ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors

T. J. Hou and X. J. Xu*

College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Received July 2, 2003

In this paper, QSPR models were developed for in vivo blood-brain partitioning data (log*BB*) of a large data set consisting of 115 diverse organic compounds. The best model is based on three descriptors: *n*-octanol/water partition coefficient calculated using the SLOGP approach, log*P*; high-charged polar surface areas based on the Gasteiger partial charges, *HCPSA*, and the excessive molecular weight larger than 360, MW_{360}. The model bears good statistical significance, $n = 78$, $r = 0.88$, $q = 0.86$, $s = 0.36$, $F = 81.5$. The actual prediction potential of the model was validated through two external validation sets of 37 diverse compounds. The predicted results demonstrate that the model bears better prediction potential than many other models and can be used for log*BB* estimations for drug and drug-like molecules. Comparison of several log*P* calculation approaches suggests that log*P* calculated by SLOGP can be used as a significant descriptor for the prediction of molecular transport properties because SLOGP gives the most similar results with CLOGP. The QSPR model indicates that larger polar surface areas have a more negative contribution to log*BB*, but the absolute partial charges on the atoms surrounded by the polar surfaces should be larger than 0.10|e|. Meanwhile, tight junction membranes limit the size of hydrophilic molecules that can cross the membrane with a molecular weight of approximately 360, because when a molecule's weight is larger than 360 it shows a negative contribution to log*BB*. The computations of molecular surface, partial charges, log*P*, and log*BB* have been accomplished using a program called Drug-BB. Moreover, to improve the efficiency of the computations of log*P*, we made an extensive reparametrization of SLOGP, and the newly developed SLOGP model is only based on simple atomic addition. Further, we developed a set of parameters to calculate the topological polar surface area (*TPSA*), thus the high-charged topological polar surface area (*HCTPSA*) could be estimated from the 2D connection information of a molecule. Adopting the new strategies, the estimations of log*P*, *HCTPSA*, and log*BB* are only based on the topological structure of a molecule and therefore, can be used for fast screening of virtual libraries having millions of molecules.

INTRODUCTION

The development of combinatorial chemistry and high-throughput screening (HTS) gives us more opportunities to synthesize and gives a rapid and effective assay to thousands upon thousands of compounds in a very short period. As discovery chemistry produces increased numbers of potential drug compounds, the use of ADME (absorption, distribution, metabolism, and excretion) properties is becoming increasingly important in the drug selection and promotion process.[1] The significant failure rate of drug candidates in late stage development is driving the need for predictive tools that can eliminate inappropriate compounds before substantial time and money are invested in testing. It has been estimated that about 50% of such failures are caused by ADME/Tox deficiencies.[2] Apparently developing effective computational models to screen ADME properties is very promising as an early screen for potential drug candidates and for the design of combinatorial libraries.

A good example that exemplifies the great utility of a predictive computational model in drug discovery is a model for predicting blood-brain barrier (*BBB*) penetration. In the case of effective central nervous system (CNS) acting drugs, the knowledge of the penetration of drugs through *BBB* is critical to screen potential therapeutic agents and to improve the side effect profile of drugs with peripheral activity.[3] *BBB* is a complex physical and biochemical interface, which is composed of tightly jointed blood capillary endothelial cells. The extent to which drug molecules cross from the blood into the brain is governed by two physiologically and anatomically related systems, *BBB* and the blood-cerebral spinal fluid (CSF) barrier, which form two pathways by which drug compounds partition between plasma and brain tissue. At the molecular level, the principal component of the barrier is the lipid bilayer of the capillary endothelial cell membrane, through which compounds have to diffuse to reach the brain. The membranes involved are tight junction membranes by brain parechymal cells. Tight junction membranes limit the size of hydrophilic molecules that can cross the membrane by paracellular diffusion. The vast majority of substances that penetrate a tight junction barrier are lipophilic molecules that cross by a transcellular route.[4] Experimental data have shown that lipophilic compounds, along with water and small polar molecules, can cross both the blood-brain and blood-CSF barriers. Hydrophilic organic molecules, including plasma proteins and larger polar molecules, cannot penetrate well.

* Corresponding author e-mail: xiaojxu@chem.pku.edu.cn.

10.1021/ci034134i CCC: $25.00
Published on Web 10/09/2003

In experiments, the relative affinity for the blood or brain tissue can be expressed in terms of the blood-brain partition coefficient, $\log BB = \log(C_{\text{brain}}/C_{\text{blood}})$, which C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively. Both in vivo and in vitro experiments have been conducted that calculated log*BB*. However, both these methodologies are laborious, expensive, and time-consuming and require a sufficient quantity of the pure compounds, often in radiolabeled form to obtain reliable experimental data, hence not amendable to high-throughput screening of therapeutic candidates.[5,6] So theoretical and computational methodologies to predict log*BB* would have a great impact on drug research and development.

Numerous attempts have been attempted to correlate *BBB* transport with physicochemical descriptors, in particular with the octanol−water partition coefficient, log*P*. Young et al. proposed a correlation between log*BB* and ΔlogP (see eq 1).[7] Δlog*P* is defined as the difference between $\log P_{\text{ow}}$ and $\log P_{\text{cychw}}$, where P_{ow} and P_{cyclw} are the octanol/water and cyclohexane/water partition coefficient, respectively. However, in some cases, log*P* shows poor correlation with log*BB*. For example, Ter Laak et al. found that the brain permeability of a series of structurally diverse histamine H1 receptor antagonists was better explained by $\log D_{\text{oct}}$ rather than by log*P*.[8] At present it is well-known that log*BB* cannot be effectively predicted only based on the hydrophobic parameters.

$$\log BB = 1.889 - 0.485\Delta\log P \qquad (1)$$

$$(n = 20,\ r = 0.831,\ s = 0.439,\ F = 40.23)$$

Kaliszan et al. reestablished the correlation of log*BB* with log*P* and refined it, employing the molecular weight as an additional descriptor of molecular bulkiness (see eq 2).[9] The authors indicated that a molecular bulkiness descriptor should be used to better account for nonspecific dispersive properties of molecules.

$$\log BB = -\,0.088 + 0.272\Delta\log P - 0.00112\mathrm{M}_m \qquad (2)$$

$$(n = 33,\ r = 0.947,\ s = 0.126,\ F = 131.1)$$

In addition to hydrophobic parameters, the descriptors related with molecular surface properties, molecular size, and hydrogen bond formation have also been found as important contributors to log*BB*. Among all these descriptors, polar surface area (*PSA*) may be the most important one. Using *PSA* as the only descriptor, Kelder et al. obtained the following simple equation for a training set of 45 compounds:[10]

$$\log BB = 1.33 - 0.032PSA \qquad (3)$$

$$(n = 45,\ r^2 = 0.84,\ F = 229)$$

A similar equation was also developed by Clark based on a training set of 55 compounds:[11]

$$\log BB = 0.55 - 0.016PSA \qquad (4)$$

$$(n = 55,\ r^2 = 0.71,\ F = 128,\ s = 0.41)$$

In an effort to account for hydrophobic contributions, Clark introduced $\log P_{\text{oct}}$ as an additional descriptor:[11]

$$\log BB = 0.139 - 0.148PSA + 0.152\mathrm{Clog}\,P \qquad (5)$$

$$(n\ = 55,\ r^2 = 0.79,\ s = 0.35,\ F = 95.8)$$

$$\log BB = 0.131 - 0.145PSA + 0.172\mathrm{Mlog}\,P \qquad (6)$$

$$(n\ = 55,\ r^2 = 0.77,\ s = 0.37,\ F = 86.0)$$

Some research observed that introducing a descriptor about hydrogen-bonding ability could improve the quality of the QSPR models. Feher et al. proposed the following regression model[12]

$$\log BB = 0.4275 - 0.0017PSA - 0.1092\log P - 0.3873n_{acc} \qquad (7)$$

$$(n\ = 61,\ r = 0.854,\ s = 0.424,\ F = 51)$$

where n_{acc} is the number of hydrogen-bond acceptors.

Abraham and co-workers constructed the following equation using a fragment-based scheme[13,14]

$$\log BB = 0.055 - 0.507\sum\alpha_2^H - 0.500\sum\beta_2^H + 0.023\log P \qquad (8)$$

$$(n\ = 49,\ r = 0.949,\ s = 0.201,\ F = 136.1)$$

$$\log BB = -\,0.038 - 0.715\sum\alpha_2^H - 0.698\sum\beta_2^H + 0.198R_2 - 0.687\pi_2^H + 0.995V_x \qquad (9)$$

$$(n\ = 57,\ r = 0.952,\ s = 0.197,\ F = 99.2)$$

where R_2 is an excess molecular refraction; π_2^H is the dipolarity/polarizability parameter; $\sum\alpha_2^H$ and $\sum\beta_2^H$ are the solute hydrogen-bond acidity and basicity, respectively; and V_x is the McGowan characteristic volume. A potential problem of their models is that the descriptors are not easily estimated for structurally diverse drug candidates.

Lomardo et al. established a correlation between log*BB* and solvation free energy calculated using semiempirical quantum chemical calculations[15]

$$\log BB = 0.43 + 0.054\Delta G_w \qquad (10)$$

$$(n = 55,\ r = 0.82,\ F = 108)$$

where ΔG_w is the free energy of solvation. This correlation provides an elegant means for good log*BB* prediction. However, computation of ΔG_w based on semiempirical calculations is time-consuming, and moreover, the precision of the current methods for ΔG_w prediction is questionable especially for complicated organic molecules.

Recently, Kaznessis et al. applied Monte Carlo simulations of compounds in water to calculate such properties as the solvent-accessible surface area (*SASA*), the solute dipole, and the hydrophilic, hydrophobic, and amphiphilic components of *SASA*.[16] Using these parameters, they obtained the following equation

$$\log BB = 0.0458 - 0.234HBAC + 0.0015MVOL + 31.610HBAC \times HBDN^{1/2}/SASA \qquad (11)$$

$$(n = 76,\ r = 0.97,\ s = 0.173,\ F = 311.307)$$

where *HBAC* is the number of hydrogen-bond acceptors; *HBDN* is the number of hydrogen-bond donors; *MVOL* is the molecular weight; and *SASA* is the solvent accessible surface area. The correlation of eq 11 seems very good, but in order to gain a better correlation, the authors identified nine strong outliers and removed them.

All of the above equations were obtained using multiple linear regression (MLR). Meanwhile, many earlier models were based on a relatively small set of molecules and were not fully validated by external prediction sets. Moreover, in most papers, to improve the correlations, the authors usually removed some compounds from the training set subjectively.

