

ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties

T. J. Hou, W. Zhang, K. Xia, X. B. Qiao, and X. J. Xu*

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

Received April 3, 2004

The correlations between Caco-2 permeability ($\log P_{\text{app}}$) and molecular properties have been investigated. A training set of 77 structurally diverse organic molecules was used to construct significant QSAR models for Caco-2 cell permeation. Cellular permeation was found to depend primarily upon experimental distribution coefficient ($\log D$) at pH = 7.4, high charged polar surface area (HCPSA), and radius of gyration (rgyr). Among these three descriptors, $\log D$ may have the largest impact on diffusion through Caco-2 cell because $\log D$ shows obvious linear correlation with $\log P_{\text{app}}$ ($r=0.703$) when $\log D$ is smaller than 2.0. High polar surface area will be unfavorable to achieve good Caco-2 permeability because higher polar surface area will introduce stronger H-bonding interactions between Caco-2 cells and drugs. The comparison among HCPSA, PSA (polar surface area), and TPSA (topological polar surface area) implies that high-charged atoms may be more important to the interactions between Caco-2 cell and drugs. Besides $\log D$ and HCPSA, rgyr is also closely connected with Caco-2 permeabilities. The molecules with larger rgyr are more difficult to cross Caco-2 monolayers than those with smaller rgyr. The descriptors included in the prediction models permit the interpretation in structural terms of the passive permeability process, evidencing the main role of lipophilicity, H-bonding, and bulk properties. Besides these three molecular descriptors, the influence of other molecular descriptors was also investigated. From the calculated results, it can be found that introducing descriptors concerned with molecular flexibility can improve the linear correlation. The resulting model with four descriptors bears good statistical significance, $n = 77$, $r = 0.82$, $q = 0.79$, $s = 0.45$, $F = 35.7$. The actual predictive abilities of the QSAR model were validated through an external validation test set of 23 diverse compounds. The predictions for the tested compounds are as the same accuracy as the compounds of the training set and significantly better than those predicted by using the model reported. The good predictive ability suggests that the proposed model may be a good tool for fast screening of $\log P_{\text{app}}$ for compound libraries or large sets of new chemical entities via combinatorial chemistry synthesis.

INTRODUCTION

For the development of bioactive molecules as therapeutic agents, oral bioavailability is often an important consideration. Therefore besides sufficient activity, an important goal for drug research is to gain sufficient understanding of the molecular properties that limit oral bioavailability to facilitate the design of new drug candidates. Now, the development of modern technologies, such as combinatorial chemistry and high throughput screening, into drug discovery has resulted in a vast increase in the number of lead compounds synthesized in pharmaceutical drug discovery setting. But in most cases such lead compounds demonstrate unfavorable biopharmaceutical properties, such as low oral bioavailability. It is believed that over 50% of the candidates failed due to ADME/Tox deficiencies during development.^{1,2} To avoid this failure at the development stage, a set of in vitro ADME screens has been implemented in many pharmaceutical companies with the aim of discarding compounds in early stage of drug discovery process. Even though the early stage in vitro ADME reduces the probability of the failure at the development stage, it is still time-consuming and resource-intensive. For this reason it is necessary to develop in silico methods for predicting drug-likeness.

Many factors will influence oral bioavailability. Among them, human intestinal absorption may be one of the most important factors. Many in vitro cell culture models have been investigated as potential tools for drug absorption.^{3,4} The most widely used in vitro model is Caco-2 cell line.³ Caco-2 cells, a well-differentiated intestinal cell line derived from human colorectal carcinoma, display many of the morphological and functional properties of the in vivo intestinal epithelial cell barrier. Extensive studies have shown that human oral drug absorption and permeability coefficient have good correlations, suggesting that the human absorption can be well predicted by this in vitro model.⁵ Caco-2 culture models have many advantages. First, it measures the transport of the drug across a cell membrane, rather than an interaction of the drug with the lipid bilayer. Second, it can measure the parallel transport routes, both passive and active. However, it possesses several limitations including long preparation time, very slow absorption times compared to the human intestine, and large interlaboratory differences in quantitative results. In light of the limitations in throughput with these systems, the development of higher throughput computational tools for the reliable prediction of Caco-2 permeability is demanding.^{6,7}

For a drug molecule across the intestinal epithelium, there are two important routes for permeation, including passive diffusion and carrier-mediated influx via active transport

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

mechanisms.⁶ It is assumed that with only a few exceptions, these orally administered drugs were transported across the intestinal epithelium predominantly by a passive transcellular process. For passive diffusion, two types exist for permeation, the paracellular and the transcellular routes. Many factors will affect the passive intestinal permeability of a molecule including lipophilicity, hydrogen bonding capacity, solute size, and et al.^{8,9} Many physicochemical descriptors have been introduced to construct the Caco-2 permeability prediction models, including polar surface area,¹⁰ hydrogen-bonding and size descriptors,¹¹ MolSurf-derived descriptors,¹² Volsurf-derived descriptors,¹³ MO-calculation,¹⁴ membrane-interaction analysis,¹⁵ and quadratic indices.¹⁶ Most of the earlier models using multiple linear regression (MLR) or partial least-squares (PLS) were based on a relatively small set of molecules and not fully validated by external prediction sets. For example, in Palm's work, the training set only includes six molecules. Some QSPR models have relative good statistical significances, but their actual predictive abilities may be questionable. The reliable models can only be produced based on enough samples.

In this study, we have investigated a number of different molecular descriptors and studied their relationships to Caco-2 permeation systematically. Our ultimate goal is to find a set of simple descriptors to estimate Caco-2 permeability as high throughput fashion. Moreover, to develop relatively universal predictive models, we used a large set of Caco-2 permeability data from a variety of sources. The whole data set includes 100 drugs or drug-like molecules. As far as we know, the data set used in this paper is much larger than those used in earlier works.

METHODS

Data Set. The apparent Caco-2 permeability coefficients for 110 structurally diverse compounds were taken from 13 references.^{17–29} The structurally heterogeneous used covers a relatively wide range of molecular size and lipophilicity. All compounds were regarded as being transported by passive diffusion. The data set was split into a training set of 77 compounds (Table 1) and a test set of 23 compounds (Table 2). The structures of the compounds were built within the SYBYL 6.9 molecular simulation package and modeled in their neutral forms.³⁰ Geometry optimization was performed by using molecular mechanism with MMFF force field,³¹ and the terminal condition was the RMS of potential energy smaller than $0.001 \text{ kcal} \cdot \text{\AA}^{-1} \cdot \text{mol}^{-1}$. For these flexible compounds, the conformational analyses were performed to determine the most stable conformers. The models were then saved in two MACCS sdf databases named training.sdf and test.sdf for further analysis. The MACCS sdf database files are available in the Supporting Information.

The Caco-2 permeability coefficients for all compounds are shown in Table 4. From Table 4, it can be found that even for the same compound, the experimental $\log P_{\text{app}}$ values from different sources exist obvious variation. This is not strange. Artursson et al. showed that there exist large variances in Caco-2 cell results due to factors such as cell passage number, culture time, type of support, and medium.³² Egan et al. conducted a small survey of published Caco-2 cell permeability data and found that average percent relative standard deviation (%RSD) was in the range of 5.6–28.3%.³³ So large cautions must be taken to ensure that a meaningful

comparison can be made because the variability in Caco-2 assay results among different laboratories is so large. To decrease the deviation produced by different assays, the Caco-2 values averaged from different experimental sources were used in the current work. Here, it should be noted that simply adopting the average value is not very suitable. For example, for 12 compounds obtained from the Hovgaard's work,²¹ the experimental $\log P_{\text{eff}}$ values for acebutolol, alprenolol, propranolol, and timolol can also be found in other references. So for compounds obtained from ref 21, if some compounds adopt the average values and the other compounds adopt the single values, the parallelism within the same experimental assay will be destroyed. Here, we treated the experimental values from ref 25 as the benchmark, because most of the $\log P_{\text{app}}$ values used in this paper are obtained from ref 25. Besides acebutolol, the difference between the $\log P_{\text{app}}$ value of alprenolol, propranolol, or timolol provided by ref 21 and that provided by ref 25 is about 0.50. So all experimental values obtained from ref 21 were adjusted by adding a constant of -0.5 . Similarly, all $\log P_{\text{app}}$ values for molecules obtained from ref 19 were adjusted by adding a constant of -0.7 . In the training set, compound PNU200603 was selected from the Nicolaas's work.²⁶ In the Nicolaas's work, six factor Xa inhibitors were investigated. Besides PNU200603, the other five inhibitors show obvious asymmetrical transport, implying active transport or metabolic process may be involved in the absorption process. So we only selected PNU200603 into the training set. Considering the difference between the average $\log P_{\text{app}}$ value of mannitol in Table 4 and $\log P_{\text{app}}$ determined by Nicolaas et al., $\log P_{\text{app}}$ of PNU200603 was adjusted by adding a constant of 0.6. We believed that the adjustment using here was reasonable because the different systematic errors from different assays were not enlarged. The $\log P_{\text{eff}}$ values for most compounds in the test set are obtained from the work of Camenisch et al.²⁹ In Camenisch's work, the Caco-2 experiments were designed based on the work of Artursson's group, since the authors want to combine their data with previously measured compounds in order to have a larger data set.¹⁸ Preliminary Caco-2 cell permeation studies were performed using atenolol, practolol, warfarin, testosterone, and mannitol. With the exception of mannitol, it was demonstrated that the Camenisch's permeability data are well correlated with and quantitatively nearly identical to data in the literature.¹⁷ From the $\log P_{\text{eff}}$ data in Table 4, we can find that the experimental values from refs 17 and 24 show high consistency, so we can say that the experimental $\log P_{\text{eff}}$ values in the training set and the test set are comparative.

