N/A
77	−1.30	−1.59	−1.57	−1.41	−1.06	N/A	N/A	N/A	N/A	N/A	N/A	N/A
78	−0.18	0.15	0.09	N/A	−0.14	N/A	N/A	N/A	N/A	N/A	N/A	N/A
79	−0.10	0.43	0.30	N/A	0.34	N/A	N/A	N/A	N/A	N/A	N/A	N/A
80	−0.29	−0.41	−0.45	N/A	−0.32	N/A	N/A	N/A	N/A	N/A	N/A	N/A
81	−0.31	−0.45	−0.47	N/A	−0.64	−0.73	N/A	N/A	N/A	N/A	N/A	N/A
82	0.36	0.37	0.09	0.68	0.28	N/A	N/A	N/A	N/A	N/A	N/A	N/A
83	0.03	0.30	0.23	−0.07	0.34	N/A	N/A	N/A	N/A	N/A	N/A	N/A
84	0.60	0.45	0.55	N/A	0.34	N/A	N/A	N/A	N/A	N/A	N/A	N/A
85	−0.06	−0.23	−0.30	N/A	0.06	N/A	N/A	N/A	N/A	N/A	N/A	N/A
86	−1.26	−0.29	−0.45	N/A	−0.77	N/A	N/A	N/A	N/A	N/A	N/A	N/A
87	−0.75	−1.41	−0.78	N/A	−1.12	−0.98	N/A	N/A	N/A	N/A	N/A	N/A
88	0.61	−0.25	−0.38	N/A	−0.53	N/A	N/A	N/A	N/A	N/A	N/A	N/A

N/A—not available.

value of 0.03, whereas it was predicted as BBB− by Hutter[27] and Cruciani and co-workers.[20] (3) Mequitazine, **140**, is a first generation antihistamic drug reported[49] to have BBB− behaviour; however, it is predicted as BBB+ by the model V4 and also by others.[20,27] (4) Another example of the utility of our model is demonstrated based on the prediction of the antibacterial quinalones,[50,51] **124**, **127**, **130**, **135**, **141**–**143** and **147**. Even though there are different opinions on the BBB transport of quinalones,[50] it is well known that they exhibit a lower concentration in brain than in most other tissues,[51] implying a low BBB permeability. Significantly, our models have classified all the quinalones, as BBB− and this observation agrees with the experimental data.[51]

The increasing order of the BBB− permeation of the quinalones is in the sequence, Sparfloxacin, **147** < Lomefloxacin, **135** < Enoxacin, **130** < Norfloxacin, **141** < Ciprofloxacin, **124** < Ofloxacin, **142** < Perfloxacin, **143** < Difloxacin, **127**. (5) Loperamide, **136**, is a well documented CNS− compound, but the model, V4, predicts it to have BBB+ behaviour, an observation similar to that of others.[20,27] Experimental data[52] suggest that the lack of CNS activity of Loperamide is due to its active removal from the brain by P-glycoprotein at the BBB. This explains the difference in the predicted and observed BBB behaviour of Loperamide. Even though the models fail to predict correctly the BBB profile of a few compound, the results of the external predictions (Test set I, II and III) using the models V1 and V4 can be considered extremely good and these models can be used in virtual screening applications.

Table 5. Virtual screening results

Compd no.	Compound name	Obs BBB sign	Pred sign V4	Pred BBB sign Hutter
117	Aldesterone	–	–	–
118	Astemizole	–	–	–
119	Atenolol	–	–	–
120	Carbidopa	–	–	–
121	Carebastine	–	+	–
122	Carmoxirol	–	+	–
123	Cetirizine	–	–	+
124	Ciprofloxacin	–	–	–
125	Cortisol	–	–	–
126	Desa_Loratidine	–	+	+
127	Difloxacin	–	–	–
128	Dopamine	–	–	–
129	Ebastine	–	+	+
130	Enoxacin	–	–	–
131	Fexofenadine	–	+	–
132	Furosemide	–	–	–
133	Isoxicam	–	–	–
134	Levodopa	–	–	–
135	Lomefloxacin	–	–	–
136	Loperamide	–	+	+
137	Loratadine	–	+	+
138	Mefloquine	–	+	+
139	Meloxicam	–	–	–
140	Mequitazine	–	+	+
141	Norfloxacin	–	–	–
142	Oflaxacin	–	–	–
143	Pefloxacin	–	–	–
144	Pirenzepine	–	–	–
145	Piroxicam	–	–	–
146	Salbutamol	–	–	–
147	Sparfloxacin	–	–	–
148	Tenoxicam	–	–	–
149	Terfenadine	–	+	–