Besides MLR, other statistical methods, especially partial least-squares (PLS), have been applied in the prediction of log*BB*. Norinder used MolSurf parametrization to calculate various properties related to the molecular valence region and combined it with PLS to develop a QSPR of log*BB* with three statistically significant components.[17] Luco also employed the PLS technique to develop a QSPR based on several topological and constitutional descriptors.[18] More recently, Crivori applied a new technique, Volsurf, to transform 3D molecule fields into descriptors and correlate them to the experimental permeation by PLS.[19] However, the PLS method generally appears to strip the QSPR from explicit physical insight, and the determination of the principle components of numerous physicochemical descriptors cannot be easily calculated for an arbitrary compound.

In our previous work of the relationships between log*BB* of 96 structurally diverse compounds with a great number of structurally derived descriptors, we found that log*P* was very crucial to log*BB*.[20] When we constructed the prediction models of log*BB*, the ALOGP approach proposed by Crippen et al.[21] was used to calculate log*P* by using the Cerius2 molecular simulations package.[22] The dependence on the commercial software prevents us from developing a procedure to estimate log*BB* as an automatic fashion. In this article, we like to present the results of our recent study by introducing a new log*P* parameter[23] and developing a simple predictive model of *BBB* penetration. For an efficient computational model, besides precision, speed should also be considered, because to be a high-throughput-screening tool it is expected to process a large number of compounds in a short period of time. To make the prediction of log*BB* more efficient, we made an extensive reparametrization of SLOGP, and the newly developed SLOGP model is only based on simple atomic addition. Further, through adopting the definition of topological polar surface area, the log*BB* calculation is only based on the topological structure of a molecule and can be accomplished very rapidly and easily.

METHODS

Data Set. The quality of a QSPR model depends strongly on the size and quality of the data set used. Variety experimental protocols have been applied in the measurement of log*BB*. To let the data used in this paper bear good comparability, all data are based on in vivo measurements taken from rat studies. Besides the compounds used in our previous work,[20] we added 20 new compounds collected from different articles.[24−27] The whole data set includes 115 diverse organic compounds, which was divided into a training set of 78 compounds (Table 1) and two test sets of 37 compounds. The first test set comprises 14 compounds from several literature sources (**B1**−**B14** in Table 2),[13,15,27] and the other one comprises 23 drugs collected by Salmien et al. (**C1**−**C23** in Table 2).[28]

The molecular geometries of all compounds were fully minimized using a molecular mechanism with a MMFF force field,[28] and the terminal condition was set as the RMS of potential energy smaller than 0.001 kcal.Å^{-1}.mol^{-1}. For these flexible compounds, the conformational analyses were performed to determine the most stable conformers. The models were then saved into two MACCS/sdf files named training.sdf and test_set.sdf for further analysis. The MACCS/sdf files are available in the Supporting Information.

Descriptors Used in MLR. (1) Hydrophobicity Descriptor. Traditionally, calculated values of the octanol/water partition coefficient have been used in the estimation of molecular transport properties. In this paper, a novel method, SLOGP, developed in our group, was used to calculate log*P* of organic molecules.[23] SLOGP estimates log*P* by summing the contribution of atom-weighted solvent accessible surface areas (*SASA*) and correction factors. Comparison of various log*P* models to the external test set demonstrates that our method bears very good accuracy and is comparable or even better than the fragment-based approaches.[23]

As being well-known, due to adopting a different training set and a different additive strategy a different log*P* prediction model may generate different predicted values for the same organic molecule. In this paper, to verify the validity of SLOGP, we compared SLOGP with CLOGP, the most popular method of log*P* prediction, using the data set studied here.[30] Meanwhile, the predictions by the other four methods, including ALOGP,[21] ALOGP98,[31] HINT,[32] and the Wildman model,[33] were compared with those of CLOGP systematically. ALOGP method is a direct, easy-to-computerize atomic constant approach to predict log*P* and is shown to exhibit a relatively robust performance. The difference between ALOGP and ALOGP 98 is the usage of different atomic hydrophobic parameters. log*P* values by ALOGP and ALOGP98 were obtained using the Cerius2 molecular simulation package.[22] HINT is a program designed for quantifying and visualizing hydrophobic and polar interactions. The log*P* calculation performed by HINT is based on the hydrophobic fragment constant approach of Hansch and Leo. In addition, there are a number of “factors” and application rules, which modify the total partition constant depending on a specific bond, chain, or branching, etc. log*P* values by CLOGP and HINT were calculated using the Sybyl molecular simulation package.[34] The Wildman’s model is based on simple atomic addition. The Wildman’s log*P* values were calculated using a homemade program. The atom typing rule and hydrophobic parameters for the Wildman model were obtained from ref 33.

(2) Hydrophilicity Descriptor. Due to the physical nature of the lipid bilayer, organic molecules, which can form favorable hydrogen-bonding or electrostatic interactions with the lipid bilayer, may have great difficulty with *BBB* penetration. To form an effective hydrogen bond or favorable electrostatic interactions, a molecule should have high electronegative atoms (oxygen, nitrogen, etc.) that are exposed on the molecular surface. Indeed it has been proven that polar surface area (*PSA*) is a very significant descriptor for drug transport properties such as human intestinal permeation and blood-brain barrier penetration. In this paper, the polar atoms include all oxygen atoms, nitrogen atoms,

Table 1. Compounds Used To Obtain the Training Set

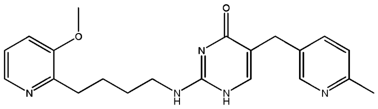
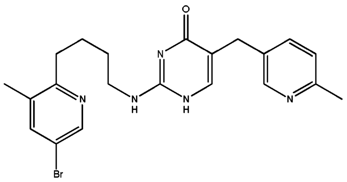
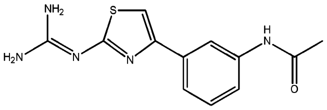
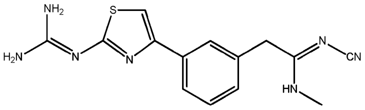
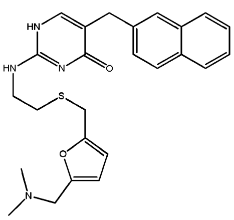
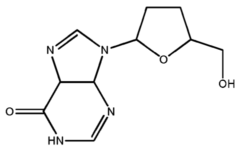
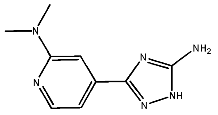
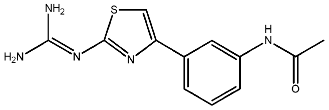
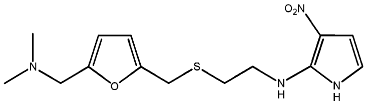
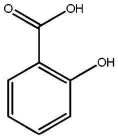
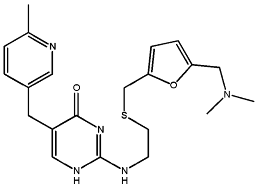
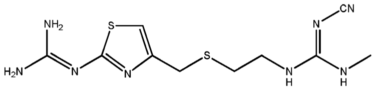
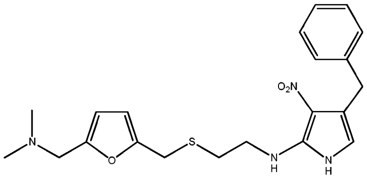
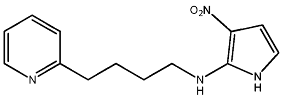
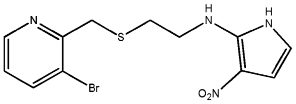
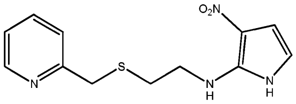
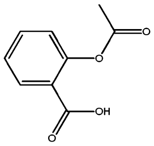
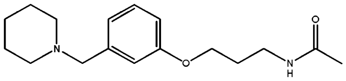
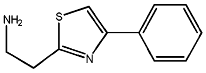
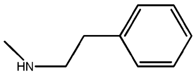
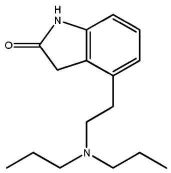
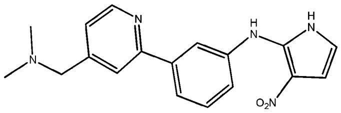
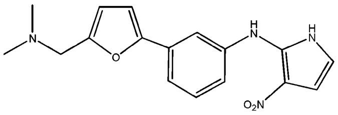
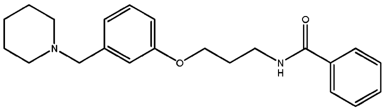
ID and name	ID and name	ID and name
1. icotidine 	2. temelastine 	3. BBcpd16 (guanidinothiazole der.) 
4. BBcpd58 (guanidinothiazole der.) 	5. SK&F 93319 	6. didanosine 
7. BBcpd10 	8. BBcpd57 (guaidinothiazole der.) 	9. BBcpd17 (ranitidine analog) 
10. salicylic acid 	11. lupitidine 	12. tiotidine 
13. BBcpd60 (ranitidine analog) 	14. zidovudine 	15. BBcpd12 (cimetidine derivative) 
16. BBcpd13 (cimetidine derivative) 	17. acetylsalicylic acid 	18. BBcpd20 (ranitidine analog) 
19. Y-G19 	20. Y-G14 	21. SKF101468 
22. BBcpd19 (ranitidine analog) 	23. BBcpd18 (ranitidine analog) 	24. BBcpd21 (ranitidine analog) 

Table I (Continued)