Descriptors Used in Multiple Linear Regression. (1) Hydrophobicity Descriptor. The experimental $\log D$ values at pH 7.4 for all compounds were obtained from the literature.^{19–21,25–29,34,35} It should be noted that the $\log D$ values for the same compound might exist with some differences in a different reference. For example, the $\log D$ values for atenolol from refs 25 and 29 are -1.29 and -2.14 , respectively. In most cases, the average $\log D$ value was used. Sometimes, the $\log D$ value for one compound could be found in three references. In these cases, if these three values do not exist with a large difference, the average value was used. But if one value exists with large differences compared with

Table 1. 2D Structures of Compounds in the Training Set

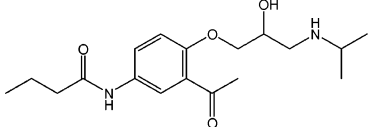
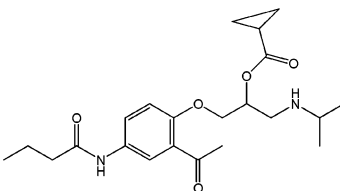
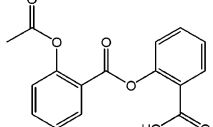
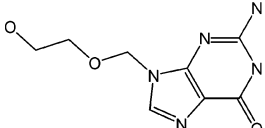
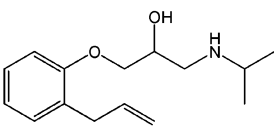
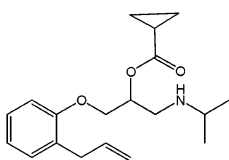
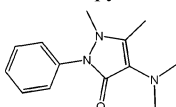
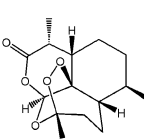
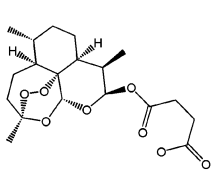
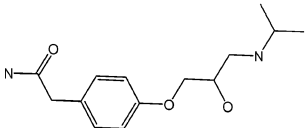
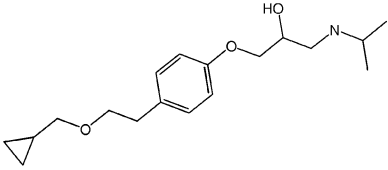
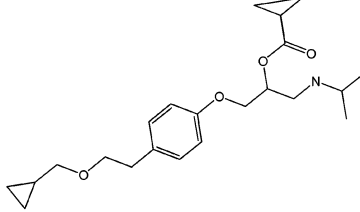
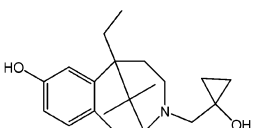
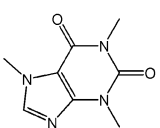
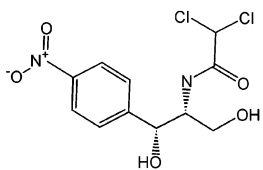
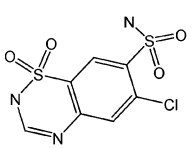
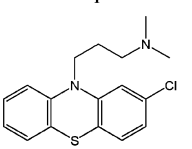
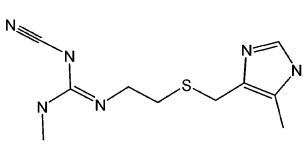
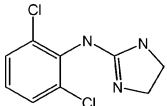
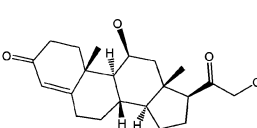
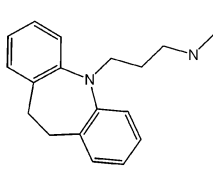
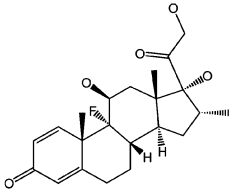
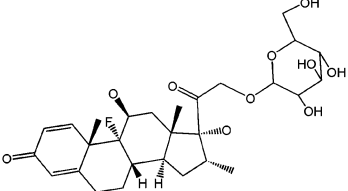
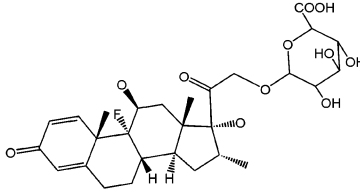
1. Acebutolol 	2. Acebutolol ester 	3. Acetylsalicylic acid 
4. Acyclovir 	5. Alprenolol 	6. Alprenolol ester 
7. Aminopyrine 	8. Artemisinin 	9. Artesunate 
10. Atenolol 	11. Betaxolol 	12. Betaxolol ester 
13. Bremazocine 	14. Caffeine 	15. Chloramphenicol 
16. Chlorothiazide 	17. Chlorpromazine 	18. Cimetidine 
19. Clonidine 	20. Corticosterone 	21. Desipramine 
22. Dexamethasone 	23. Dexamethasone-β-D-glucoside 	24. Dexamethasone-β-D-glucuronide 

Table 1 (Continued)

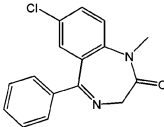
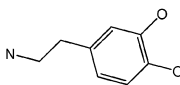
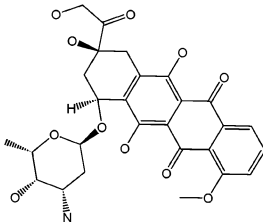
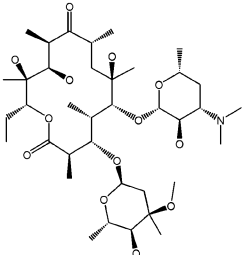
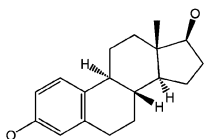
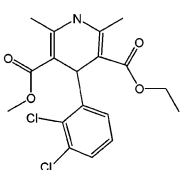
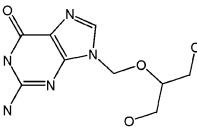
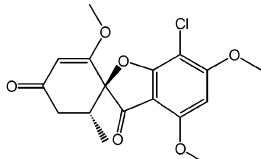
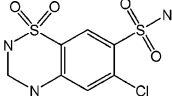
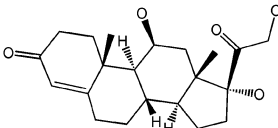
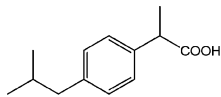
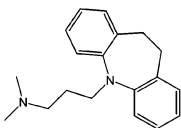
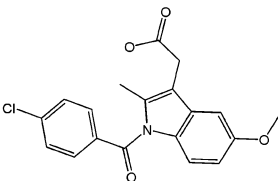
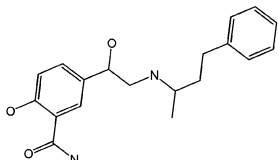
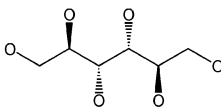
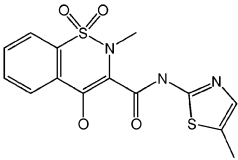

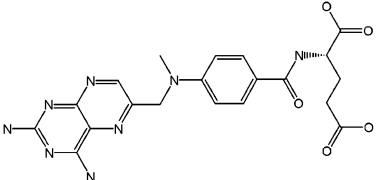
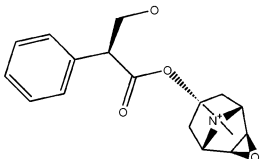
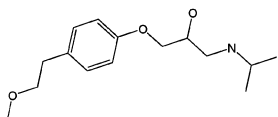
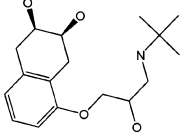
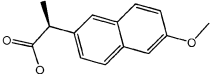
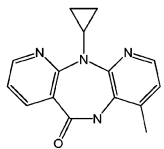
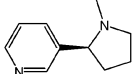
25. Diazepam 	26. Dopamine 	27. Doxorubicin 
28. Erythromycin 	29. Estradiol 	30. Felodipine 
31. Ganciclovir 	32. Griseofulvin 	33. Hydrochlorothiazide 
34. Hydrocortisone 	35. Ibuprofen 	36. Imipramine 
37. Indomethacin 	38. Labetalol 	39. Mannitol 
49. Meloxicam 	41. Methanol 	42. Methotrexate 
43. Methylscopolamine 	44. Metoprolol 	45. Nadolol 
46. Naproxen 	47. Nevirapine 	48. Nicotine 

Table 1 (Continued)

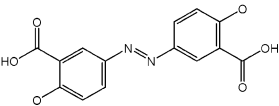
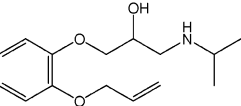
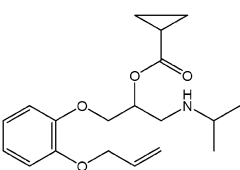
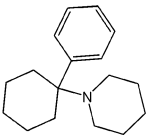
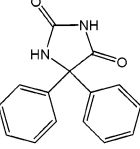
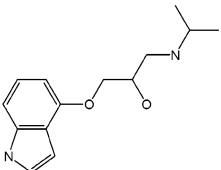
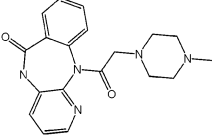
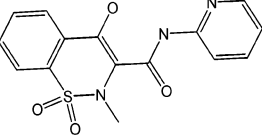
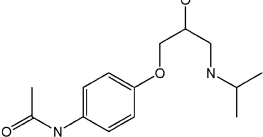
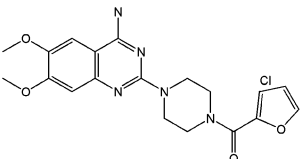
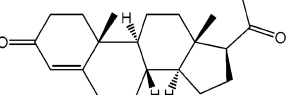
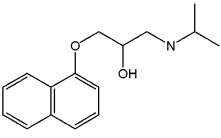
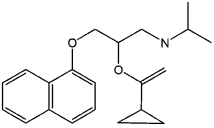
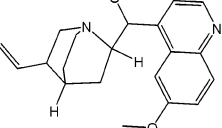
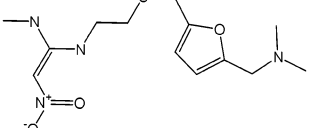
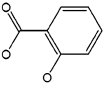
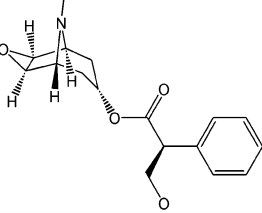
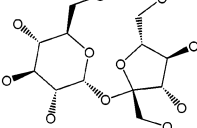
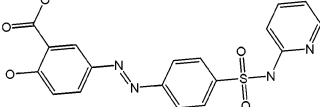
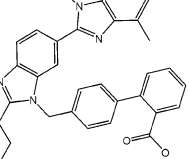
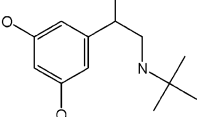
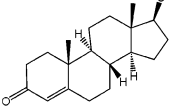
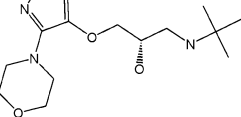
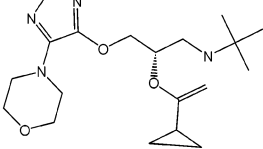
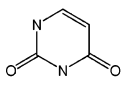

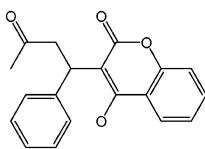
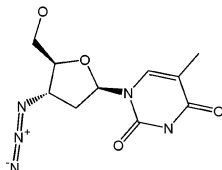
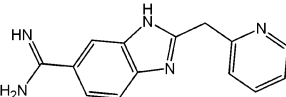
49. Olsalazine 	50. Oxprenolol 	51. Oxprenolol ester 
52. Phencyclidine 	53. Phenytoin 	54. Pindolol 
55. Pirenzepine 	56. Piroxicam 	57. Practolol 
58. Prazocin 	59. Progesterone 	60. Propranolol 
61. Propranolol ester 	62. Quinidine 	63. Ranitidine 
64. Salicylic acid 	65. Scopolamine 	66. Sucrose 
67. Sulphasalazine 	68. Telmisartan 	69. Terbutaline 
70. Testosterone 	71. Timolol 	72. Timolol ester 

Table 1 (Continued)

73. Uracil 	74. Urea 	75. Warfarin 
76. Zidovudine 	77. PNU200603 	

the other two values, this value was treated as outlier and rejected. All experimental $\log D$ values are presented in Tables 4 and 5.