(*continued on next page*)

Table 5 (*continued*)

Compd no.	Compound name	Obs BBB sign	Pred sign V4	Pred BBB sign Hutter
150	EMD60400	–	–	N/A
151	GR94839_A	–	–	N/A
152	GR94839_B	–	–	N/A
153	GR94839_C	–	–	N/A
154	GR94839_D	–	–	N/A
155	GR94839_E	–	–	N/A
156	GR94839_F	–	–	N/A
157	GR94839_G	–	–	N/A
158	GR94839_H	–	–	N/A
159	GR94839_I	–	–	N/A
160	GR94839_L	–	–	N/A
161	GR94839	–	–	N/A
162	ICI204448	–	–	N/A
163	ICI205640	–	–	N/A
164	SB204454	–	–	N/A
165	SB204457	–	–	N/A
166	SB204459	–	–	N/A
167	SB205563	–	–	N/A
168	SB205605	–	–	N/A
169	Apomorphine_R	+	+	–
170	Clobazam	+	+	+
171	Diphenhydramine	+	+	+
172	Doxylamine	+	+	+
173	Estradiol	+	+	+
174	Haloperidol	+	+	–
175	Morphine	+	–	+
176	Naltrexone	+	–	–
177	Nordazepam	+	–	+
178	Perphenazine	+	+	+
179	Progesterone	+	+	+
180	Promethazine	+	+	+
181	Rivastigmine	+	+	+
182	Roxindole	+	+	–
183	Testosterone	+	+	+
184	Thiopental	+	+	–
185	BRL52537	+	+	N/A
186	BRL52580	+	+	N/A
187	BRL52656	+	+	N/A
188	BRL52871	+	+	N/A
189	BRL52974	+	+	N/A
190	BRL53080	+	+	N/A
191	BRL53087	+	+	N/A
192	GR45809	+	+	N/A
193	GR85571	+	–	N/A
194	GR88377	+	+	N/A
195	GR89696	+	+	N/A
196	GR89696_et	+	+	N/A
197	GR89696_pr	+	+	N/A
198	GR91272	+	–	N/A
199	ICI197067	+	+	N/A
200	ICI199441	+	+	N/A
201	Levallorphan	+	+	N/A
202	Nalorphine	+	–	N/A
203	RP60180	+	+	N/A
204	Sankyo	+	+	N/A
205	SB201708	+	–	N/A
206	SB204484	+	–	N/A
207	Tifluadom	+	–	N/A
208	U50488	+	+	N/A

3. Conclusion

In this paper, we have described global predictive models based on three and four descriptors, for blood–brain barrier (BBB) permeation using the largest dataset of 116 diverse drugs and drug-like compounds and 324 molecular descriptors as of date, to the best of our knowledge. For this aim, VSMP, a systematic variable selection method along with multiple linear regression (MLR), is employed for the first time. Unlike many of the reported approaches, we have reported multiple models with various combinations of descriptors. The reported models possess excellent predictive power (both internal and external) and their utility for virtual screening applications is demonstrated. Finally, as the models reported herein are based on computed properties, they appear as valuable tools for virtual screening, where selection and prioritization of candidates is required.

4. Methods

4.1. Datasets and BBB permeation property

A database of 116 drugs and drug-like molecules with known logBB data from various literature sources[14–28] is created (Supplementary material) and the members of this database constitute training set, test sets I and II.

4.2. Training set

Eighty-eight drugs and drug-like molecules constitute the training set. The names and their corresponding logBB values are shown in Table 2. The training set consists of compounds with a wide range of molecular size and complexity, ranging from small nonpolar molecules such as N_2 to small organics such as Theophylline and caffeine to the larger drugs Indinavir and Verapamil. The logBB values range from −2.15 (BBCPD11) to 1.64 (Org12962) and the molecular weights range from 28.0134 (Nitrogen) to 599.8148 (Indinavir).

4.3. Test sets I and II

A total of 28 compounds were set aside as test set for validation studies. All the data used in the training set, test sets I and II are based on in vivo measurements obtained from rat studies.