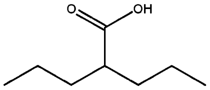
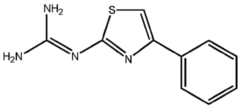
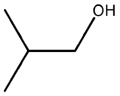
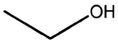
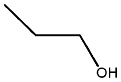
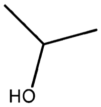
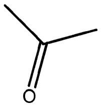
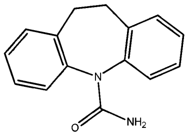
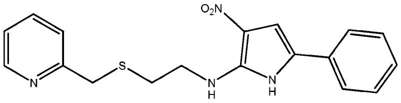
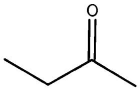
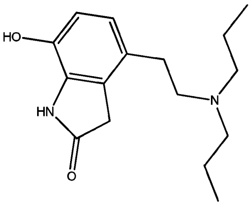
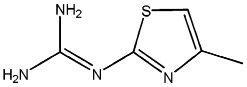
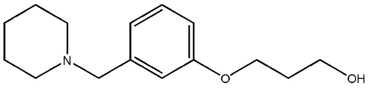
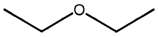
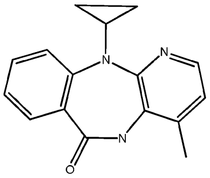
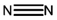
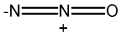
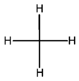
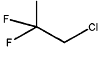
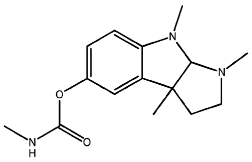
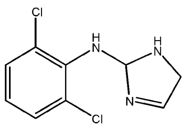
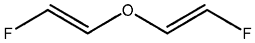
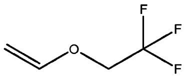
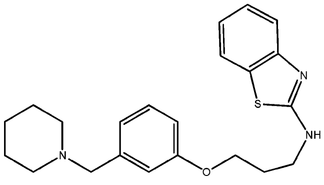
ID and name	ID and name	ID and name
25. valproic acid 	26. BBcpd15 (guanidinothiazole der.) 	27. 2-methylpropanol 
28. ethanol 	29. 1-propanol 	30. 2-propanol 
31. propanole 	32. carbamazepine 	33. BBcpd14 (cimetidine derivative) 
34. butanone 	35. SKF89124 	36. ICI 17148 
37. BBcpd22 (ranitidine analog) 	38. diethyl ether 	39. nevirapine 
40. nitrogen 	41. nitrous oxide 	42. methane 
43. 1,1,1-trifluoro-2-chloroethane 	44. physostigmine 	45. clonidine 
46. di-(2-fluroethene) ether 	47. fluroxene 	48. zolantidine (ranitidine analog) 

Table I (Continued)

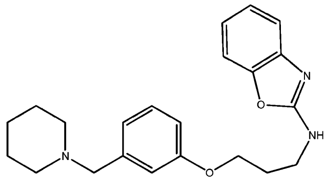
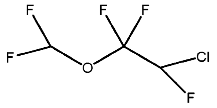
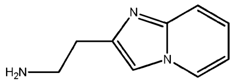
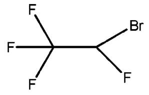
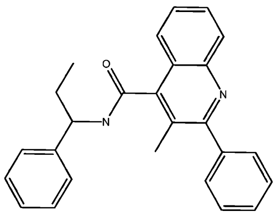
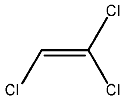
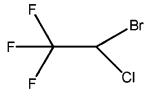
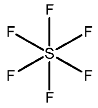
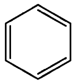
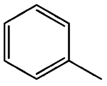
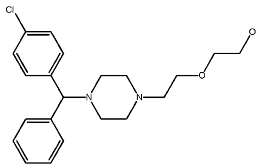
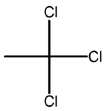
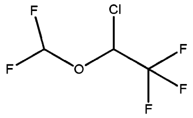
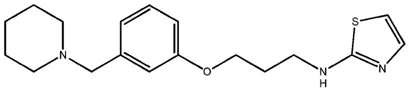
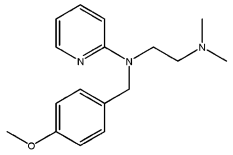
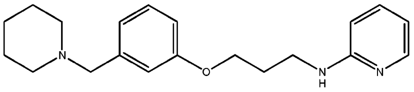
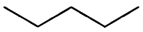
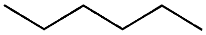
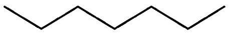
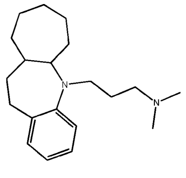
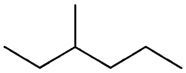
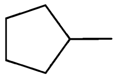
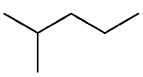
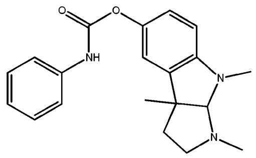
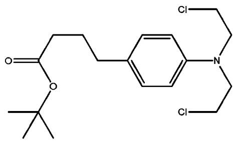
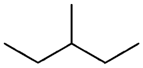
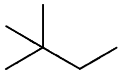
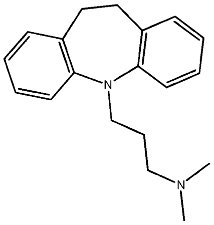
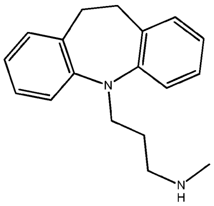
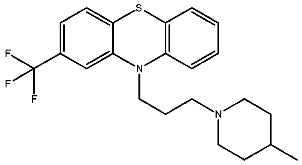
ID and name	ID and name	ID and name
49. BBcpd26 (ranitidine analog) 	50. enfluane 	51. Y-G20 
52. teflurane 	53. SB-222200 	54. trichloroethene 
55. halothane 	56. sulfur hexafluoride 	57. benzene 
58. toluene 	59. hydroxyzine 	60. 1,1,1-trichloroethane 
61. isofluane 	62. BBcpd24 (ranitidine analog) 	63. mepyramine 
64. BBcpd23 (ranitidine analog) 	65. pentane 	66. hexane 
67. heptane 	68. amitryptalline 	69. 3-methylhexane 
70. methylcyclopentane 	71. 2-methylpentane 	72. phenserine 

Table I (Continued)

ID and name	ID and name	ID and name
73. terbutylchlorambucil 	74. 3-methylpentane 	75. 2,2-dimethylbutane 
76. imipramine 	77. desipramine 	78. trifluoroperazine 

and sulfur atoms. By definition, *PSA* evaluation requires 3D molecular conformation and atomic surface area. Here, molecular solvent accessible surface areas were calculated using the MSMS program[35] and the probe radius was set to 0.5 Å, according to the definition in SLOGP, and thus the calculations of log*P* and *PSA* can share the same output of surface calculation. It should be noted that the surface areas of hydrogen connected with the polar atoms are included in *PSA*.

Generally, as a polar atom, it should be highly electronegative and possess high charge density. If the charge density on an oxygen atom or a nitrogen atom is very low, this atom may not produce a strong hydrogen bond or favorable electrostatic interactions with other polar atoms. Thus, to make a more close connection between the polar atom and the partial charge, we used a new definition named "high-charged polar atom". According to our definition, only polar atoms with high charge densities belong to high-charged polar atoms. Here, the Gasteiger method was used to calculate the partial charges,[36] and the *PSA* surrounding those polar atoms with absolute partial charges larger than 0.1|e| was treated as the high-charged polar surface area (*HCPSA*). In Gasteiger calculations, only the connectivities of the atoms are considered so only the topology of a molecule is important.

The number of hydrogen-bond donors and acceptors were obtained using the Patty rules,[37] which were interpreted by OELIB.[38] We defined a parameter file to store features of atoms that can form hydrogen bonds. These atoms were divided into three categories: hydrogen-bond donor (HBD), hydrogen-bond acceptor (HBA), and polar atom (POL) that has a lone electron pair and a polar hydrogen atom and can be treated as a hydrogen-bond donor or a hydrogen-bond acceptor at the same time.

(3) Molecular Bulkiness Descriptors. It is obvious that the rate of passive paracelluar transport depends strongly on molecular size. The simplest descriptor concerned with molecular size is molecular weight (*MW*). Certainly, *MW* usually correlates very well with other two descriptors: molecular volume and molecular surface area. Here, molecular volume and molecular solvent-accessible surface area (*SASA*) were estimated using the MSMS program.[35]

log*BB* Prediction based on Topological Structures. The calculations of log*P* and *HCPSA* need surface area calculation based on the 3D representation of a molecule, so the computations of these two descriptors may be relatively time-consuming. To improve the efficiency of log*BB* prediction, we tried to calculate these two descriptors only based on 2D topological information. To improve the calculation efficiency of log*P* and *HCPSA*, we adopted the following strategies:

(1) Reparametrization of SLOGP. In version 1.0 of SLOGP,[23] log*P* is calculated by summing the contribution of atom-weighted solvent accessible surface areas (*SASA*) and correction factors, so 3-D structure and molecular surface calculations should be necessary. In the revised version of SLOGP, log*P* of a molecule is calculated from the additions of atoms and correction factors, and it can be described by

$$\log P = \sum_i b_i n_i + \sum_j c_j B_j \tag{12}$$

where b_i and c_i are regression coefficients; n_i is the number of the *i*th atom type; and B_j is the number of the *j*th correction factor.

Because we do not connect *SASA* with log*P*, we may need to define more atom types to represent the atoms with different exposures to solvent. The final atom classification system includes 112 atom types, not 100 in SLOGP v1.0. The atom types were determined by using the SMARTS system. In SLOGP v2.0, we also considered two correction factors, including hydrophobic carbon and intramolecular hydrogen bond, implemented in SLOGP v1.0. More detailed descriptions of these two correction factors can be found in ref 23. The data set used for parametrization includes 1850 organic molecules, the same as those used in our previous work. The new atom typing rule and the corresponding hydrophobic contributions can be found in SLOGP v2.0.

(2) High-Charged Topological Polar Surface Area (*HCTPSA*). Recently a new protocol to generate *PSA* based solely on molecular topological information was proposed by Ertl et al.[39] The procedure calculates *PSA* from 2D molecular bonding information only. The result was termed topological polar surface area (*TPSA*). In Ertl's work the target for fitting is the van der Waals surface area, while in