As a comparison, the predicted $\log P$ values were also calculated. The $\log P$ values for all compounds were calculated using the SLOGP method developed in our group.³⁶ SLOGP estimates the $\log P$ values by summing the contribution of atom-weighted solvent accessible surface areas (SASA) and correction factors. Comparison of various procedures of $\log P$ calculation for the external test set demonstrates that our method bears very good accuracy and is comparable or even better than the fragment-based approaches.

(2) Hydrophilicity Descriptor. Due to the physical nature of the lipid bilayer, organic molecules, which can form favorable hydrogen-bonding or electrostatic interactions with lipid bilayer, may exist with great difficulties in Caco-2 penetration. To cross a membrane, a drug molecule needs to break hydrogen bonds with its aqueous environment. The more potential hydrogen bonds a molecule can make, the more energy this bond breaking costs, and so high hydrogen-bonding potential is an unfavorable property that is often related to low permeability and absorption. A variety of computational approaches have addressed the problem of estimating hydrogen-bonding capacity, ranging from simple heteroatom (O and N) counts, the consideration of molecules in terms of the number of hydrogen-bond acceptors and donors, and more sophisticated measures that take into account such parameters as free-energy factors and (dynamic) polar surface area (PSA). 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.^{6,7} Here, the polar atoms include all oxygen atoms, nitrogen atoms, sulfur atoms, and hydrogen atoms connected with them. By definition, PSA evaluation requires 3D molecular conformation and atomic surface area. Here, molecular solvent accessible surface areas were calculated using the MSMS program and the probe radius was set to 0.5 Å.³⁷

Generally, as a polar atom, it should be highly electronegative and possess high density of charge. If the charge density on an oxygen atom or a nitrogen atom is very low, it may not produce strong hydrogen bond or electrostatic interactions with other polar atoms. In our previous work, we proposed a new concept of high-charged polar surface area (HCPSA).³⁸ According to our definition, only polar atoms with high charge densities belong to high-charged polar

atoms. Here, the Gasteiger–Marsili method was used to calculate the partial charges,³⁹ 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).

As a comparison, the topological polar surface area (TPSA) was calculated. The concept of TPSA was originally proposed by Ertl et al.⁴⁰ The procedure calculates TPSA from 2-D molecular bonding information only, so it allows PSA calculations to be implemented in virtual screening approaches. Here, the SMARTS definition and the contributions for different atom types were obtained from the Ertl's work.

The number of hydrogen-bond donors (n_{HBD}) and acceptors (n_{HBA}) were obtained using the Patty rules,⁴¹ which were interpreted by OELIB.⁴² 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 hydrogen-bond acceptor.

(3) Molecular Bulkiness Descriptors. Molecular size is a very important factor influencing the diffusion in biological membrane and continuous fluid media.⁴³ For diffusion in water this dependence is relatively small, while for trans-cellular diffusion in biological membranes, a rather strong dependence on molecular size could be observed. The paracellular diffusion through the tight junctions of the biological membrane is, because of the restricted size of these aqueous pores, also dependent upon molecular size.⁴⁴ The simplest descriptor concerned with molecular size is molecular weight (MW). However, molecular weight may be so simple to describe the shape of a molecule. For example, there are two molecules with similar molecular weights; if one molecule is thin and the other one is global, the Caco-2 permeability of the thin molecule may be easier than the global one.

Compared with MW, two descriptors of molecular volume and molecular surface area should be more suitable to describe the shape of a molecule. Here, molecular volume (V) and molecular solvent-accessible surface area (SASA) were estimated using the MSMS program.³⁷ Moreover, radius of gyration (rgyr) was calculated using the following equation

$$rgyr = \sqrt{\frac{1}{M} \sum_{i=1}^n m_i r_i^2} \quad (1)$$

Table 2. 2D Structures of Compounds in the Test Set

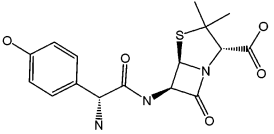
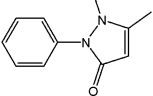
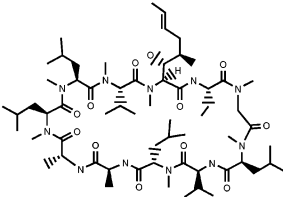
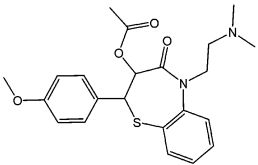
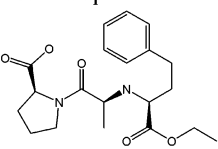
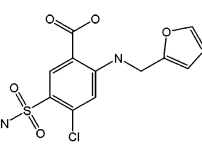
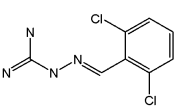
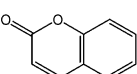
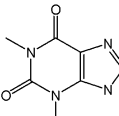
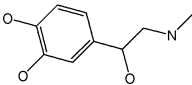
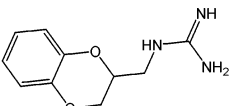
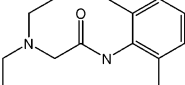
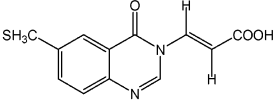
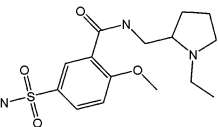
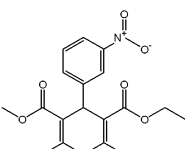
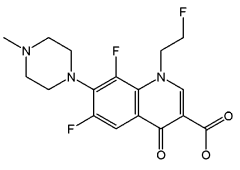
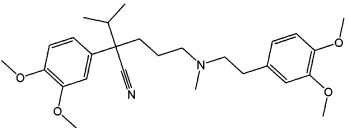
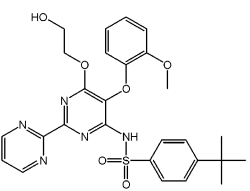
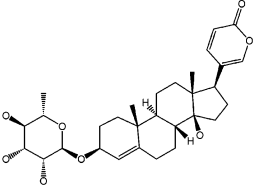
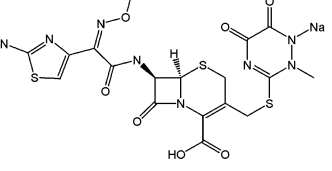
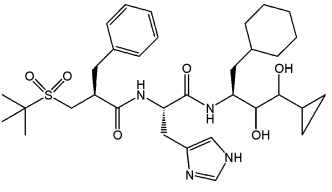
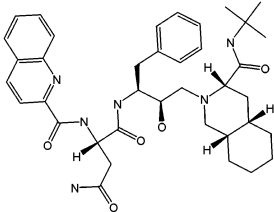
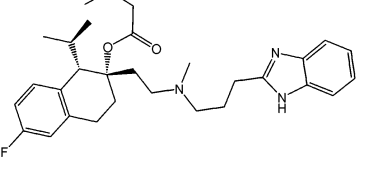
A1. Amoxicillin 	A2. Antipyrine 	A3. Cyclosporine 
A4. Diltiazem 	A5. Enalapril 	A6. Furosemide 
A7. Guanabenz 	A8. Coumarin 	A9. Theophylline 
A10. Epinephrine 	A11. Guanoxan 	A12. Lidocaine 
A13. Tiacrilast 	A14. Sulpiride 	A15. Nitrendipine 
A16. Fleroxacin 	A17. Verapamil 	A18. Bosentan 
A19. Proscillaridin 	A20. Ceftriaxone 	A21. Remikiren 
A22. Saquinavir 	A23. Mibefradil 	

Table 3. Experimental $\log P_{\text{eff}}$ Values for Molecules in the Data Set^a

no.	A	B	C	D	E	F	G	H	I	J	K	L	M	$\log P_{\text{eff}}$
1					-5.35				-6.29					-5.83
2					-4.11									-4.61
3	-5.62							-4.51	-5.04					-5.06
4								-5.70	-6.60					-6.15
5	-4.39	-4.39			-4.12				-4.60					-4.62
6					-3.97									-4.47
7									-4.44					-4.44
8						-4.52								-4.52
9						-5.40								-5.40
10	-6.70	-6.69					-6.3		-6.28					-6.50
11					-4.31									-4.81
12					-4.02									-4.52
13									-5.10					-5.10
14								-4.30	-4.51					-4.41
15								-4.69						-4.69
16									-6.72					-6.72
17									-4.70					-4.70
18							-6.30	-5.51	-5.86					-5.89
19								-4.52	-4.66					-4.59
20	-4.26								-4.67					-4.47
21								-4.67	-4.61					-4.64
22	-4.90		-4.57					-4.63	-4.91					-4.75
23			-6.36											-6.36
24			-5.94											-5.94
25								-4.15	-4.48					-4.32
26									-5.03					-5.03
27								-6.80						-6.80
28								-5.43						-5.43
29									-4.77					-4.77
30	-4.64													-4.64
31				-5.57					-6.42					-5.99
32									-4.44					-4.44
33							-5.82		-6.29					-6.06
34	-4.67			-4.45					-4.85					-4.66
35								-4.28						-4.28
36								-4.85						-4.85
37									-4.69					-4.69
38									-5.03					-5.03
39	-6.75			-5.49				-6.19	-6.42					-6.21
40									-4.71					-4.71
41				-3.88										-3.88
42								-5.92						-5.92
43									-6.16					-6.16
44	-4.57	-4.57							-4.63					-4.59
45									-5.41					-5.41
46				-4.13										-4.13
47									-4.52					-4.52
48									-4.71					-4.71
49	-6.96													-6.96
50					-4.18									-4.68
51					-4.01									-4.51
52									-4.61					-4.61
53									-4.57					-4.57
54									-4.78					-4.78
55									-6.36					-6.36
56									-4.45					-4.45
57	-6.05	-6.05												-6.05
58								-4.36						-4.36
59				-4.10					-4.63					-4.37
60	-4.38	-4.38			-4.08			-4.56	-4.66					-4.58
61					-3.98									-4.48
62								-4.69						-4.69
63									-6.31					-6.31
64	-4.92								-4.66					-4.79
65									-4.93					-4.93
66									-5.77					-5.77
67	-6.89								-6.52					-6.71
68									-4.82					-4.82
69	-6.42								-6.33					-6.38
70	-4.29							-4.14	-4.60					-4.34
71					-4.35				-4.89					-4.85
72					-4.10									-4.60
73									-5.37					-5.37

Table 3 (Continued)

no.	A	B	C	D	E	F	G	H	I	J	K	L	M	$\log P_{\text{eff}}$
74									-5.34					-5.34
75	-4.42								-4.68					-4.55
76									-5.16					-5.16
77										-6.85				-6.85
A1											-6.10			-6.10
A2											-4.55			-4.55
A3											-6.05			-6.05
A4											-4.53	-4.30	-4.31	-4.38
A5											-5.64			-5.64
A6											-6.92		-6.09	-6.51
A7											-4.86		-4.14	-4.50
A8													-4.11	-4.11
A9													-4.35	-4.35
A10													-6.02	-6.02
A11													-4.71	-4.71
A12													-4.21	-4.21
A13													-4.90	-4.90
A14													-6.16	-6.16
A15													-4.77	-4.77
A16													-4.81	-4.81
A17													-4.58	-4.58
A18													-5.98	-5.98
A19													-6.20	-6.20
A20													-6.88	-6.88
A21													-6.13	-6.13
A22													-6.26	-6.26
A23													-4.87	-4.87

^a References: A (Artusson 1990), B (Artursson 1991), C (Haeberlin 1993), D (Rubas 1993), E (Hovgaard 1995), F (Augustijns 1996), G (collett 1996), H (Yee 1997), I (Yazdanian 1998), J (Nicolaas 2001), K (Zhu 2002), L (Saha 2002), M (Camenisch 1998).