4.4. Test set III

Test set III represents 92 drugs and drug-like compounds, like, Quinalones, antibacterials, etc., that are structurally different are used to perform virtual screening based on the models reported in the present study. They are selected from the report of Crivori et al.[20] and of the 92 compounds, 52 are BBB− logBB compounds and 40 are BBB+ compounds. '+' sign was assigned if their logBB value is greater than '0' and '–' sign was assigned if their logBB value is less than '0'. It contains different chemical classes such as antipsycotics, dopaminergics, antihistamines of first and second generation, anxiolotics, hypnotics, opioids, β-blockers, CCK antagonists, acetylcholineesterase inhibitors, quinolone antibacterials, NSAIDs, antiparkinsonians, iron chelators and hormones.

4.5. VSMP

All subset regression (ASR) is one of the variable selection procedures that can select the best descriptor subset from the given dataset, but is computationally intensive. Liu et al.[32] introduced two controllable parameters, namely, the inter-correlation coefficient between pairs of independent descriptors (r_{int}) and cross-validation correlation coefficient (Q^2) into ASR, to accelerate the speed of computation. We have implemented the algorithm and in our study we have set the inter-correlation coefficient between descriptors as 0.75.

4.6. Software

All the programs used in the present study are developed in-house, are a part of the in-house product, 'Biosuite',[53] and are used to perform the following: (1) draw the 2D structures of the compounds, (2) calculate the molecular descriptors, (3) descriptor selection and (4) multiple linear regression and validation.

4.7. Molecular descriptors calculation

A total of 324 descriptors[34–37] are calculated using the QSAR module of the in-house software, 'Biosuite',[53] of which 7 are physicochemical descriptors, 7 are geometrical descriptors, 11 are structural descriptors and the remaining 299 are topological descriptors. The calculated descriptors are stored as a text file, formatted into a table using an in-house program.

4.8. Cross-validations

We have used 'Leave-One-Out (LOO)' method,[38] the simplest and commonly used cross-validation approach, in our studies. In this approach, the property value for a given compound in the training set is predicted using the regression equation derived from the data of the remaining compounds. The PRESS (predictive residual sum of squares) statistic is computed using the formula

$$\mathrm{PRESS} = \sum_{i=1}^{N} (y_i - y_i')^2$$

where y_i' is the predicted logBB value calculated after eliminating the ith compound and y_i is the experimental logBB value. The Q value is given by

$$Q = \sqrt{1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

Acknowledgements

We thank Drs. Rajgopal Srinivasan and B. Gopalakrishnan, Advanced Technology Centre, Tata Consultancy Services Limited and Mr. Akash Khandelwal for constructive discussions during the preparation of the manuscript.

Supplementary data

The following information are provided in the Supplementary material: (1) Tables of structures of the 116 compounds used in the analysis and their corresponding logBB values. (2) Ninety-two compounds used for virtual screening studies and their corresponding logBB sign. (3) Results of the cross-validations and external validations of the models V2, V3, V5 and V6. (4) Confidence score of logBB prediction of compounds in the training set. (5) Plots of cross-validated logBB values using models V2, V3, V5 and V6 versus experimental logBB values. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2005.01.061.

References and notes

1. Cavalla, D. *Pharm. Sci. Technol. Today* **1998**, 189–190.
2. Kennadt, T. *Drug Discovery Today* **1997**, *2*, 436–444.
3. Eddy, E. P.; Maleef, B. E.; Hart, T. K.; Smith, P. L. *Adv. Drug Delivery Rev.* **1997**, *23*, 185–198.
4. Reichel, A.; Begley, D. J. *Pharm. Res.* **1998**, *15*, 1270–1274.
5. Colmenarejo, G.; Alvarez-Prdraglio, A.; Lavandera, J.-L. *J. Med. Chem.* **2001**, *44*, 4370–4378.
6. van de waterbeemd, H. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*(1), 33–43.
7. Viswanathan, V. N.; Balan, C.; Hulme, C.; Cheetham, J. C.; Sun, Y. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*(3), 400–406.
8. Chaturvedi, P. R.; Decker, C. J.; Odinecs, A. *Curr. Opin. Chem. Biol.* **2001**, *5*, 452–463.
9. Banik, G. M. *Curr. Drug Discovery* **2004**, 31–34.
10. Segall, Matthew D. *Future Drug Discovery* **2004**, 81–84.
11. De Vries, H. E.; Kuiper, J.; De Boer, A. G.; Van Berkel, T. J. C.; Breimer, D. D. *Pharmacol. Rev.* **1997**, *49*, 143–155.
12. van de waterbeemd, H.; Gifford, E. *Nature Rev. Drug Discovery* **2003**, *2*, 192–204.
13. Clark, D. E.; Grootenhuis, P. D. J. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*(3), 382–390.
14. 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*, 656–671.
15. Lombardo, F.; Blake, J. F.; Curatolo, W. J. *J. Med. Chem.* **1996**, *39*, 4750–4755.
16. Abraham, M. H.; Takacs-Novak, K.; Mitchell, R. C. *J. Pharma. Sci.* **1997**, *86*, 310–315.
17. Norinder, U.; Sjoberg, P.; Osterberg, T. *J. Pharma. Sci.* **1998**, *87*, 952–959.
18. Clark, D. E. *J. Pharma. Sci.* **1999**, *88*, 815–821.
19. Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. *Pharma. Res.* **1999**, *16*, 1514–1519.
20. Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. *J. Med. Chem.* **2000**, *43*, 2204–2216.
21. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714–3717.
22. Keseru, G. M.; Molnar, L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 120–128.
23. Liu, R.; Sun, H.; So, S.-S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1623–1632.
24. Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651–666.
25. Hou, T.; Xu, X. *J. Mol. Model.* **2002**, *8*, 337–349.

26. Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.
27. Hutter, M. C. *J. Comput. Aided Mol. Des.* **2003**, *17*, 415–433.
28. Abraham, M. H. *Eur. J. Med. Chem.* **2004**, *39*, 235–240.
29. Wessel, M. D.; Jurs, P. C.; Tolan, J.; Muskal, S. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.
30. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294.
31. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393–401.
32. Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964–969.
33. Izrailev, S.; Agrafiotis, D. K. *SAR QSAR Environ. Res.* **2002**, *13*, 417–423.
34. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, 2000.
35. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley Interscience: New York, 2000.
36. *From Chemical Topology to 3D Molecular Geometry*; Balaban, A. T., Ed.; Plenum: New York, 1997.
37. Kubinyi, H.; Folkers, G.; Martin, Y. T. In *3D QSAR in Drug Design*; Kluwer ESCOM, 1996–1998; Vol. 1–3.
38. Wold, S. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.
39. Hardman, J. G.; Limbird, L. E. *Goodman & Gillman*, 10th ed.; McGraw-Hill, 2001; pp 233–234.
40. Hardman, J. G.; Limbird, L. E. *Goodman & Gillman*, 10th ed.; McGraw-Hill, 2001; pp 653–654.
41. Hardman, J. G.; Limbird, L. E. *Goodman & Gillman*, 10th ed.; McGraw-Hill, 2001; p 653.
42. Hardman, J. G.; Limbird, L. E. *Goodman & Gillman*, 10th ed.; McGraw-Hill, 2001; pp 413–419.
43. Hardman, J. G.; Limbird, L. E. *Goodman & Gillman*, 10th ed.; McGraw-Hill, 2001; pp 589–590.
44. Walters, W. P.; Stahl, M. T.; Murcko, M. A. *Drug Discovery Today* **1998**, *3*, 160–178.
45. Fox, S.; Farr-Jones, S.; Yund, M. A. *J. Biomol. Screening* **1999**, *4*, 183–186.
46. Oprea, T. I. *Molecules* **2002**, *7*, 51–62.
47. Bajorath, J. *Nature Rev. Drug Discovery* **2002**, *1*, 882–894.
48. Fischer, S.; Renz, D.; Schaper, W.; Karliczek, G. F. *Anesthesiology* **1995**, *82*, 451–458.
49. Basak, S. C.; Gute, B. D.; Drewes, L. R. *Pharma. Res.* **1996**, *13*, 775–778.
50. Ooie, T.; Terasaki, T.; Suzuki, H.; Suglyama, Y. *J. Pharmaco. Exper. Ther.* **1997**, *283*, 293–304.
51. Jaehde, U.; Goto, T.; de Boer, A. G.; Breimer, D. D. *Eur. J. Pharmacol. Sci.* **1993**, *1*, 49–55.
52. Schinkel, A. H.; Wagenaar, E.; Carla, A. A. M.; van deemter, L. *J. Clin. Invest* **1981**, *2*, 103–123.
53. Tata Consultancy services, 'TCS Launches TATA Bio-Suite, Bioinformatics software to facilitate life science research through information technology', Press Release, June 7th, 2004.