Table 2. Compounds Comprising the Test Sets

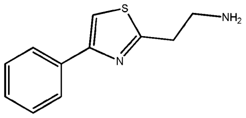
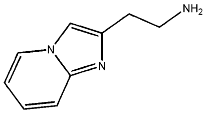
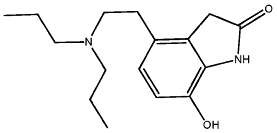
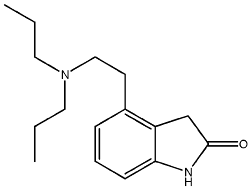
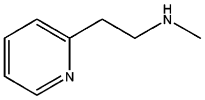
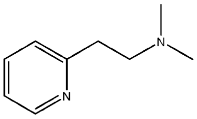
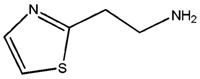
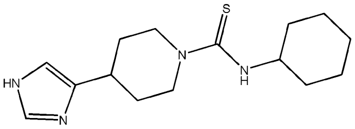
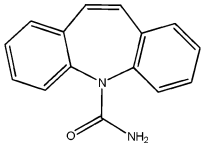
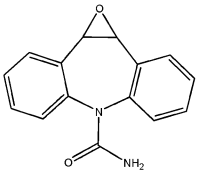
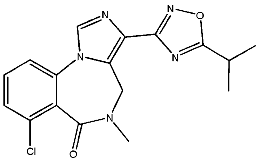
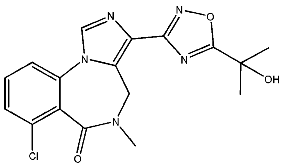
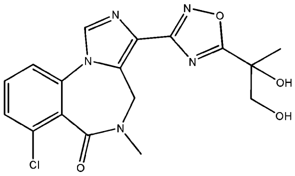
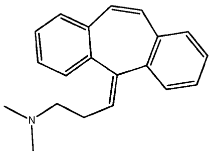
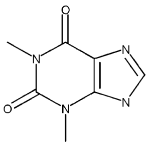
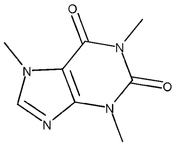
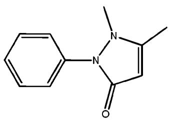
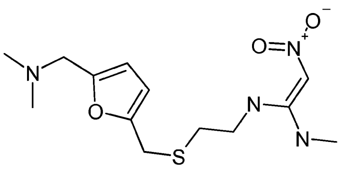
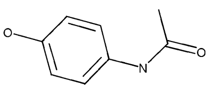
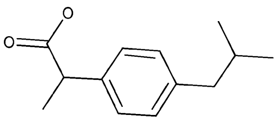
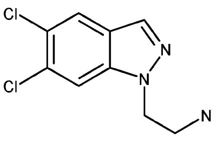
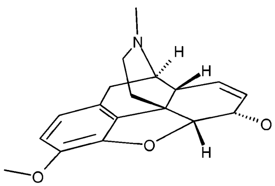
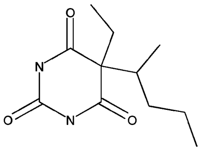
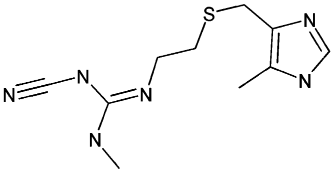
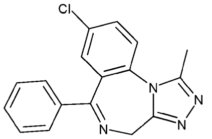
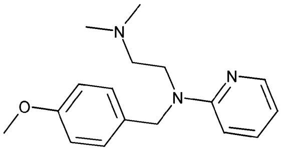
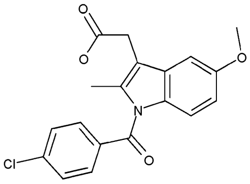
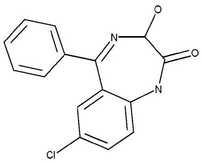
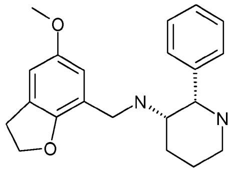
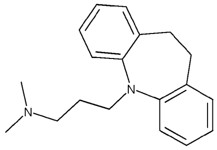
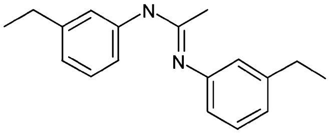
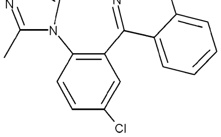
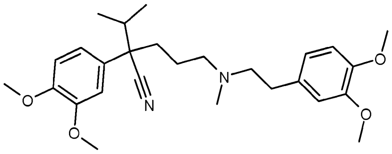
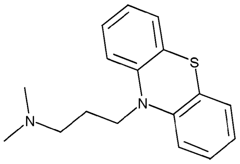
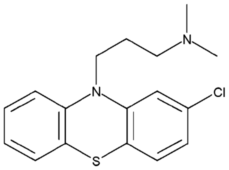
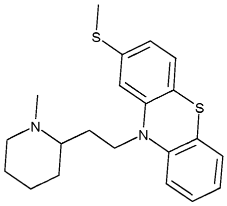
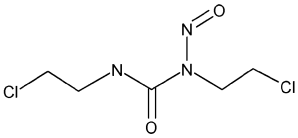
ID	ID	ID
B1 	B2 	B3 
B4 	B5 	B6 
B7 	B8 	B9 
B10 	B11 	B12 
B13 	B14 	C1 
C2 	C3 	C4 
C5 	C6 	C7 
C8 	C9 	C10 

Table II (Continued)

ID	ID	ID
C11 	C12 	C13 
C14 	C15 	C16 
C17 	C18 	C19 
C20 	C21 	C22 
C23 		

the current work the solvent-accessible surface area was used. Certainly, different procedures of surface calculations, different van der Waals atomic radii, or even different calculation parameters may generate different *PSA*. So, here, we developed a new set of atomic parameters to calculate *TPSA*. *PSA* calculated by MSMS was used as the target in fitting (see eq 13)

$$PSA = \sum_i n_i.s_i \tag{13}$$

where *PSA* is the traditionally calculated *PSA* based on 3D molecular structure using MSMS; n_i is the frequency of fragment *i* in the molecule; and s_i is the surface contribution of type *i*.

The definition for atom types using in fitting is based on SMARTS, and all types are united-atomic model (see Table 3). The training set includes 20 000 organic compounds randomly selected from the Available Chemical Database (ACD-3D).[40] Each compound should satisfy the rule of 5 proposed by Lipinski,[41] such as molecular weight smaller than 600, CLOGP smaller than 5.0, number of hydrogen-bond donors smaller than 5, and number of hydrogen-bond acceptors smaller than 10.

(3) High-Throughput log*BB* Prediction. After reparametrization of SLOGP and *TPSA*, the calculations of log*P* and *HCTPSA* are based on 2-D molecular bonding information only, so using log*P*, *HCTPSA*, and *MW*, we developed a new regression model which can predict log*BB* as a high-throughput fashion. All compounds manipulation, processing of SMARTS, input of parameters, identification of polar fragment, estimation of log*P*, *HCTPSA*, and log*BB*, were accomplished by using an in-house program named Drug-HBB written in C++.

RESULTS AND DISCUSSION

The program, Drug-BB, was developed in C++. The program reads a single molecule or multiple molecules

Table 3. Atomic Contributions (Å^2) to PSA

SMARTS	contribution	SMARTS	contribution
[#8;H1]	24.584	[n;H0](:*)(:*) −*	6.203
[#8](−*) −*	12.475	[n]:[n]	15.242
[#8;r]	18.205	[NH3]	36.928
[#8]=*	17.867	[N;H0](−*)(=*)=*	0.000
[#8]−N=O	23.875	[N;H0](−*)(−*)=*	0.000
[#8]=N−O	23.875		
[o]	24.854	[#16;H1]	46.908
		[#16;H0](−*) −*	29.273
[#7;H2]	33.645	[#16]=*	39.344
[#7;H1]=*	19.902	[#16]=O	25.149
[#7;H1](−*)−*	7.619	[#16](=*)=*	5.915
[#7;H0](−*)(−*)−*	0.000	[s]	35.737
[#7;H0](−*)=*	14.200		
[#7;H0]#*	19.924	[#15](−*)(−*)−*	16.564
[n]	17.807	[#15](−*)=*	25.376
[n;r5][c]	17.097	[#15](−*)(−*)(−*)=*	5.261
[nH]	29.817	[#15;H1](−*)(−*)=*	20.123

[a] Description: * represents any atom; n represents aromatic nitrogen; o represents aromatic oxygen; s represents aromatic sulfur; − represents a single bond; = represents a double bond; # represents a triple bond; : represents an aromatic bond. [b] Oxygen in a ring. [c] Nitrogen in five-membered ring.

(represented in single SYBYL/mol2 file, single MACCS/mol file, SYBYL/mol2 database file, or MACCS/sdf database file), performs atom typing, charge assignment, and surface estimation, and then calculates log*P*, *HCPSA*, and log*BB*. The program, Drug-HBB, was developed in C++. The difference between Drug-HBB and Drug-BB is that Drug-HBB calculates log*P*, *HCTPSA*, and log*BB* based on molecular topological information only. The program, SLOGP v2.0, was also released. Now, SLOGP can give two log*P* values for each molecule based on two different additive models.

Descriptors in QSPR Models. (1) log*P*. It was found that log*P* was an important factor although itself correlates with log*BB* poorly. Here, a direct fitting of log*P* values with log*BB* of the compounds in the training set produced an *r* of approximately 0.5:

$$\log BB = -\,0.552 + 0.236 \log P \tag{14}$$

$$(n = 78,\ r = 0.492,\ s = 0.649,\ F = 24.3)$$

In the past tremendous efforts by theoretical chemists led to several useful computational methods for estimating log*P* of organic compounds. Among all these log*P* methods, CLOGP is the oldest computational procedure, actively managed, commercially distributed, and perhaps the most widely used. Here, the SLOGP model developed in our group was used, so we expected to know if SLOGP could give effective predictions for those compounds in Tables 1 and 2. The correlation between the log*P* values predicted by CLOGP and those by SLOGP are shown in Figure 1a. From Figure 1a, it can be found that the log*P* values predicted by these two approaches show high linear correlation ($r = 0.93$).

Furthermore, we predicted the log*P* values using four other approaches including ALOGP, ALOGP98, HINT, and the Wildman model. The linear correlations of log*P* by these four methods and those by CLOGP are 0.84, 0.92, 0.78, and 0.90, respectively. As shown in Figure 1, the log*P* values predicted by these six approaches indeed exist with obvious differences. For example, the log*P* values of compound 6 predicted by CLOGP, SLOGP, the Wildman model, HINT, ALOGP, and ALOGP98 are −1.92, −1.05, −0.54, 1.57, −0.10, and −1.52, respectively. From the mean square deviation and the linear correlation coefficient, SLOGP gave the most similar results with CLOGP. Because the experimental log*P* values are unavailable, we cannot give a conclusion that for each compound SLOGP gives the best prediction, but the comparison at least demonstrates that SLOGP yields acceptable estimations for the studied compounds.

(2) *HCPSA*. We first carried out a simple linear regression of the 78 compounds in the training set using *TPSA* as the only descriptor. The resulting equation and statistics are

$$\log BB = 0.571 - 0.0156PSA \tag{15}$$

$$(n = 78,\ r = 0.753,\ s = 0.490,\ F = 100.4)$$

Compared with the Clark's results shown in eq 3, our fitting is worse. The main reason for the difference of fittings given by us and Clark is the usage of different training sets. In Clark's work, they used a training set of 55 compounds, but in our work, we used a training set of 78 compounds.