Table 4. Symbols and Descriptions of Other Molecular Descriptors Used in Correlation

symbol	descriptor	symbol	description
N_{atom}	number of all atoms	N_{bond}	number of all bonds
N_{aromatic}	number of aromatic atoms	N_{hydro}	number of hydrophobic atoms
f_{roth}	fraction of rotatable bonds	density	mass density
dipole	dipole moment	Hf	heat of formation
W	Weiner index	Z	Hosoya index
Zagreb	Zagreb index	E_{solv}	solvation free energy using the SAWSA model

where m_i is the mass of the i th atom, r_i are the center-of-mass coordinates in the system, and M_T is the total nuclear mass.

It should be noted that some descriptors including V, SASA, and rgyr are conformation-dependent. For all molecules the energy-lowest structures were considered. In preliminary calculations using a set of low-energy conformations from molecular dynamics it was found that the standard deviations of these three descriptors were usually in the order of 5%. We do not consider this as critical for the present analyses. Furthermore, for conformationally flexible compounds the actual conformation penetrating a membrane is difficult to estimate. The ensemble-averaged property of a molecule does not equal to that of the actual conformation in a membrane.

(4) Other Molecular Descriptors. To investigate the potential influence of other molecular descriptors to Caco-2 permeabilities, we introduced many other descriptors to linear correlation. These descriptors include six constitutional and geometrical descriptors, three topological descriptors, two physicochemical descriptors, and one solvation descriptor. The symbol and descriptions of these descriptors are listed in Table 4. The solvation descriptor was calculated using the SAWSA model developed in our group,⁴⁵ and the other descriptors were calculated using a revised drug-BB program developed in our group.³⁸

Statistics of the Models. The statistics of a model was mainly tested by its linear correlation coefficient (r) and leave-one-out cross-validation correlation coefficient (q). Cross-validated q^2 is defined as $q^2 = (\text{SSY} - \text{PRESS}) / \text{SSY}$, where SSY is the sum of squared deviations of the dependent variable values from their mean, and PRESS is the prediction error sum of squares obtained from the leave-one-out cross-validation procedure. Moreover, the standard deviation value (s) and the Fisher value (F) were reported for each model.

RESULTS AND DISCUSSION

Descriptors in QSPR Models. (1) Polar Surface Area.

In 1992, Van de Waterbeemd et al. correlated the PSAs of a series of CNS drugs to membrane transport first.⁴⁶ Thenceforward, PSA has become a very popular parameter for the prediction of molecular transport properties, particularly intestinal absorption and blood-brain barrier penetration. Here, we first carried out a simple linear regression of the 77 training set compounds using polar surface area (PSA) or topological polar surface area (TPSA) as the only descriptor. The resulting equations (eqs 2 and 3) and statistics are listed in Table 6.

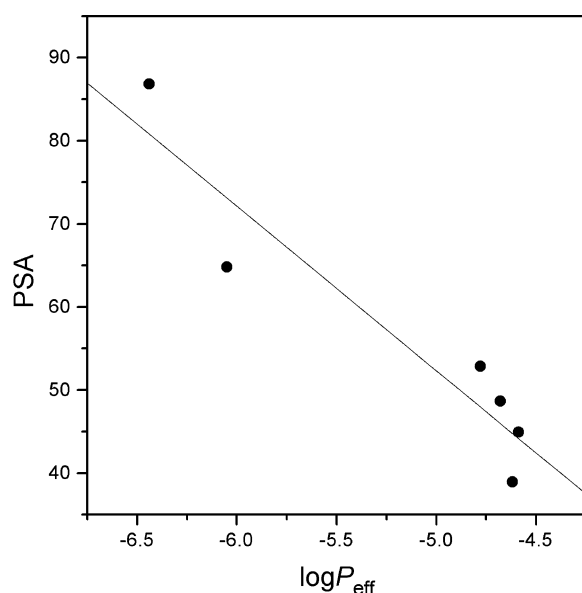
Compared with TPSA, PSA can obtain better correlation with Caco-2 permeabilities. Palm et al. ever performed a correlation between dynamic polar vdW surface areas and

Table 5. Molecules Descriptors and Predicted Values of $\log P_{\text{eff}}$ for Compounds in the Training Set

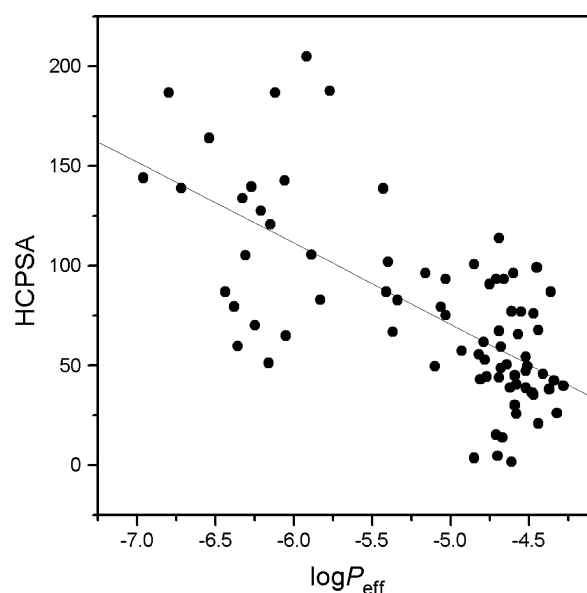
name	$\log P_{\text{eff}}$ (exp)	$\log D$	rgyr	rgyr _d	HCPSA	TPSA	N_{rotb}	$\log P_{\text{eff}}$ (calc)	residue
acebutolol	-5.83	-0.09	4.64	4.51	82.88	87.66	0.31	-5.30	-0.53
acebutolol ester	-4.61	1.59	5.12	5.03	77.08	93.73	0.29	-4.89	0.28
acetylsalic acid	-5.06	-2.25	3.41	3.24	79.38	89.9	0.23	-5.77	0.71
acyclovir	-6.15	-1.80	3.37	3.23	120.63	114.76	0.21	-5.91	-0.24
alprenolol	-4.62	1.38	3.68	3.69	38.92	41.49	0.29	-4.58	-0.04
alprenolol ester	-4.47	2.78	3.84	3.88	35.53	47.56	0.27	-4.39	-0.08
aminopyrin	-4.44	0.63	2.97	2.97	20.81	26.79	0.17	-4.63	0.19
artemisinin	-4.52	2.22	2.75	2.75	54.27	53.99	0.07	-4.47	-0.05
artesanate	-5.40	-0.88	4.02	3.62	102.05	100.52	0.16	-5.64	0.24
atenolol	-6.44	-1.81	4.58	4.52	86.82	84.58	0.29	-5.85	-0.59
betazolol ester	-4.81	0.28	5.41	5.27	43.02	50.72	0.27	-5.20	0.39
betazolol ₁	-4.52	0.63	5.64	5.39	47.14	56.79	0.26	-5.13	0.61
bremazocine	-5.10	1.66	3.43	3.38	49.56	43.7	0.15	-4.57	-0.53
caffeine	-4.41	0.02	2.47	2.47	45.55	58.44	0.12	-4.89	0.48
chloramphenicol	-4.69	1.14	3.75	3.73	113.73	115.38	0.28	-5.11	0.42
chlorothiazide	-6.72	-1.15	3.11	3.11	138.76	118.69	0.08	-5.87	-0.85
chlorpromazine	-4.70	1.86	3.74	3.69	4.60	6.48	0.14	-4.38	-0.32
cimetidine	-5.89	-0.36	4.26	4.24	105.44	88.89	0.33	-5.55	-0.34
clonidine	-4.59	0.78	2.79	2.79	30.03	36.42	0.08	-4.69	0.10
corticosterone	-4.47	1.78	3.68	3.71	75.95	74.6	0.10	-4.78	0.31
desipramine	-4.67	1.57	3.4	3.42	13.8	15.27	0.11	-4.46	-0.21
dexamethas	-4.75	1.89	3.6	3.66	90.74	94.83	0.13	-4.77	0.02
dexamethas- β -D-glucoside	-6.54	0.58	5.67	5.28	163.95	173.98	0.17	-5.83	-0.71
dexamethas- β -D-glucuronide	-6.12	-1.59	5.75	5.23	186.88	191.05	0.17	-6.55	0.43
diazepam	-4.32	2.58	3.28	3.28	25.93	32.67	0.06	-4.45	0.13
dopamine	-5.03	-0.80	2.67	2.68	75.13	66.48	0.23	-5.27	0.24
doxorubici	-6.80	-0.16	4.85	4.9	186.78	206.07	0.18	-6.00	-0.80
erythromycin	-5.43	1.26	4.99	5.01	138.69	193.91	0.21	-5.13	-0.30
estradiol	-4.77	2.24	3.44	3.44	44.34	40.46	0.06	-4.57	-0.20
felodipine	-4.64	3.48	3.39	3.48	50.34	64.63	0.22	-4.44	-0.20
ganciclovir	-6.27	-0.87	3.7	3.48	139.45	134.99	0.25	-5.79	-0.48
griseofulvin	-4.44	2.47	3.37	3.37	67.55	71.06	0.16	-4.59	0.15
hydrochlorothiazide	-6.06	-0.12	3.11	3.11	142.85	118.36	0.08	-5.62	-0.44
hydrocortisone	-4.66	1.48	3.72	3.79	93.37	94.83	0.12	-4.94	0.28
ibuprophen	-4.28	0.68	3.45	3.36	39.86	37.3	0.24	-4.78	0.50
imipramine	-4.85	2.52	3.44	3.45	3.56	6.48	0.13	-4.28	-0.57
indomethacin	-4.69	1.00	4.16	3.16	67.13	68.53	0.19	-5.00	0.31
labetalol	-5.03	1.24	4.61	4.46	93.29	95.58	0.24	-5.09	0.06
mannitol	-6.21	-2.65	2.48	2.59	127.46	121.38	0.44	-5.87	-0.34
meloxicam	-4.71	0.03	3.34	3.36	93.21	99.6	0.16	-5.27	0.56
methanol	-4.58	-0.70	0.84	0.84	25.64	20.23	0.20	-4.67	0.09
methotrexate	-5.92	-2.53	5.33	5.18	204.96	210.54	0.26	-6.79	0.87
methylscopolamine	-6.16	-1.14	3.67	3.74	51.29	59.06	0.16	-5.37	-0.79
metoprolol	-4.59	0.51	4.59	4.53	44.88	50.72	0.30	-4.99	0.4
nadolol	-5.41	0.68	4.37	4.1	86.73	81.95	0.24	-5.15	-0.26
naproxen	-4.83	0.42	3.38	3.43	76.98	46.53	0.19	-5.09	0.26
nevirapine	-4.52	1.81	2.94	2.94	36.68	58.12	0.05	-4.49	-0.03
nicotine	-4.71	0.41	2.5	2.5	15.1	16.13	0.07	-4.65	-0.06
olsalazine	-6.96	-4.50	4.62	4.37	144.08	139.78	0.27	-6.97	0.01
oxprenolol	-4.68	0.45	3.63	3.56	48.62	50.72	0.31	-4.84	0.16
oxprenolol ester	-4.51	1.98	3.87	3.9	49.58	56.79	0.29	-4.45	-0.06
phenacyclidine	-4.61	1.31	2.91	2.91	1.49	3.24	0.04	-4.42	-0.19
phenytoin	-4.57	2.26	2.97	2.97	65.63	58.2	0.06	-4.60	0.03
pindolol	-4.78	0.19	3.71	3.71	52.80	57.28	0.23	-5.02	0.24
pirenzepine	-6.36	-0.46	3.55	3.40	59.71	68.78	0.08	-5.30	-1.06
piroxicam	-4.45	-0.07	3.17	3.26	99.19	99.6	0.13	-5.31	0.86
pnu200603	-6.25	-4.00	3.89	3.79	69.89	91.44	0.15	-6.37	0.12
practolol	-6.05	-1.40	4.02	4.09	64.79	70.59	0.29	-5.50	-0.55
prazocin	-4.36	1.88	4.96	4.99	86.76	106.95	0.15	-5.05	0.69
progesterone	-4.37	3.48	3.58	3.62	38.10	34.14	0.07	-4.54	0.17
propranolol	-4.58	1.55	3.63	3.53	40.42	41.49	0.22	-4.57	-0.01
propranolol ester	-4.48	3.02	4.13	4.06	36.21	47.56	0.22	-4.50	0.02
quinidine	-4.69	2.04	3.25	3.30	43.77	45.59	0.14	-4.46	-0.23
ranitidine	-6.31	-0.12	5.13	4.57	105.15	86.26	0.33	-5.60	-0.71
salicylic acid	-4.79	-1.44	2.14	2.14	61.71	57.53	0.19	-5.29	0.50
scopolamine	-4.93	0.21	3.63	3.49	57.35	62.30	0.15	-5.07	0.14
sucrose	-5.77	-3.34	3.49	3.54	187.69	189.53	0.28	-6.56	0.79
sulfasalazine	-6.33	-0.42	5.68	5.53	133.67	141.31	0.20	-6.06	-0.27
telmisartan	-4.82	2.41	5.29	5.01	55.48	72.94	0.15	-4.85	0.03
terbutaline	-6.38	-1.07	3.15	3.15	79.52	72.72	0.29	-5.36	-1.02
tesosterone	-4.34	3.11	3.33	3.33	42.35	37.3	0.06	-4.53	0.19
timolol	-4.85	0.03	4.02	4.01	100.74	79.74	0.24	-5.35	0.50
timolol ester	-4.60	1.74	3.98	4.13	96.25	85.81	0.23	-4.82	0.22
uracil	-5.37	-1.11	1.84	1.84	66.72	58.20	0.00	-5.23	-0.14
urea	-5.34	-1.64	1.23	1.23	82.72	69.11	0.29	-5.29	-0.05
warfarine	-4.68	0.64	3.45	3.50	59.47	63.60	0.15	-4.95	0.27
zidovudine	-5.16	-0.58	3.14	3.13	96.33	103.59	0.18	-5.43	0.27