When a drug molecule passes through brain parechymal cells, different polar atoms should give a different unfavorable contribution to log*BB*, even if these two atoms bear similar *PSA* because the partial charges on these two atoms may exist quite difference. So considering the effect of partial charges, we divided the *PSA* into two categories according to the values of partial charges: high-charged polar surface area (*HCPSA*) and low-charged polar surface area (*LCPSA*). *HCPSA* is the *PSA* given by polar atoms with absolute partial charges larger than 0.1|e|, and *LCPSA* is that given by polar atoms with absolute partial chargers smaller than 0.1|e|. W constructed a correlation between log*BB* with *HCPSA*, and the resulting equation is

$$\log BB = 0.589 - 0.0177HCPSA \tag{16}$$

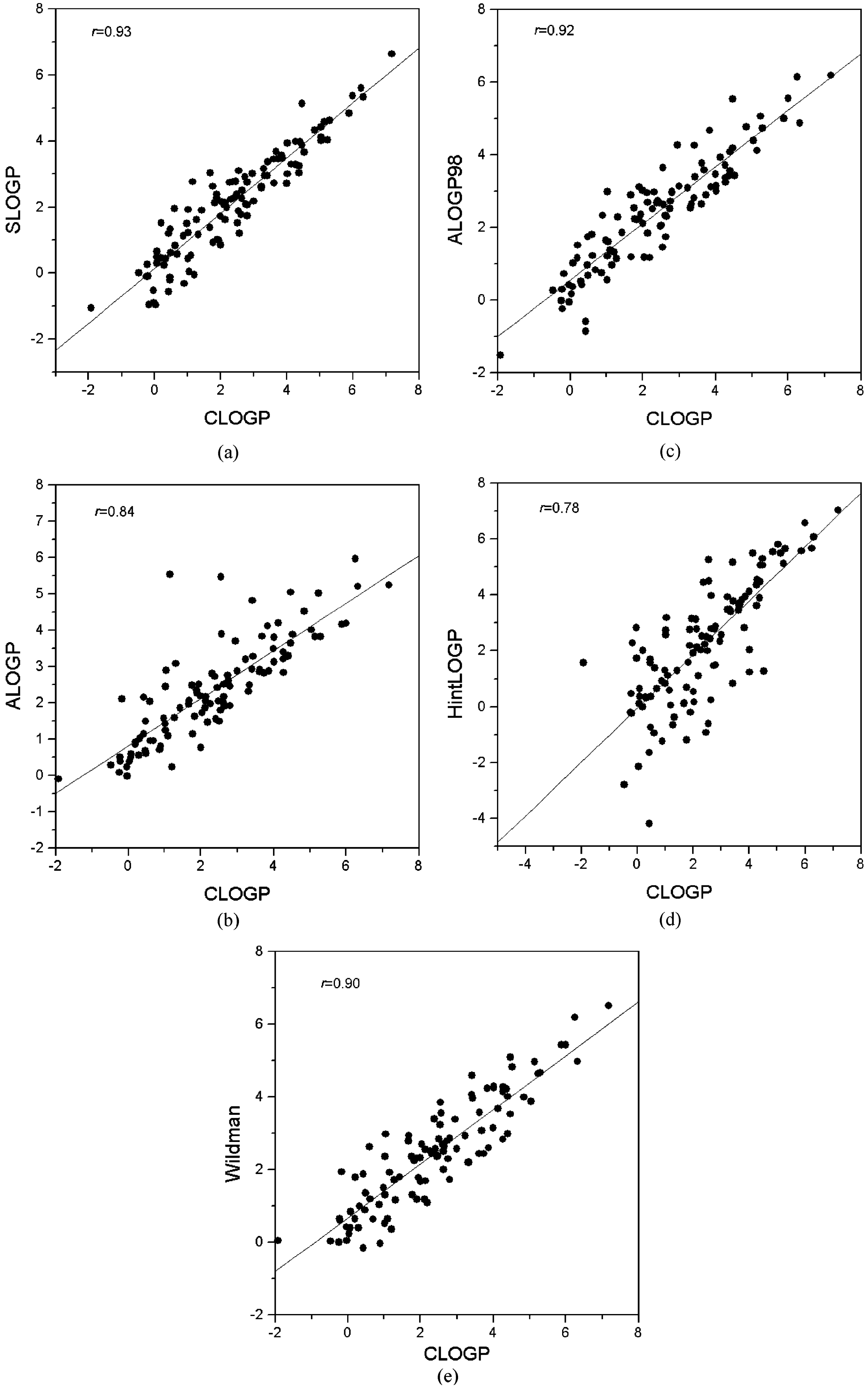
$$(n = 78,\ r = 0.779,\ s = 0.468,\ F = 117.4)$$

As shown in eqs 15 and 16, using the new parameter, the correlation coefficient and the Fisher value were improved obviously. This parameter provides a solid physical picture of the molecular mechanisms, indicating that only high-charged polar surface areas impact the *BBB* permeation. Here, the cutoff value for the calculation of *HCPSA* was defined as 0.1. In fittings, a systematical search was used to change this value from 0 to 0.2 using a step of 0.01, and finally we found that the value of 0.1 could generate the best linear model. The values of *PSA* and *HCPSA* for the compounds in the training set are listed in Table 4. From Table 4 it can be found that 63 compounds have identical *PSA* and *HCPSA*, while the other 15 compounds have smaller *HCPSA* than *PSA*.

Then, we considered both log*P* and *HCPSA* in MLR. It is interesting to find that the partitioning of compounds between the blood and brain compartments can be effectively described by a combination of log*P* and *HCPSA*.

$$\log BB = 0.219 + 0.139\log P - 0.0158HCPSA \tag{17}$$

$$(n = 78,\ r = 0.827,\ s = 0.422,\ F = 81.1)$$



(a)

(c)

(b)

(d)

(e)

Figure 1. Correlation between the calculation logP values by CLOGP with (a) SLOGP, (b) ALOGP, (c) ALOGP98, (d) HintLOGP, and (e) the Wildman's model.

Table 4. Experimental and Computed log*BB* Values for Compounds in the Training Set

ID	logBB_{exp}	log*P*	*PSA*	*HCPSA*	MW_{360}	logBB_{cal}	residue	ID	logBB_{exp}	log*P*	*PSA*	*HCPSA*	MW_{360}	logBB_{cal}	residue
1	**−2.00**	**1.72**	**76.79**	**76.79**	**19.46**	**−0.88**	**−1.12**	40	0.03	0.00	67.06	0.00	0.00	0.08	−0.05
2	−1.88	3.01	69.36	69.36	82.36	−1.41	−0.47	41	0.03	−0.57	80.98	0.00	0.00	−0.03	0.06
3	−1.57	0.85	99.74	89.70	0.00	−0.96	−0.61	42	0.04	0.53	0.00	0.00	0.00	0.19	−0.15
4	−1.54	1.02	125.07	109.41	0.00	−1.19	−0.35	43	0.08	1.90	0.00	0.00	0.00	0.46	−0.38
5	−1.30	3.55	58.74	58.74	88.59	−1.25	−0.05	44	0.08	0.99	41.68	41.68	0.00	−0.28	0.36
6	−1.30	−1.05	92.57	92.57	0.00	−1.37	0.07	45	0.11	1.92	35.44	35.44	0.00	−0.01	0.12
7	−1.17	0.44	85.06	85.06	0.00	−0.98	−0.19	46	0.13	1.19	7.00	7.00	0.00	0.22	−0.09
8	−1.15	0.92	103.47	93.40	0.00	−0.99	−0.16	47	0.13	1.62	7.56	7.56	0.00	0.30	−0.17
9	−1.12	1.88	83.65	79.01	0.00	−0.61	−0.51	48	0.14	4.84	32.92	32.92	21.54	0.29	−0.15
10	−1.10	2.21	57.49	57.49	0.00	−0.26	−0.85	49	0.22	4.58	41.30	41.30	5.48	0.35	−0.13
11	−1.06	1.38	69.25	69.25	53.55	−1.33	0.27	50	0.24	2.78	8.15	8.15	0.00	0.52	−0.28
12	−0.82	−0.32	136.16	118.01	0.00	−1.57	0.75	51	0.25	−0.21	43.93	43.93	0.00	−0.55	0.80
13	−0.73	3.66	80.50	77.09	54.53	−1.00	0.27	52	0.27	2.13	0.00	0.00	0.00	0.50	−0.23
14	−0.72	−0.97	133.67	99.88	0.00	−1.45	0.73	53	0.30	5.61	34.22	34.22	20.49	0.44	−0.14
15	−0.67	3.10	69.05	69.05	0.00	−0.24	−0.43	54	0.34	2.26	0.00	0.00	0.00	0.53	−0.19
16	−0.66	2.16	83.75	79.52	0.00	−0.56	−0.10	55	0.35	2.75	0.00	0.00	0.00	0.63	−0.28
17	−0.50	1.22	64.61	64.61	0.00	−0.55	0.05	56	0.36	3.03	0.00	0.00	0.00	0.68	−0.32
18	−0.46	1.52	39.37	39.37	0.00	−0.15	−0.31	57	0.37	2.06	0.00	0.00	0.00	0.49	−0.12
19	−0.43	2.27	41.22	41.22	0.00	−0.02	−0.41	58	0.37	2.50	0.00	0.00	0.00	0.58	−0.21
20	−0.42	0.49	25.21	25.21	0.00	−0.16	−0.26	59	0.39	2.72	33.23	33.23	14.91	−0.04	0.43
21	−0.30	2.06	37.60	37.60	0.00	−0.02	−0.28	60	0.40	2.39	0.00	0.00	0.00	0.56	−0.16
22	−0.28	2.96	82.81	79.44	0.00	−0.40	0.12	61	0.42	2.62	7.64	7.64	0.00	0.50	−0.08
23	−0.27	3.00	81.86	77.82	0.00	−0.37	0.10	62	0.44	3.28	33.4	33.4	0.00	0.28	0.16
24	−0.24	3.29	36.47	36.47	0.00	0.24	−0.48	63	0.49	2.58	18.09	18.09	0.00	0.35	0.14
25	−0.22	2.09	44.28	44.28	0.00	−0.10	−0.12	64	0.69	3.24	33.22	33.22	0.00	0.28	0.41
26	−0.18	1.73	72.57	62.45	0.00	−0.42	0.24	65	0.76	2.96	0.00	0.00	0.00	0.67	0.09
27	−0.17	0.57	23.98	23.98	0.00	−0.13	−0.04	66	0.80	3.47	0.00	0.00	0.00	0.77	0.03
28	−0.16	−0.10	24.93	24.93	0.00	−0.27	0.11	67	0.81	3.97	0.00	0.00	0.00	0.87	−0.06
29	−0.16	0.24	24.94	24.94	0.00	−0.20	0.04	68	0.83	4.43	4.54	4.54	0.00	0.90	−0.07
30	−0.15	0.30	23.81	23.81	0.00	−0.18	0.03	69	0.90	3.98	0.00	0.00	0.00	0.87	0.03
31	−0.15	−0.09	20.56	20.56	0.00	−0.21	0.06	70	0.93	3.16	0.00	0.00	0.00	0.71	0.22
32	−0.14	2.77	43.34	43.34	0.00	0.05	−0.19	71	0.97	3.46	0.00	0.00	0.00	0.77	0.20
33	−0.12	3.93	81.53	79.23	8.46	−0.33	0.21	72	1.00	3.30	36.56	36.56	0.00	0.24	0.76
34	−0.08	0.44	18.73	18.73	0.00	−0.08	0.00	73	1.00	4.03	19.64	19.64	0.32	0.61	0.39
35	−0.06	1.62	60.90	60.90	0.00	−0.42	0.36	74	1.01	3.47	0.00	0.00	0.00	0.77	0.24
36	−0.04	−0.06	74.17	64.04	0.00	−0.79	0.75	75	1.04	3.45	0.00	0.00	0.00	0.77	0.27
37	−0.02	1.78	34.58	34.58	0.00	−0.03	0.01	76	1.07	4.01	4.32	4.32	0.00	0.82	0.25
38	0.00	1.12	7.18	7.18	0.00	0.21	−0.21	77	1.20	3.88	14.58	14.58	0.00	0.65	0.55
39	0.00	2.50	40.83	40.83	0.00	0.03	−0.03	78	1.44	6.63	4.46	4.46	46.52	0.68	0.76

The obvious characteristics of a molecule with large *PSA* or *HCPSA* is that this molecule should have a strong tendency to form hydrogen bonds, because the atoms in hydrogen bonds should have highly electronegative atoms (oxygen, nitrogen, etc.). Fesher et al. even used the descriptor of the number of hydrogen-bond acceptors and found that it could improve the correlation of the equation (see eq 7). Here, besides log*P* and *HCPSA*, we used two descriptors including n_{HBA} and n_{HBD} in MLR, and the resulting equations are as follows:

$$\log BB = 0.221 + 0.139 \log P - 0.0156 HCPSA - 0.00463 n_{HBA} \quad (18)$$

$$(n = 78, r = 0.827, s = 0.422, F = 53.3)$$

$$\log BB = 0.212 + 0.139 \log P - 0.013 HCPSA - 0.150 n_{HBA} \quad (19)$$

$$(n = 78, r = 0.830, s = 0.422, F = 54.6)$$

From eqs 18 and 19, it can be found that considering these two descriptors, the statistical significances of the correlations did not have effective improvement. In fact, the descriptor *HCPSA* shows significant correlation with n_{HBA} or n_{HBD}. The correlation between *HCPSA* and n_{HBA} is 0.81, and that between *HCPSA* and n_{HBD} is 0.88, indicating that the descriptor related to the hydrogen bonds may be replaced by the descriptor *HCPSA*. As to the equation developed by Fesher et al., the improvement of fitting by n_{acc} may be caused by random correlation.

(3) Molecular Weight. Besides hydrophilicity and hydrophobicity, the bulkiness property of a molecule should be considered. Molecular bulkiness properties, such as molecular weight, molecular volume, or molecular surface, have been introduced by some research. First, we added the descriptor, molecular weight (*MW*), in correlation, and the obtained equation is

$$\log BB = 0.225 + 0.174 \log P - 0.0138 HCPSA - 0.000709 MW \quad (20)$$

$$(n = 78, r = 0.829, s = 0.423, F = 54.2)$$

Compared with eq 17, the statistical significance of eq 20 nearly does not have any improvement. It seems that introduction of a bulkiness descriptor cannot improve the correlation effectively. We do not think that molecular bulkiness does not affect the *BBB* permeation, but its effect on log*BB* should be quite different from those of other molecular features such as log*P* or *PSA*. This difference can be easily interpreted. The cavity or channel among the tight junction membranes should be limited, so only these molecules with suitable size are able to cross the membranes. When the size of a molecule is less than that of the cavity or channel, the influence of molecular bulkiness should not be obvious. When the size of a molecule is larger than a threshold, the influence of molecular bulkiness begins to go into effect. If we directly introduce bulkiness descriptors into correlation, they are considered to be additive, which is questionable. To discover the effective range of *MW*, we applied a spline model for *MW*. The spline model was denoted with angled brackets. For example, ⟨*MW-a*⟩ was equal to zero if the value of *MW-a* was negative; otherwise, it was equal to *MW-a*. The regression with splines allows the incorporation of features that do not have a linear effect over their entire range. To determine the best value of *a*, a systematical search was used to change this value from 100 to 400 using a step of 10. The best equation is presented below:

$$\log BB = 0.00845 + 0.197 \log P - 0.0135 HCPSA - 0.0140 < MW - 360 > \quad (21)$$

$$(n = 78, r = 0.876, s = 0.364, F = 81.5)$$

Compared with eq 20, the statistical significance of eq 21 is improved significantly. The threshold value of 360 in eq 21 demonstrates that larger molecular weight produces lower permeation rates, but the effect takes effect only when the molecular weight is larger than 360.

Since *MW* correlates with molecular sizes, other similar descriptors such as molecular volume or molecular surface should possess similar features with *MW*. Similarly, we constructed two linear equations including *V* and *SASA*, respectively:

$$\log BB = - 0.00740 + 0.207 \log P - 0.0135 HCPSA - 0.0166 < V - 290 > \quad (22)$$

$$(n = 78, r = 0.872, s = 0.370, F = 78.4)$$

$$\log BB = 0.291 + 0.138 \log P - 0.0098 HCPSA - 0.00969 < SASA - 450 > \quad (23)$$

$$(n = 78, r = 0.867, s = 0.377, F = 74.4)$$

Although *MW*, *V*, and *SASA* are all descriptor related to molecular size, judging from the statistical parameters of the linear equations, the best linear model is eq 21.

Model Validation. Equation 21 only includes three descriptors; moreover, from the calculation of the correlation matrix of the parameters, we found that all descriptors in eq 21 were independent. Although introduction of other descriptors may improve the correlation, we think that when the training set is limited the addition of more descriptors may introduce more possibility of random correlation. The leave-one-out (LOO) method was used to calculate the statistical quantity *q* of eq 21. The calculated *q* (0.858) shows that eq 21 is reliable. A plot of the calculated log*BB* versus observed log*BB* values for the training set is shown in Figure 2. The observed log*BB* values, calculated, and residuals are listed in Table 4.

The actual prediction power of eq 21 was validated by two external test sets. The first validation set includes the BB ratio of eight H1-receptor (**B1**−**B8**) histamine antagonist/agonist and six miscellaneous CNS agents (**B9**−**B14**). The observed and predicted log*BB* are shown in Table 5. As may be seen from Table 5, the predictions to compounds **B1**−**B14** in the test set are very good, whereas the log*BB* value for compound **B1** is strongly underestimated by this model. The log*BB* value for compound **B2** is also underestimated, but the residual is below 1.0. It should be noted that this set of compounds has also been used by the PLS model reported by Luco et al.,[18] the three-descriptor linear model reported by Feher et al.,[12] and the four-descriptor linear model reported by Hou et al.[20] Here, the estimation error exceeds ±1.00 and was considered as prediction failure. In Luco's work, compounds **B1** and **B2** were not predicted correctly. In Feher's work, two compounds including **B1** and **B11** were highly overestimated. In Hou's work, compounds **B1** and **B2** were also determined as outliers. Here, using our model, only compound **B1** has the estimation error larger than 1.0. If we treated **B1** as an outlier, the mean unsigned error is 0.16, which is lower than that reported by Hou et al. (0.41), Luco et al. (0.25), or Feher et al. (0.40). That is to say, the prediction potential of eq 21 is much better than several other models.

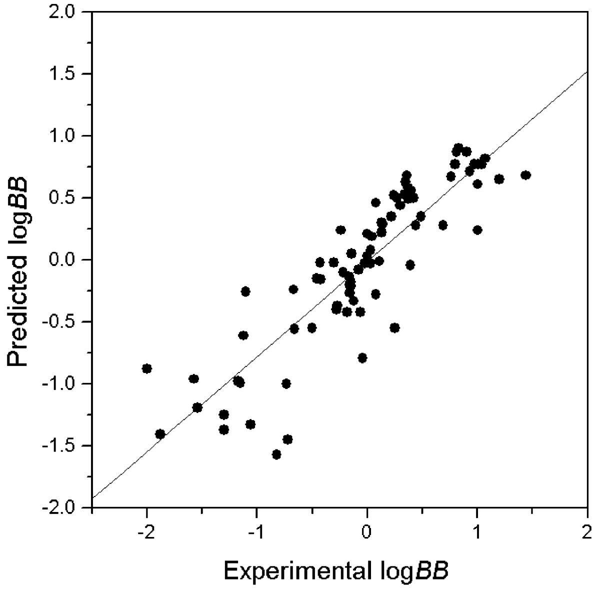


Figure 2. Comparison of experimental log*BB* with calculated log*BB* for the compounds in the training set using eq 21.

The second validation test included 23 structurally diverse compounds. The observed and predicted log*BB* data are shown in Table 5. Inspection of these results shows that that the linear model performs reasonably well, and only one compound **C13** was strongly underestimated and may be considered as an outlier. Most compounds in test set 2 are also included in the test set 2 used by Feher et al.[12] In Feher's work, compounds **C1**, **C2**, **C8**, and **C14** were not estimated correctly. But in the current work, for all these four compounds, eq 21 gave good predictions. In our previous work, two compounds including **C14** and **C19** were strongly underestimated. Considering all compounds in test set 2, the mean unsigned error using eq 21 is 0.43, which is a little better than that (0.48) in our previous work. The plot of calculated log*BB* versus observed log*BB* values for the tested compounds is shown in Figure 3. The observed log*BB* values, calculated, and residuals are listed in Table 5. The good predictions for the tested compounds confirm the significance of the three selected descriptors and the model based on them.