Table 6. Multivariate Prediction Models for Caco-2 Permeabilities

$\log P_{\text{eff}} = -4.28 - 0.011\text{PSA}$ ($n=77$, $r=0.664$, $q=0.647$, $s=0.566$, $F=60.0$)	(2)
$\log P_{\text{eff}} = -4.36 - 0.010\text{TPSA}$ ($n=77$, $r=0.636$, $q=0.612$, $s=0.583$, $F=50.9$)	(3)
$\log P_{\text{eff}} = -4.27 - 0.011\text{HCPSA}$ ($n=77$, $r=0.678$, $q=0.661$, $s=0.558$, $F=63.9$)	(4)
$\log P_{\text{eff}} = -4.49 - 0.279n_{\text{HBD}}$ ($n=77$, $r=0.670$, $q=0.647$, $s=0.562$, $F=60.9$)	(5)
$\log P_{\text{eff}} = -4.45 - 0.152n_{\text{HBA}}$ ($n=77$, $r=0.523$, $q=0.484$, $s=0.644$, $F=28.3$)	(6)
$\log P_{\text{eff}} = -5.469 + 0.236\log P$ ($n=77$, $r=0.471$, $q=0.437$, $s=0.657$, $F=24.0$)	(7)
$\log P_{\text{eff}} = -5.265 + 0.310\log D$ ($n=77$, $r=0.706$, $q=0.691$, $s=0.536$, $F=74.6$)	(8)
$\log P_{\text{eff}} = -5.303 + 0.406 < -1.8 < \log D < 2.0 >$ ($n=77$, $r=0.714$, $q=0.698$, $s=0.531$, $F=78.0$)	(9)
$\log P_{\text{eff}} = -5.102 + 0.166\log D_{\text{calc}}$ ($n=44$, $r=0.462$, $q=0.349$, $s=0.723$, $F=11.4$)	(10)
$\log P_{\text{eff}} = -4.719 + 0.281 < -1.8 < \log D < 2.0 > -0.00699\text{HCPSA}$ ($n=77$, $r=0.798$, $q=0.774$, $s=0.460$, $F=64.9$)	(11)
$\log P_{\text{eff}} = -4.513 + 0.359 < -1.8 < \log D < 2.0 > -0.00413\text{HCPSA} - 0.00149\text{MW}$ ($n=77$, $r=0.811$, $q=0.781$, $s=0.450$, $F=46.8$)	(12)
$\log P_{\text{eff}} = -4.300 + 0.325 < -1.8 < \log D < 2.0 > -0.00495\text{HCPSA} - 0.160\text{rgyr}$ ($n=77$, $r=0.817$, $q=0.789$, $s=0.443$, $F=49.0$)	(13)
$\log P_{\text{eff}} = -4.386 + 0.343 < -1.8 < \log D < 2.0 > -0.00530\text{HCPSA} - 0.00131\text{SASA}$ ($n=77$, $r=0.813$, $q=0.782$, $s=0.447$, $F=47.6$)	(14)
$\log P_{\text{eff}} = -4.475 + 0.357 < -1.8 < \log D < 1.8 > -0.00507\text{HCPSA} - 0.00174V$ ($n=77$, $r=0.813$, $q=0.781$, $s=0.448$, $F=47.4$)	(15)
$\log P_{\text{eff}} = -4.277 + 0.330 < -1.8 < \log D < 2.0 > -0.00494\text{HCPSA} - 0.170\text{rgyr}_d$ ($n=77$, $r=0.817$, $q=0.789$, $s=0.444$, $F=48.7$)	(16)
$\log P_{\text{eff}} = -4.392 + 0.252 < -1.8 < \log D < 2.0 > -0.00484\text{HCPSA} - 0.193\text{rgyr} + 1.060f_{\text{roth}}$ ($n=77$, $r=0.824$, $q=0.790$, $s=0.439$, $F=38.1$)	(17)
$\log P_{\text{eff}} = -4.358 + 0.317 < -1.8 < \log D < 2.0 > -0.00558\text{HCPSA} - 0.179\text{rgyr} + 1.074f_{\text{roth}}$ ($n=74$, $r=0.845$, $q=0.812$, $s=0.405$, $F=43.1$)	(18)

**Figure 1.** Correlation between $\log P_{\text{eff}}$ and PSA for six β -adreno-receptor antagonists.

Caco-2 permeability coefficient and found that they were well linearly correlated. Compared with the model developed by Palm et al., eq 2 seems very poor. But it should be noted that in Palm's work, only six β -adrenoreceptor antagonists were used in correlation.¹⁰ If we only used these several molecules, high linear correlation could also be obtained (see Figure 1). These six antagonists bear similar basic structures, so the change of Caco-2 permeabilities may be simply explained by one molecular descriptor. It is obvious that the model based on PSA cannot be treated as a universal principle to predict Caco-2 permeabilities.

**Figure 2.** Correlation between $\log P_{\text{eff}}$ and HCPSA of 77 compounds in the training set.

In our previous work of blood-brain partitioning prediction,³⁸ we proposed the concept of high-charged polar surface area (HCPSA). Using HCPSA, we constructed eq 4 (Table 6). Compared with eq 2, eq 4 shows a better correlation, implying that only high-charged polar surface areas impact the Caco-2 permeation. Figure 2 shows the linear correlation between Caco-2 permeabilities and HCPSA.

Here, we used two descriptors including n_{HBA} and n_{HBD} in MLR, and the resulting equations (eqs 5 and 6) are listed in Table 6. It is interesting to find that the number of hydrogen donor shows a high correlation with Caco-2

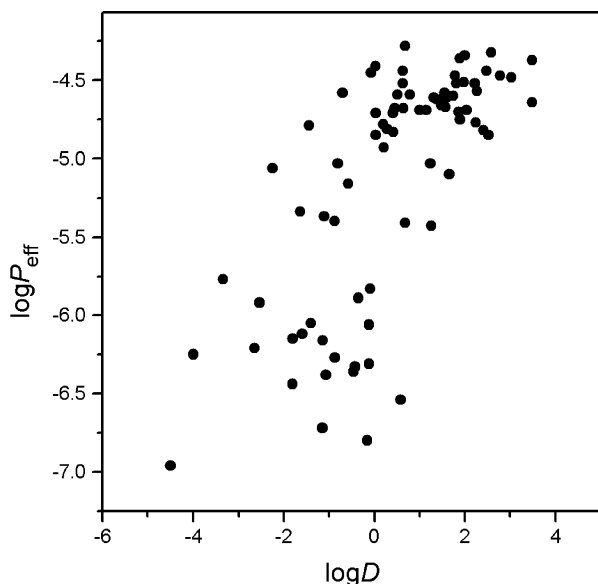


Figure 3. Correlation between $\log P_{\text{eff}}$ and $\log D$ of 77 compounds in the training set.

permeability. The correlation between n_{HBD} and Caco-2 permeabilities is a little worse than that between HCPSA and Caco-2 permeabilities. But compared with PSA, TPSA, and HCPSA, the descriptor of n_{HBA} shows poorer correlations. In fact, the descriptor HCPSA shows significant correlation with n_{HBD} . The correlation between HCPSA and n_{HBD} is 0.84, indicating that the descriptor of hydrogen bonds may be partly replaced by the descriptor HCPSA.

logD. Drug lipophilicity is widely used as a predictor of membrane permeability since it is assumed that drug partitioning into the (lipophilic) cell membranes is a rate-determining process for passive membrane permeation.⁶ The standard for expressing lipophilicity is the partition coefficient P (or $\log P$ to have a more convenient scale) in an octanol/water system. Here, a direct fitting of $\log P$ values with $\log BB$ values of the compounds in the training set produced an r of approximately 0.47 (eq 7 in Table 6).