High-Throughput log*BB* Prediction. To improve the efficiency of log*BB* prediction, we performed reparametri-

Table 5. Experimental and Predicted log*BB* Values for Compounds Comprising Test Sets Using Equation 21

ID	logBB_{exp}	*HCPSA*	log*P*	MW_{360}	logBB_{cah}	residue
B1[c]	**−1.30**	**40.85**	**2.37**	**0.00**	**−0.02**	**−1.28**
B2	−1.40	44.10	−0.12	0.00	−0.56	−0.84
B3	−0.43	59.67	2.13	0.00	−0.33	−0.10
B4	0.25	36.28	2.77	0.00	0.12	0.13
B5	−0.30	24.56	0.65	0.00	−0.13	−0.17
B6	−0.06	14.05	0.82	0.00	0.05	−0.11
B7	−0.42	43.62	0.26	0.00	−0.48	0.06
B8	−0.16	35.12	2.18	0.00	0.02	−0.18
B9	0.00	47.70	2.91	0.00	−0.01	0.01
B10	−0.34	62.95	1.95	0.00	−0.41	0.07
B11	−0.30	50.02	1.21	0.00	−0.38	0.08
B12	−1.34	71.21	0.03	13.80	−1.09	−0.25
B13	−1.82	92.94	−0.96	29.80	−1.81	−0.01
B14	0.89	3.76	4.32	0.00	0.88	0.01
r						0.94
MUE[a]						0.16
SSE[b]						0.88
rmse						0.26
C1	−0.29	62.68	−0.89	0.00	−0.97	0.68
C2	−0.06	45.53	−0.53	0.00	−0.66	0.60
C3	−0.10	20.56	1.51	0.00	0.09	−0.19
C4	−1.23	70.08	1.33	0.00	−0.63	−0.60
C5	−0.31	54.76	0.61	0.00	−0.56	0.25
C6	−0.18	43.69	3.68	0.00	0.20	−0.38
C7	0.11	42.68	2.27	0.00	−0.07	0.18
C8	0.55	43.79	1.50	0.00	−0.23	0.78
C9	0.12	79.95	2.05	0.00	−0.63	0.75
C10	−1.42	78.16	0.46	0.00	−0.92	−0.50
C11	0.04	40.54	3.1	0.00	0.13	−0.09
C12	0.5	19.35	2.62	0.00	0.33	0.17
C13[c]	**−1.26**	**67.29**	**3.37**	**0.00**	**−0.19**	**−1.07**
C14	0.61	64.09	2.23	0.00	−0.37	0.98
C15	0.39	37.72	2.73	0.00	0.09	0.30
C16	1.30	3.94	4.11	0.00	0.84	0.46
C17	1.20	9.72	5.32	0.00	1.00	0.20
C18	0.36	21.71	3.36	0.00	0.44	−0.08
C19	−0.7	51.08	5.13	94.61	−0.89	0.19
C20	1.23	4.01	3.99	0.00	0.81	0.42
C21	1.06	3.98	4.62	0.00	0.94	0.12
C22	0.24	3.71	5.37	10.59	0.95	−0.71
C23	−0.52	49.26	1.16	0.00	−0.38	−0.14
r						0.78
MUE[a]						0.39
SSE[b]						4.97
rmse						0.48

[a] *MUE* represents mean unsigned error. [b] *SSE* represents sum of square error. [c] **B1** and **C13** are not included in the calculations of *r*, *MUE*, *SSE*, and rmse.

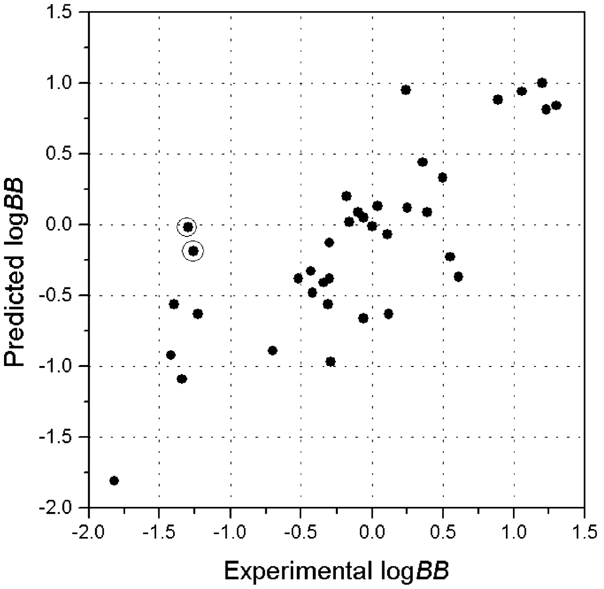


Figure 3. Comparison of experimental log*BB* with calculated log*BB* for the compounds in the test set using eq 21. (Two compounds with predicted errors larger than 1.0 are marked with circles.)

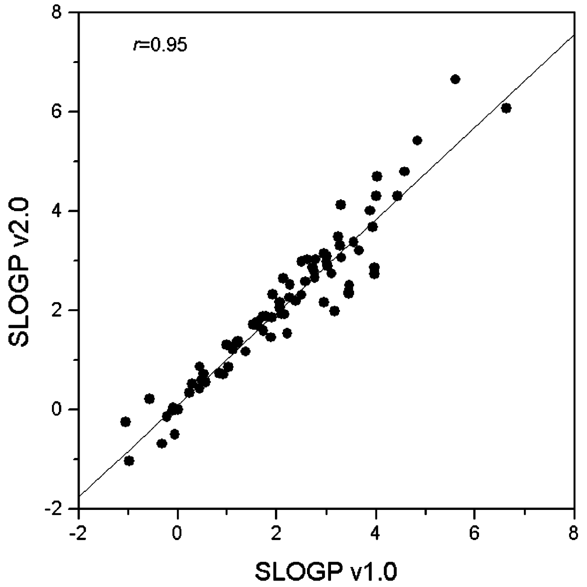


Figure 4. Correlation between the calculation logP values by addition of atom-weighted surface area in SLOGP v1.0 and simple atomic addition in SLOGP v2.0.

zation of SLOGP and proposed a set of parameters of topological polar surface. Using the new strategies, the calculations of log*P* and *HCPSA* are only based on the 2D topological information of a molecule.

(1) SLOGP v2.0 Model. The final model for log *P* calculations was obtained by correlating the total *SASA* of 112 atom types and the frequencies of two correlation factors with the experimental log*P* values. Using the new atom typing rule, we obtained a good prediction model ($n = 1850$, $r = 0.988$, $SD = 0.374$, $F = 662.574$). This model is only a little worse than that model based on addition of the atom-weighted surface area proposed by us ($n = 1850$, $r = 0.988$, $SD = 0.368$, $F = 702.218$). The new atoms typing rule and the corresponding parameters have been introduced into SLOGP v2.0.

To compare the parameters and the parameters in SLOGP v1.0, we predicted log*P* values of the compounds in the training set. Figure 4 shows the correlation between the calculated values by simple atomic addition in SLOGP v2.0 and those by addition of atom-weighted surface area in SLOGP v1.0. The log*P* values calculated by those two methods show significant correlation, which was indicated by the high correlation ($r = 0.95$) shown in Figure 4.

(2) *TPSA*. The statistical analysis provides very good correlation between 3D *PSA* and *TPSA* with the following statistical parameters: $r^2 = 0.990$, $SD = 6.43$. The atomic contributions obtained from eq 13 are listed in Table 3. The contributions for different atom types are different from those provided by Ertl et al. It is not strange because in this paper the solvent accessible molecular surface areas were used in fitting, while in Ertl's work, the van der Waals surface area were used. Using the parameters of *TPSA*, we calculated the high-charged topological polar surface area (*HCTPSA*). The calculated results are highly correlated with *CHPSA* ($r = 0.99$), which indicates that *CHTPSA* provides the same quality as the computationally much more demanding 3D *HCPSA*.

It seems that the calculated results of *HCPSA* and *HCTPSA* are quite similar. Actually, for small organic molecules, these two kinds of models are not very different, because nearly all atoms in small compounds are exposed to solvent. Here, it should be noted that although for small organic molecules the performances of the two surface models are not very different, we think *HCPSA* based on 3D molecular structure should be a more universal model especially for large molecules such as peptide. Sometimes, if some atoms in a molecule are surrounded by other atoms and located in the interior of a molecule, these atoms contribute little or even nothing to surface area, and so the calculated results of *HCTPSA* may deviate the correct value significantly.

(3) High-Throughput log*BB* Prediction. After we reparametrized SLOGP and proposed the *HCTPSA* parameters, three parameters in eq 21 were independent with the 3D molecular structure. Using the new parameters, we performed a new linear correlation and got the following equation:

$$\log BB = 0.1256 + 0.160 \log P - 0.0133 HCPSA - 0.0148 < MW - 360 > \quad (24)$$

$$(n = 78, r = 0.862, s = 0.375, F = 66.9)$$

Compared with eq 21, the correlation of eq 24 is a little worse, but eq 24 can be used as very high throughput, because the only processing step required is the identification of atom typing rules and the corresponding atomic log*P* and *HCTPSA* parameters.

To check the actual prediction potential of eq 24, the log*BB* values for two test sets were predicted using eq 24. The calculated log*P*, *HCTPSA*, and predicted log*BB* data are shown in Table 6. To these two test sets, the predictions using eq 24 are quite similar to those using eq 21. Using eq 24, two compounds **B1** and **C13** were also highly overestimated. In fact, the correlation between the predicted results using eq 21 and those using eq 24 are very high ($r = 0.97$), implying that those two models can give consistent results. Not considering any outlier, eq 24 gives an absolute mean error of 0.40. The prediction is a little worse than that given by eq 21 (absolute mean error is 0.36), but the calculation efficiency of Drug-HBB is much higher than that of Drug-BB. Based on 3D molecular structures, the Drug-BB program is able to process about 1000 molecules/min on a standard 1.2G-MHz PC, while based on topological information only, the Drug-HBB program is able to process about 5000 molecules/min on a standard 1.2G-MHz PC.

Table 6. Experimental and Predicted log*BB* Values for Compounds Comprising Test Sets Using Equation 24

ID	logBB_{exp}	*HCTPSA*	log*P*	MW_{360}	logBB_{cah}	residue
B1[c]	**−1.30**	**40.48**	**2.52**	**0.00**	**−0.01**	**−1.29**
B2	−1.40	50.87	−0.14	0.00	−0.57	−0.83
B3	−0.43	50.24	1.69	0.00	−0.27	−0.16
B4	0.25	27.35	2.17	0.00	0.11	0.14
B5	−0.30	18.59	0.60	0.00	−0.03	−0.27
B6	−0.06	10.00	0.89	0.00	0.13	−0.19
B7	−0.42	40.48	0.15	0.00	−0.39	−0.03
B8	−0.16	37.65	1.55	0.00	−0.13	−0.03
B9	0.00	45.24	2.49	0.00	−0.08	0.08
B10	−0.34	53.98	1.59	0.00	−0.34	−0.00
B11	−0.30	49.46	3.18	0.00	−0.02	−0.28
B12	−1.34	72.34	2.42	13.80	−0.64	−0.70
B13	−1.82	95.23	1.31	29.80	−1.33	−0.49
B14	0.89	1.80	4.05	0.00	0.75	0.14
r						0.96
MUE[a]						0.21
SSE[b]						0.99
rmse						0.28
C1	−0.29	57.54	−0.73	0.00	−0.76	0.47
C2	−0.06	43.39	−0.57	0.00	−0.54	0.48
C3	−0.10	37.25	2.04	0.00	−0.04	−0.06
C4	−1.23	66.56	1.56	0.00	−0.51	−0.72
C5	−0.31	48.43	0.71	0.00	−0.41	0.10
C6	−0.18	38.04	2.85	0.00	0.08	−0.26
C7	0.11	52.19	2.42	0.00	−0.18	0.29
C8	0.55	40.34	1.15	0.00	−0.23	0.78
C9	0.12	66.24	1.27	0.00	−0.55	0.67
C10	−1.42	89.67	0.09	0.00	−1.05	−0.37
C11	0.04	34.48	3.97	0.00	0.30	−0.26
C12	0.5	18.71	2.58	0.00	0.29	0.21
C13[c]	**−1.26**	**61.89**	**3.91**	**0.00**	**−0.07**	**−1.19**
C14	0.61	59.91	2.44	0.00	−0.28	0.89
C15	0.39	36.44	2.76	0.00	0.08	0.31
C16	1.30	3.61	4.30	0.00	0.77	0.53
C17	1.20	10.40	3.70	0.00	0.58	0.62
C18	0.36	22.77	3.76	0.00	0.42	−0.06
C19	−0.7	53.76	4.37	94.61	−1.17	0.47
C20	1.23	3.61	4.32	0.00	0.77	0.46
C21	1.06	3.61	4.94	0.00	0.87	0.19
C22	0.24	3.61	−0.73	10.59	0.83	−0.59
C23	−0.52	42.50	−0.57	0.00	−0.26	−0.26
r						0.79
MUE[a]						0.41
SSE[b]						4.89
rmse						0.47

[a] *MUE* represents mean unsigned error. [b] *SSE* represents sum of square error. [c] **B1** and **C13** are not included in the calculations of *r*, *MUE*, *SSE*, and rmse.

CONCLUSION

In the current work, based on a large set of organic compounds linear correlation models were developed to estimate blood-brain partitioning values. The best linear model includes three descriptors: *n*-octanol/water partition coefficient calculated using the SLOGP approach, log*P*; high-charged polar surface areas based on Gasteiger partial charges, PSA_{HC}; and the excessive molecular weight larger than 360, MW_{360}. These three descriptors give a meaningful physical picture of the molecular mechanisms involved in *BBB* permeation: a hydrophobic molecule can penetrate the *BBB* barrier easier; larger polar surface areas have more negative contribution to log*BB* values, but the contributions are only limited to those atoms with high-charge densities; and a larger molecule will lead to worse *BBB* penetration ability, but this bulk effect may take effect when the molecular weight is larger than 360. The predictions to the external test sets demonstrate that this model bears good performance and can be used for estimation of log*BB* values for drug and drug-like molecules.

To improve the efficiency of prediction, we made an extensive reparametrization of SLOGP and developed a new set of parameters to calculate topological polar surface area. Based on the new procedures, the calculations of log*P*, *HCTPSA*, and log*BB* are only based on the topological structure of a molecule and can be performed as real high-throughput fashion.

ACKNOWLEDGMENT

This project is supported by the National Natural Science Foundation of China (NSFC 29992590-2 and 29873003).

Supporting Information Available: The new atomic parameters for calculations of log*P* based on simple atomic addition have been incorporated into program SLOGP v2.0. In SLOGP v2.0, a parameter file name hydro_NA.prm was used to store the SMARTS definitions of atom types and the corresponding hydrophobic parameters. The calculations of log*P* based on SLOGP v1.0, *HCPSA*, and log*BB* based on the 3D structure of a molecule have been incorporated into a program named Drug-BB. The calculations of log*P* based on SLOGP v2.0, HCTPSA, and log*BB* based on topological information of a molecule have been incorporated into a program name Drug-HBB. In Drug-HBB, a parameter file name typsa.prm is used to store the definition of SMARTS definitions of atom types and the corresponding surface parameters. The SLOGP v2.0, Drug-BB, and Drug-HBB have been tested on IRIX and Linux operation systems, and all the programs can be obtained from authors upon request or downloaded from the Web site www.cadd.chem.pku.edu.cn. The structures in the training set and the test set are saved in MACCS/sdf database format. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Chaturvedi, P. R.; Decker, C. J.; Odinecs, A. Prediction of pharmacokinetic properties using experimental approaches during early drug discovery. *Curr. Opin. Chem. Bio.* **2001**, *5*, 452−463.

(2) Caldwell, G. W. Compound optimization in early- and late-phase drug discovery: Acceptable pharmacokinetic properties utilizing combined physicochemical, in vitro and in vivo screens. *Curr. Opin. Drug Discus*s. **200**0, *3*, 30−41.

(3) Guyton, A. C. *Textbook of Medical Physiology*, 8th ed.; W. B. Saunders Co.: Philadelphia, 1991; pp 683−684.

(4) Davson, H.; Segal, M. *Physiology of the CSF and Blood Brain Barriers*; CRC Press: Boca Raton, FL, 1996; p 8.

(5) Eddy, E. P.; Maleef, B. E.; Hart, T. K. Smith, P. L. In vitro models to predict blood-brain barrier permeability. *Adv. Drug Deliv. Rev.* **1997**, *23*, 185−198.

(6) Reichel, A.; Begley, D. J. Potential of immobilized artificial membranes for predicting drug penetration across the blood-brain barrier. *Pharm. Res.* **1998**, *15*, 1270−1274.

(7) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders: D.; Smith, L. R.; Sore, N. E.; Wilks, T. J. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *J. Med. Chem.* **1988**, *31*, 656−671.

(8) Ter Laak, A. M.; Tsai, R. S.; Donne-Op den Kelder, G. M.; Carrupt, P.-A.; Testa, B.; Timmermann, H. Lipophilicity and hydrogen-bonding capacity of H1-antihistaminic agents in relation to their central sedative side-effects. *Eur. J. Pharm. Sci.* **199**4, 2, 373−384.

(9) Kaliszan, R.; Markuszewski, M. Brain/blood distribution described by a combination of partition coefficient and molecular Mass. *Int. J. Pharm.* **1996**, *145*, 9−16.

(10) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Re*s. **1999**, *1*6, 1514−1519.

(11) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sc*i. **1999**, *8*8, 815−821.

(12) Feher, M.; Sorial, E.; Schmidt, J. M. A simple model for the prediction of blood-brain partitioning, *Int. J. Pharm.* **2000**, *201*, 239−247.

(13) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen bonding. 33. factors that influence the distribution of solute between blood and brain. *J. Pharm. Sci.* **1994**, *83*, 1257−1268.

(14) Abraham, M. H.; Chada, H. S.; Mitchell, R. C. Hydrogen-bonding part 36. determination of blood brain distribution using octanol−water partition coefficients. *Drug Des. Discov.* **1995**, *13*, 123−131.

(15) Lombardo, F.; Blake, J. F.; Curatolo, W. J. Computation of brain-blood partitioning of organic solute via free energy calculations. *J. Med. Chem.* **1996**, *39*, 4750−4755.

(16) Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. Prediction of blood-brain partitioning using monte carlo simulations of molecules in water. *J. Comput.-Aid. Mol. Des.* **2001**, *15*, 697−708.

(17) Norinder, U.; Sjoberg, P.; Osterberg, T. Theoretical calculation and prediction of brain-blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.* **1998**, *87*, 952−959.

(18) Luco, J. M. Prediction of the Brain-Blood Distribution of a Large Set of Drugs from Structurally Derived Descriptors Using Partial Least-Squares (PLS) Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396−404.

(19) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach. *J. Mol. Struct. (THEOCHEM).* **2000**, *503*, 17−30.

(20) Hou, T. J.; Xu, X. J. ADME Evaluation in drug discovery. 1. Applications of genetic algorithms on the prediction of blood-brain partitioning of a large set drugs. *J. Mol. Model.* **2002**, *8*, 337−349.

(21) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure−activity relationships i. partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(22) Cerius2 4.5, Molecular Simulation Inc., San Diego, USA, 2001.

(23) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 2. Prediction of Partition Coefficient by Atom-additive Approach Based on Atom-weighted Solvent Accessible Surface Areas, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058−1067.

(24) Yazdanian, M.; Glynn, S. L. In vitro blood-brain barrier permeability of nevirapine compared to other HIV antiretroviral agents. *J. Pharm. Sci.* **1998**, *87*, 306−310.

(25) Lin, J. H.; Chen, I.; Lin, T. Blood-brain barrier permeability and in vivo activity of partial agonists of benzodiazepine receptor: a study of L-663,581 and its metabolites in rats. *J. Pharmacol. Exp. Ther.* **1994**, *271*, 1197−1202.

(26) Van Belle, K. Brain, Liver, and blood distribution kinetics of carbamazepine and its metabolic interaction with clomipramine in rats: a quantitative microdialysis study. *J. Pharmacol. Exp. Ther.* **1995**, *272*, 1217−1222.

(27) Calder, J. A. D.; Ganelline, R. Predicting the brain-penetrating capability of histaminergic compounds. *Drug Des. Discov.* **1994**, *11*, 259−268.

(28) Salminen, T.; Pulli, A.; Taskinen, J. Relationship between immobilised artificial membrane chromatographic retention and the brain penetration of structurally diverse drugs. *J. Pharm. Biomed. Anal.* **1997**, *15*, 469−477.

(29) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(30) Leo, A. Calculating loP_{oct} from structures. *Chem. Rev.* **1993**, *93*, 1281−1306.

(31) Ghose, A. K., Viswanadhan, V. N. and Wendoloski, J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem.* **1998**, *102*, 3762−3772.

(32) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aid. Mol. Des.* **1991**, *5*, 545−552.

(33) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(34) SYBYL 6.5 User Guide; Tripos Inc., St. Louis, MO, 1999.

(35) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.

(36) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(37) Bush, B. L.; Sheridan, R. P. PATTY: A Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.

(38) James, C. A.; Weininger, D.; Delany, J. Daylight theory manual Daylight 4.62, Daylight Chemical Information Systems, Inc., Los Altos, 2001.

(39) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contribution and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(40) ACD-3D database, MDL Information Systems, Inc.; San Leandro, CA.

(41) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.

CI034134I