From regression coefficient, the correlation between $\log P_{\text{eff}}$ and $\log P$ is not very high. In fact, $\log P$ values can only be a first estimate of the lipophilicity of a compound in a biological environment. For many organic molecules, they will exist in a different ionizable state in different pH. To consider the influence of ionization to lipophilicity, the distribution coefficient ($\log D$) instead of $\log P$ was used in fitting (eq 8). The plot of measured Caco-2 permeability data versus experimental $\log D$ values is shown in Figure 3. The distribution coefficients refer to the same pH as the permeability data, and therefore some of the compounds are charged under these conditions.

Compared with eq 7, the correlation coefficient and the Fisher value of eq 8 were improved significantly. Obviously, for partition processes in the body, the distribution coefficient D for which an aqueous buffer at pH 7.4 (blood pH) is used in the experimental determination often provides a more meaningful description of lipophilicity, especially for ionizable compounds.

Some previous efforts to correlate $\log D$ to Caco-2 cell permeability coefficients yielded mixed results. For example, Hilgers et al. have reported a sigmoidal relationship,⁴⁷ whereas Artursson observed a poor linear correlation.³ From

Figure 3, it can be found the relationship between Caco-2 and $\log D$ could not be simply described by a linear regression, while it also cannot be described effectively by a sigmoidal fitting. A sigmoidal curve should possess two plateau regions at the low and the high values of variable. From Figure 3, it can be found that only compounds with high permeabilities form a distinct plateau region. For example, when the $\log D$ values are higher than approximately 2.0, all the compounds have high cellular permeability coefficients independent of the distribution coefficient or net charges of the compounds at pH 7.4. However, for compounds below this threshold value, it seems that more lipophilic compounds have higher Caco-2 cell permeability coefficient in general. So, it is obvious that the influence of $\log D$ to Caco-2 permeabilities will fall into a certain range. To discover the effective range of $\log D$, we applied a two-parameter spline model for $\log D$. The spline model was denoted with angled brackets. For example, $\langle b < \log D < a \rangle$ was equal to $\log D$ if $\log D$ is smaller than a or larger than b ; otherwise, it will be a or b when $\log D$ is larger than a or smaller than b . 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 or b , a systematical search was used to change this value from -3 to 3 using a step of 0.1 . From the calculated results, we could find that the correlation is very sensitive to the variation of b but not a . The best equation is presented as eq 9 in Table 6.

Compared with eq 8, the statistical significance of eq 9 is improved obviously. The threshold values of -1.8 and 2.0 demonstrate when larger $\log D$ values produce higher permeation rates, but the effect takes effect only when distribution coefficient is larger than -1.8 and lower than 2.0 .

In eqs 8 and 9, the experimental $\log D$ values were used. In the development of a theoretical prediction model, we certainly expect that all variables in a model are theoretically derived descriptors, so the prediction is experimentally irrespective. It is interesting to compare the performance of the experimental with calculated $\log D$ values, and thus we performed a correlation between $\log P_{\text{eff}}$ and the predicted $\log D$ values. From the work reported by Opera and Gottfries, we obtained the calculated $\log D$ values of 44 compounds in the training set.⁴⁸ The calculated $\log D$ values were obtained from ACD labs at pH = 7.4.⁴⁹ It should be noted that the calculated $\log D$ value reported by Opera et al. for prazosin may be questionable. We calculated $\log D$ of prazosin by using ACD labs at pH = 7.0, and the obtained value is -0.1 . Using the calculated $\log D$ values reported by Opera and Gottfries, 4 compounds showed deviations greater than 2.0 log unit (olsalazine, ganciclovir, timolol, and mannitol). But the predicted values for olsalazine, ganciclovir, and timolol would be greatly improved when we used ACD/logD version 8.0. The predicted $\log D$ values for olsalazine, ganciclovir, and timolol calculated by using ACD/logD version 8.0 are -4.7 , -2.1 , and -1.4 , respectively. Using the calculated $\log D$ values, the obtained correlation is shown as eq 10 in Table 6.

If we use the experimental $\log D$ values, the linear correlation of eq 10 can be improved significantly ($r=0.693$, $s=0.588$, $F=38.9$). It is obvious that the performance of the calculated $\log D$ values is not satisfactory. Figure 4 shows the linear correlation between experimental and calculated

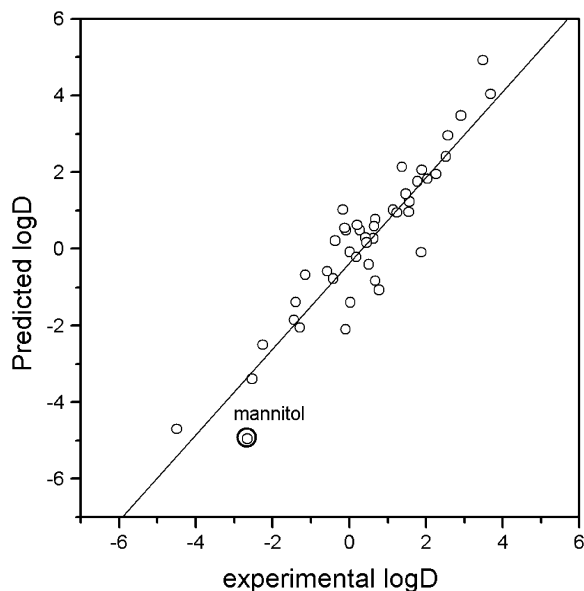


Figure 4. Correlation between experimental and calculated logD values of 44 compounds selected from the training set (compounds with predicted error larger than 2.0 are surrounded by circles).

logD values ($r=0.92$). Really, the ACD logD model predicts well for most of the 44 compounds selected from training set, but as shown in Figure 4, 1 compound showed deviation greater than 2.0 log unit and 7 compounds showed deviations greater than 1.0 log unit. These deviations may be brought by many reasons, such as logP prediction, pK_a prediction, or even experimental errors. Now, several approaches have been developed for pK_a predictions, including ACD/ pK_a (ACD), Pallas/ pK_a (Compudrug), and SPARC.⁵⁰ But until now, these methods cannot provide very reliable prediction for some complicated organic molecules. Obviously, in the prediction of $\log P_{\text{eff}}$ the experimental logD is more reliable than the calculated values.

Then, we considered two parameters including logD and HCPSA in the same equation (eq 11 in Table 6). It is interesting to find that the Caco-2 permeabilities can be effectively described by a combination of logD and HCPSA.

(3) Radius of Gyration. Besides hydrophilicity and hydrophobicity, the bulkiness property of a molecule should be considered. In the literature, molecular weight is often used as a simple accessible molecular size descriptor. First, we introduced molecular weight (MW) to correlation, and the obtained equation is shown as eq 12 in Table 6.

Compared with eq 11, the statistical significance of eq 12 shows obvious improvement. The negative contribution of MW demonstrates that larger molecules are unfavorable to achieve good penetration through Caco-2 monolayers.

Certainly we think that besides molecular weight, the size of a molecule may be described well by other descriptors, such as radius of gyration, molecular surface area, or molecular volume. Here, we introduced radius of gyration, molecular surface area, and molecular volume to linear correlation and constructed three equations (eqs 13–15 in Table 6).

It is obvious that introducing a descriptor concerned with molecular bulkiness, the correlation can be improved. The volume shows the best correlation with MW ($r=0.95$), while radius of gyration shows the poorest correlation with MW ($r=0.77$). Moreover, the correlation between MW and

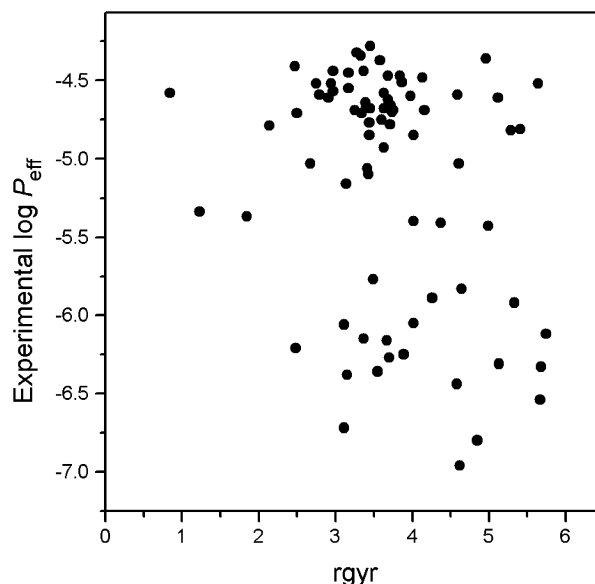


Figure 5. Correlation between rgyr and $\log P_{\text{eff}}$ of 77 compounds in the training set.

molecular surface is high ($r=0.89$). It is obvious that MW reasonably substitutes for V. Among all these descriptors, radius of gyration can give the best description of the shape of a molecule. Judging from the cross-validation coefficient of the linear equations, the best linear model is eq 13. The plot of the measured Caco-2 permeability values versus rgyr is shown in Figure 5. It can be found that the permeability is not simply correlated to the radius of gyration of the compounds ($r=-0.29$). It is not strange considering the difficulties of describing the shape of a molecule. Actually, the shape or bulkiness effect of a molecule cannot be simply described by a simple molecule descriptor.

For small molecules, the influence of molecular flexibility to rgyr may be ignored. However, for larger compounds, such as peptides, it may be better to define a hydrodynamic size or radius. Palm et al. ever proposed the concept of dynamic polar surface area (PSA_d),¹⁰ which is the polar van der Waals surface area Boltzmann-averaged over all low-energy conformers of a molecule. Here, to investigate the influence of molecular flexibility to rgyr, we proposed a concept of dynamic rgyr ($rgyr_d$). The conformational ensemble for each molecule was sampled by using molecular dynamics (MD). All MD simulations were carried out at 300 K with MMFF94 force field.³¹ The time step of the simulations was 1 fs with a cutoff 12 Å for the nonbonded interactions. MD simulation procedures for each molecule involved (1) 5 ps of MD simulations for equilibrium and (2) 200 ps of MD simulations for data collection. In the data collection stage, every 500 fs, the snapshot was recorded in the trajectory file. The dynamic radius of gyration was obtained by averaging the rgyr values over all conformers. Instead of rgyr with $rgyr_d$, we obtained eq 16.

Using $rgyr_d$ in correlation, the correlation was not improved. The correlation coefficient between rgyr and $rgyr_d$ is 0.99, implying that the influence of molecular flexibilities may be negligible.

(4) The Influence of Other Descriptors. The descriptors listed in Table 4 were added into eq 13 in order. From the calculated results, we can find that introducing one descriptor can improve the linear correlation, and the obtained equation

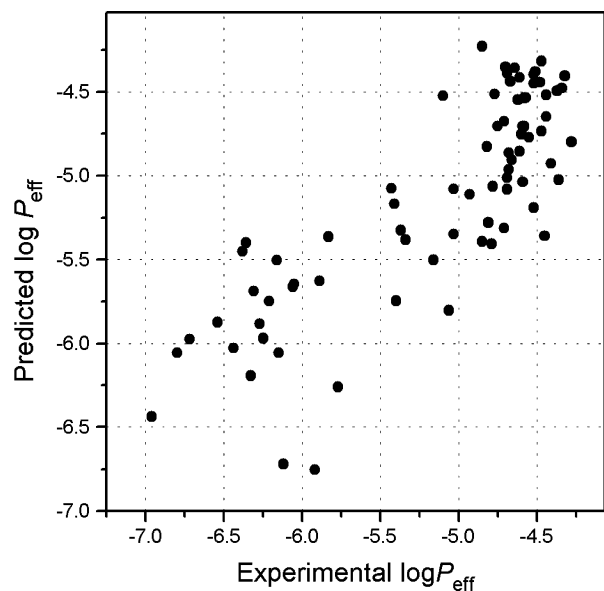


Figure 6. Correlation between experimental and calculated $\log P_{\text{eff}}$ values of 77 compounds in the training set.

is shown as eq 17 in Table 6. But according to leave-one-out cross-validation coefficient, eqs 17 and 13 do not exist with large differences. When we used eqs 13 and 17 to predict the Caco-2 permeabilities of compounds in the test set, it could be found that eq 17 bear better predictive ability. The correlation between experimental $\log P_{\text{app}}$ values and calculated values using eq 17 is 0.78, while that between experimental $\log P_{\text{app}}$ values and calculated values using eq 13 is only 0.73.

The descriptor of f_{rotb} applied in eq 17 defined the flexibility of a molecule. The contributions of the descriptor to $\log P_{\text{eff}}$ demonstrate that high flexible molecule is favorable to Caco-2 permeation. The results are reasonable. Compared with rigid molecules, flexible molecules are easier to adjust its conformation when they penetrate the Caco-2 monolayers.

Equation 17 only includes four descriptors. Moreover, from the calculation of the correlation matrix of the parameters, we found that all descriptors in eq 17 were independent. Although introduction of other descriptors may improve the correlation, the addition of more descriptors may introduce more possibility of random correlation when the training set is limited. The molecular parameters and the predicted Caco-2 permeability coefficients using eq 17 for the compounds in the training set are listed in Table 5. The linear relationships between experimental and predicted $\log P_{\text{eff}}$ values using eq 17 are depicted in Figure 6. In Table 5, it can be found that there are two points (pirenzepin and terbutaline) with predicted error larger than 1.0. If these three compounds were considered as outliers and eliminated from the training set, the statistical parameters of eq 17 were improved effectively (see eq 18).

Certainly, lacking specific information to explain why these molecules behave as outliers, their exclusion from the model was not justified in the present study. Furthermore, the removal of these two compounds had no major effect on the predictions made for test set.

Model Validation. The calculated q (0.78) shows that eq 18 is reliable. Certainly, the high value of LOO q appears to be the necessary but not the sufficient condition for the models to have a good predictive power. Golbraikh et al.

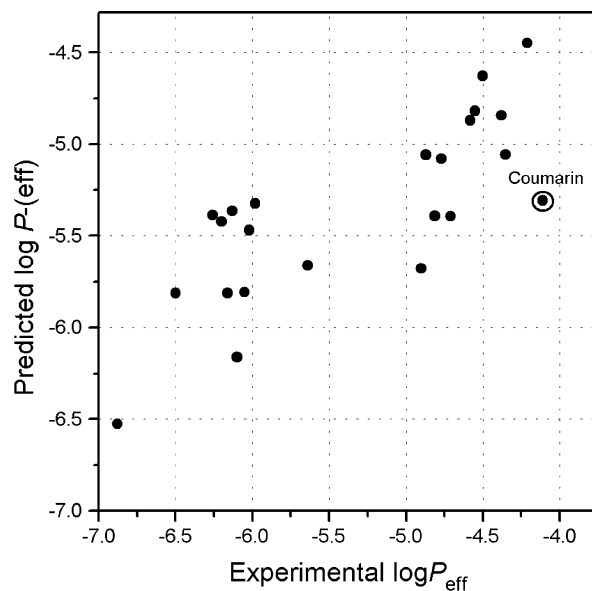


Figure 7. Correlation between experimental and calculated $\log P_{\text{eff}}$ values of 23 compounds in the test set.

ever emphasized that the actual predictive ability of a QSAR model can only be estimated using an external test set of compounds that were not used for building the model.⁵¹ Here, the actual prediction powers of eq 17 were validated by an external test set of 23 compounds. The predicted values using eq 17 are shown in Table 6. The plot of calculated $\log P_{\text{eff}}$ based on eq 17 versus observed values for the tested compounds is shown in Figure 7. Besides coumarin, the predicted error for the other tested molecules is smaller than 1.0. The good predictions for the tested compounds confirms the significance of the selected molecular descriptors and the model based on them.

It is informative to compare the predictive performance of reported $\log P_{\text{eff}}$ models with models developed in this paper, but now with this kind of comparison exists some difficulties. Most of the earlier models were based on a relatively small set of molecules and were not fully validated by external prediction sets. For example, Palm et al. proposed the correlation between $\log P_{\text{app}}$ and PSA_d based on six β -adrenoreceptor blocking agents.¹⁰ Using the same data set, Segarra et al. developed a linear PLS model based on the GRID calculations.⁵² Moreover, PLS regression of MolSurf descriptors was also used to model the dataset modeled by Palm et al.¹² Van de Waterbeemd et al. have developed a set of linear regression models based on 17 compounds.¹¹ The QSAR models developed for 17 compounds had a correlation coefficient less than 0.89. Recently, Fujiwara and co-workers constructed an equation using five molecular descriptors from MO calculations. The authors found that a feed-forward back-propagation neural network was used to improve the correlation. The best linear model proposed by Fujiwara et al. has a regression coefficient of 0.79.¹⁴ The data set used by Fujiwara et al. includes 87 compounds, which is much larger than those used by other researches. We think that the model developed by Fujiwara et al. may be more reliable than some other models, but the predictive ability of the Fujiwara's model was not validated by external tested molecules.

Recently, Ponce et al. proposed several prediction models based on quadratic indices.¹⁶ The training set used by Ponce

Table 7. Molecules Descriptors and Predicted Values of $\log P_{\text{eff}}$ for Compounds in the Test Set

name	$\log P_{\text{eff}}$ (exp)	HCPSA	$\log D$	rgyr	N_{rotb}	$\log P_{\text{eff}}$ (calc)	
						eq 18 ^a	Ponce's model ^b
amoxicillin	-6.10	158.58	-1.69	4.20	0.22	-6.16	
antipyrine	-4.55	20.49	0.34	2.74	0.11	-4.82	
cyclosporine	-6.05	181.48	2.31	6.55	0.21	-5.81	
diltiazem	-4.38	73.52	2.22	4.15	0.19	-4.84	-3.17
enalapril	-5.64	91.59	-0.90	4.48	0.25	-5.66	
furosemide	-6.50	125.96	-1.21	3.77	0.21	-5.81	-8.74
guanabenz	-4.50	75.04	1.67	3.01	0.27	-4.63	-6.68
coumarin	-4.11	26.91	-1.39	2.25	0.00	-5.31	-3.17
theophylline	-4.35	61.59	-0.02	2.36	0.09	-5.06	-4.65
epinephrine	-6.02	77.99	-2.59	2.76	0.27	-5.47	-6.44
guanoxan	-4.71	84.07	-0.83	2.93	0.17	-5.39	-6.69
lidocaine	-4.21	24.46	1.63	3.23	0.26	-4.45	-4.83
tiacrilast	-4.90	100.57	-1.05	3.92	0.21	-5.68	-3.89
sulpiride	-6.16	110.95	-1.15	4.23	0.21	-5.81	-7.76
nitrendipine	-4.77	99.98	0.97	3.58	0.23	-5.08	-4.87
fleroxacin	-4.81	63.28	-0.29	3.93	0.13	-5.39	-3.95
verapamil	-4.58	61.58	2.07	5.14	0.29	-4.87	-3.17
bosentan	-5.98	117.36	1.30	4.84	0.23	-5.32	-6.03
proscillaridin	-6.20	116.19	2.48	5.69	0.12	-5.42	-5.63
ceftriaxone	-6.88	220.00	-1.23	5.2	0.23	-6.53	-8.03
remikiren	-6.13	131.95	2.75	5.66	0.24	-5.36	-8.33
saquinavir	-6.26	130.93	3.00	5.58	0.20	-5.39	-9.32
mibefradil	-4.87	51.90	3.66	5.91	0.21	-5.06	-4.83
<i>r</i>						0.78	0.79
UME ^c						0.49	1.15

^a Prediction using eq 18. ^b Prediction using the Ponce's model.
^c UME represents unsigned mean error.

et al. is the same 17 molecules used by Van de Waterbeemd et al. The best model bears very good statistical significantly ($r=0.93$). Moreover, authors used an external test set of 20 drugs to evaluate the predictive power of the model. The 20 drugs are also used in our work. Among these 20 drugs, one (imipramine) is included in the training set, and the others are included in the test set. The predicted values for the tested molecules using the Ponce's model are listed in Table 7. Inspection of performance of our model shows that it can perform reasonably well for most compounds, and only one compound (coumarin) was overestimated. Using the Ponce's model, there are 10 compounds with a predicted error larger than 1.0. Considering all 19 common compounds, the unsigned mean error using eq 17 is 0.56, which is significantly better than that using the Ponce's model (UME=1.15). Although the statistical parameters of the Ponce's model are very good, its predicted power is very poor. We think that the data set used by Ponce et al. is so limited. The rather limited data do not cover a wide range chemical space and necessarily implied a larger sample size. In a much smaller region of chemical space, some local models may produce good fitting between property and descriptors, but the local models may not have the capability to extrapolate predictions into regions of chemical space for which the modes have no data and where the relationships between property and descriptors may change. For example, if we performed a correlation between the $\log P_{\text{eff}}$ values of the 17 compounds used by Van de Waterbeemd et al.¹¹ and the four descriptors used in this paper, the correlation coefficient is higher than 0.91. So based on the limited data, the high statistical qualities of the models are sometimes caused by random correlation or uncertain factors, and then the obtained models are not very meaningful from the viewpoint of prediction.

CONCLUSION

In the current work, based on a large set of drug or drug-like molecules linear correlation models were developed to estimate Caco-2 permeability data. Four molecules descriptors are crucial to Caco-2 permeation: experimental distribution coefficient, $\log D$; high-charged polar surface areas based on Gasteiger partial charges, HCPSA; the radius of gyration, rgyr; and the fraction of rotatable bonds. These four descriptors give meaningful physical picture of the molecular mechanisms involved in Caco-2 permeation: a hydrophobic molecule can penetrate Caco-2 monolayers easier, but when the distribution coefficient is larger than 2.0, the influence of hydrophobicity is not very significant; larger polar surface areas have more negative contribution to Caco-2 permeability values, but the contributions are only limited to those atoms with high-charge densities; larger and rigid molecules will lead to worse Caco-2 cell penetration ability. The predictions to the external test set demonstrate that this model bears good performance and can be used for estimation of Caco-2 permeability for drug and drug-like molecules.

Certainly, the models proposed in the current work are far from perfect, because the data set used here is limited. A sufficiently large set of experimental data relating to this endpoint for the validation is crucial in the development of the prediction models. So, based on increasing data, the learning/modeling will need to be an ongoing, iterative process in which the models are continuously refined.

ACKNOWLEDGMENT

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

Supporting Information Available: The structures in the training set and the test set (saved in MACCS/SDF database format) and all molecular descriptors used in predictions model in the SDF database files. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Kennedy, T. Managing the drug discovery/development interface. *Drug Discov. Today* **1997**, 2, 436–444.
- (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 Discov.* **2000**, 3, 30–41.
- (3) Artursson, P. Cell Cultures as Models for Drug Absorption Across the Intestinal Mucosa. *Cri. Rev. Ther. Carrier Syst.* **1991**, 8, 305–330.
- (4) Auaroni, A.; Hochman, J. Development of Intestinal Cell Culture Models for Drug Transport and Metabolism Studies. *Adv. Drug Del. Rev.* **1996**, 22, 3–52.
- (5) Artursson, P.; Karlsson, J. Correlation between Oral Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cells. *Biochem. Biophys. Res. Com.* **1991**, 175, 880–885.
- (6) Stenberg, P.; Luthman, K.; Artursson, P. Virtual Screening of Intestinal Drug Permeability. *J. Controlled Release* **2000**, 65, 231–243.
- (7) Egan, W. J.; Lauri, G. Prediction of Intestinal Permeability. *Adv. Drug Delivery Rev.* **2002**, 54, 273–289.
- (8) Chan, O.; Stewart, B. H. Physicochemical and drug-delivery consideration for oral drug bioavailability. *Drug Discov. Today* **1996**, 1, 461–473.
- (9) Conradi, R. A.; Burton, P. S.; Borchardt, R. T. Physicochemical and biological factors that influence a drug's cellular permeability by passive diffusion. *Method Princ. Med. Chem.* **1996**, 4, 233–252.
- (10) Palm, K. P.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, A. P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, 85, 32–39.
- (11) Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.* **1996**, 15, 480–490.

- (12) Norinder, U.; Osterber, T.; Artursson, P. Theoretical Calculation and Prediction of Caco-2 Cell Permeability Using MolSurf Parametrization and PLS Statistics. *Pharm. Res.* **1997**, *14*, 1786–1791.
- (13) 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.
- (14) Fujiwara, S. I.; Yamashita, F.; Hashida, M. Prediction of Caco-2 Cell Permeability Using a Combination of MO-Calculation and Neural Networks. *Int. J. Pharm.* **2002**, *237*, 95–105.
- (15) Kulkarni, A.; Han, Y.; Hopfinger, J. Prediction Caco-2 Cell Permeation Coefficients of Organic Molecules Using Membrane-Interaction QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 331–342.
- (16) Ponce, Y. M.; Cabrera Pérez, M. A.; Zaldivar, V. R.; Ofori, E.; Montero, L. A. Total and Local Quadratic Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”. Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003**, *4*, 512–536.
- (17) Artursson, P. Epithelial Transport of Drugs in Cell Culture. I: A Model for Studying the Passive Diffusion of Drug Over Intestinal Absorptive (Caco-2) Cell. *J. Pharm. Sci.* **1990**, *79*, 476–482.
- (18) Artursson, P.; Karlsson, J. Correlation Between Oral Drugs Absorption in Humans and Apparent Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cell. *Biochem. Biophys. Res. Com.* **1991**, *175*, 880–885.
- (19) Haeberlin, B.; Rubas, W.; Nolen, H. W. III; Friend, D. R. In Vitro Evaluation of Dexamethasone- β -D-Glucuronide for Colon-Specifics of a Human Colonic Epithelial (Caco-2) Cell Line to Colon of Rabbit, Monkey, and Dog Intestine and Human Drug Absorption. Drug Delivery. *Pharm. Res.* **1993**, *10*, 1553–1562.
- (20) Rubas, W.; Jezyk, N.; Grass, G. M. Comparison of the Permeability Characteristics. *Pharm. Res.* **1993**, *10*, 113–118.
- (21) Hovgaard, L.; Brøndsted, H.; Buur, A.; Bundgaard, H. Drug Delivery Studies in Caco-2 Monolayers. Synthesis, Hydrolysis, and Transport of O-Cyclopropane Carboxylic Acid Ester Prodrugs of Various β -Blocking Agents. *Pharm. Res.* **1995**, *12*, 387–392.
- (22) Augustijns, P.; D’Hulst, A.; Daele, J. V.; Kinget, R. Transport of Artemisinin and Sodium Artesunate in Caco-2 Intestinal Epithelial Cells. *J. Pharm. Sci.* **1996**, *85*, 577–579.
- (23) Collett, A.; Sims, E.; Walker, D.; He, Y. L.; Ayrton, J.; Rowland, M.; Warhurst, G. Comparison of HT29–18-C-1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. *Pharm. Res.* **1996**, *13*, 216–221.
- (24) Yee, S. In Vitro Permeability Across Caco-2 Cells (Colonic) Can Predict In Vivo (Small Intestinal) Absorption in Man-Fact or Myth. *Pharm. Res.* **1997**, *14*, 763–766.
- (25) Yazdani, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating Partitioning and Caco-2 Cell Permeability of Structurally Diverse Small Molecular Weight Compounds. *Pharm. Res.* **1998**, *15*, 1490–1494.
- (26) Nicolaas, G. M. Schipper; Österberg, T.; Wrangé, U.; Westberg, C.; Sokolowski, A.; Rai, R.; Young, W.; Sjöström. In Vitro Intestinal Permeability of Factor Xa Inhibitors: Influence of Chemical Structure on Passive Transport and Susceptibility to Efflux. *Pharm. Res.* **2001**, *18*, 1735–1741.
- (27) Zhu, C. Y.; Jiang, L.; Chen, T. M.; Hwang, K. K. A Comparative Study of Artificial Membrane Permeability Assay for High Throughput Profiling of Drug Absorption Potential. *Eur. J. Med. Chem.* **2002**, *37*, 399–407.
- (28) Saha, D.; Kou, J. H. Effect of Bovine Serum Albumin on Drug Permeability Estimation Across Caco-2 Monolayers. *Eur. J. Pharm. Biopharm.* **2002**, *54*, 319–324.
- (29) Camenisch, G.; Alsenz, J.; Waterbeemd, Han van de; Folkers, G. Estimation of Permeability by Passive diffusion through Caco-2 Cell Monolayers Using the Drug’s Lipophilicity and Molecular Weight. *Eur. J. Pharm. Sci.* **1998**, *6*, 313–319.
- (30) SYBYL 6.5 User Guide; Tripos Inc., St. Louis, MO, 1999.
- (31) Halgren, T. A. Merck molecular force field .1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (32) Artursson, P.; Palm, K.; Luthman, K. Caco-2 Monolayers in Experimental and Theoretical Predictions of Drug Transport. *Adv. Drug Deliv. Rev.* **1996**, *22*, 67–84.
- (33) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877.
- (34) Balon, K.; Riebeschl, B. U.; Müller, B. W. Determination of liposome partitioning of ionizable drugs by titration. *Pharm. Res.* **1999**, *16*, 882–888.
- (35) Derety, E.; Feher, M.; Schmidt, J. M. Rapid prediction of human intestinal absorption. *Quant. Struct.-Act. Relat.* **2002**, *21*, 493–506.
- (36) 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.
- (37) Sanner, M. F.; Olson, A. J.; Spohner, Reduced surface: An efficient way to compute molecular surfaces. *J. Biopolymers* **1996**, *38*, 305–320.
- (38) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.
- (39) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity – A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (40) 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.
- (41) 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.
- (42) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual Daylight 4.62, Daylight Chemical Information Systems, Inc., Los Altos, 2001.
- (43) Walter, A.; Gutknecht, J. Permeability of small nonelectrolytes through lipid bilayer membranes. *J. Membr. Biol.* **1986**, *90*, 207–217.
- (44) Leahy, D. E.; Lynch, J.; Taylor, D. C. Mechanisms of absorption of small molecules. In *Novel drug delivery and its therapeutic application*; Prescott, L. F., Nimmo, W. S., Eds.; John Wiley Sons: New York, pp 33–44.
- (45) Hou, T. J.; Qiao, X. B.; Zhang, W.; Xu, X. J. Empirical aqueous solvation models based on accessible surface areas with implicit electrostatics. *J. Phys. Chem. B* **2002**, *106*, 11295–11304.
- (46) Van der Waterbeemd, H.; Kansy, M. Hydrogen-bonding capacity and brain penetration. *Chimia* **1992**, *46*, 299–303.
- (47) Hilgers, A. R.; Conradi, R. A.; Burton, P. S. Caco-2 cell monolayers as a model for drug transport across the intestinal mucosa. *Pharm. Res.* **1990**, *7*, 902–910.
- (48) Opera, T. I.; Gottfries, J. Toward minimalistic modeling of oral drug absorption. *J. Mol. Graph. Model.* **1999**, *17*, 261–274.
- (49) See the ACD labs Web page <http://www.acdlabs.com/products/physchem-lab/logd/for details>.
- (50) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. A rigorous test for SPARC’s chemical reactivity models: estimation of more than 4300 ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.
- (51) Golbraikh, A.; Tropsha, A. Beware of q²!. *J. Mol. Graphic Modell.* **2002**, *20*, 269–276.
- (52) Segarra, V.; López, M.; Ryder, H.; Palacios, J. M. Prediction of drug permeability based on grid calculations. *Quant. Struct.-Act. Relat.* **1999**, *18*, 474–481.

CI049884M