

Tingjun Hou · Xiaojie Xu

ADME evaluation in drug discovery

1. Applications of genetic algorithms to the prediction of blood–brain partitioning of a large set of drugs

Received: 17 June 2002 / Accepted: 12 September 2002 / Published online: 18 December 2002
© Springer-Verlag 2002

Abstract In this study, the relationships between the brain–blood concentration ratio of 96 structurally diverse compounds with a large number of structurally derived descriptors were investigated. The linear models were based on molecular descriptors that can be calculated for any compound simply from a knowledge of its molecular structure. The linear correlation coefficients of the models were optimized by genetic algorithms (GAs), and the descriptors used in the linear models were automatically selected from 27 structurally derived descriptors. The GA optimizations resulted in a group of linear models with three or four molecular descriptors with good statistical significance. The change of descriptor use as the evolution proceeds demonstrates that the octane/water partition coefficient and the partial negative solvent-accessible surface area multiplied by the negative charge are crucial to brain–blood barrier permeability. Moreover, we found that the predictions using multiple QSPR models from GA optimization gave quite good results in spite of the diversity of structures, which was better than the predictions using the best single model. The predictions for the two external sets with 37 diverse compounds using multiple QSPR models indicate that the best linear models with four descriptors are sufficiently effective for predictive use. Considering the ease of computation of the descriptors, the linear models may be used as general utilities to screen the blood–brain barrier partitioning of drugs in a high-throughput fashion.

Keywords Blood–brain partitioning · ADME · Genetic algorithm · QSPR

Introduction

In the case of effective central nervous system (CNS) acting drugs, the knowledge of the penetration of drugs

through the blood–brain barrier (BBB) is critical to screen potential therapeutic agents and improve the side effect profile of drugs with peripheral activity. [1] The BBB is a complex physical and biochemical interface, which morphologically is based on tightly jointed blood capillary endothelial cells. [2] For penetration regulated by the BBB, passive diffusion is the dominant process of cerebrovascular transport. At the molecular level, the principal component of the barrier is the lipid bilayer of the capillary endothelial cell membrane, through which compounds must diffuse to reach the brain.

In experiments, the relatively affinity for the blood or brain tissue can be expressed in terms of the blood–brain partition coefficient, $\log BB = \log(C_{\text{brain}}/C_{\text{blood}})$, where C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood respectively. Both in vivo and in vitro experiments have been conducted that determined $\log BB$. [3, 4, 5] However, both these methodologies are laborious, expensive and time-consuming and require a sufficient quantity of the pure compounds, often in radiolabeled form, to obtain reliable experimental data. They are hence not amenable to high-throughput screening of therapeutic candidates. Thus, theoretical and computational methodologies to predict the blood–brain coefficient would have a great impact on drug research and development.

Several attempts to correlate BBB transport with physico-chemical descriptors, in particular with the octanol–water partition coefficient, as $\log P$, have been reported. [6, 7] However, in some cases, such as in the brain penetration by H₂-receptor histamine antagonist, $\log P$ shows poor correlation with $\log BB$. [8] In addition to $\log P$, the descriptors concerned with molecular size and hydrogen bond formation have also been found to be important contributors to $\log BB$. [9, 10, 11, 12] Unfortunately, these earlier models were based on a small set of molecules and were not fully validated by external prediction sets. Using a larger dataset, Lombardo established a correlation between $\log BB$ and the solvation-free energy from semiempirical quantum chemical calculations. [13]

T. Hou · X. Xu (✉)
College of Chemistry and Molecular Engineering,
Peking University, Beijing 100871, PR China
e-mail: xiaojxu@chem.pku.edu.cn

Norinder et al. [14] used MolSurf [15] parameterization to calculate various properties related to the molecular valence region, and combined it with the partial least squares to latent structures (PLS) method to develop a QSPR of log BB with three statistically significant components. Luco also employed the PLS technique to develop a QSPR based on several topological and constitutional descriptors. [16] More recently, Cruciani et al. applied a new technique, Volsurf, to transform 3D molecule fields into descriptors and correlated them to the experimental permeation by the PLS procedure. [17] However, the PLS method generally appears to strip the QSPR of explicit physical insight, and the determination of the principle components of numerous physicochemical descriptors cannot be easily calculated for an arbitrary compound.

Recently, Kaznessis et al. constructed a QSPR with a physically significant descriptor for 76 structurally diverse molecules. [18] Monte Carlo simulations of compounds in water were performed to calculate such properties as the solvent-accessible surface area (SASA), the solute dipole, and the hydrophilic, hydrophobic and amphiphilic components of the SASA. But in the statistical analysis, nine strong outliers were identified and subjectively removed in order to gain better correlation. Moreover, the quantity of the QSPR models obtained was only judged by the coefficient of linear regression, not by external prediction sets or even cross-validation.

It is well known that BBB permeation depends on multiple factors. H-bonding capacity, local hydrophobicity, molecular size, lipophilicity and even flexibility are important parameters that play important roles in BBB permeation. A molecule can be described by many molecular descriptors. Moreover, many kinds of molecular descriptors are interconnected, and the information in a molecule may be duplicated over many of the descriptors. Thus, it is relatively difficult for us to select the most appropriate descriptors to construct the best QSPR models. In this paper, a QSPR analysis technique based on a GA developed in our group has been applied to optimize the process of descriptor selection. [19, 20, 21] The QSAR method based on a GA was first proposed by Rogers and Hopfinger [22] and Kubinyi. [23, 24] QSAR or QSPR based on a GA can find a group of reliable QSAR or QSPR models from a large number of samples very efficiently. Moreover, from the analysis of the variable use as the evolution proceeds, we may obtain the crucial molecular properties relevant to activity. To the best of our knowledge, no attempt has been made to use a GA to derive multiple QSPR models to predict BBB permeation. The main purpose of this work is to obtain a group of QSPR models based on simple molecular structural descriptors that could be used to predict BB ratios for a wide range of new therapeutic agents.

Materials and methods

Biological data

The brain–blood concentration ratios for the 59 compounds comprising the training set were taken from the studies reported by Young et al. [8] (1–30, part h in Table 1) and Abraham et al. [11] (31, part h in Table 1, C1–C23, part f in Table 2). In addition, five acidic compounds were included in the training set (32–36, part i in Table 1), and their log BB values were taken from Salminen et al. [10] and Greenwood et al. [25]

Table 1 Some compounds used to obtain the training set

1	
2	
3	
4	
5	
6	
7	

Table 1 (continued)

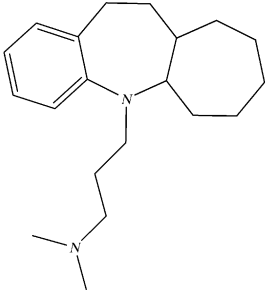
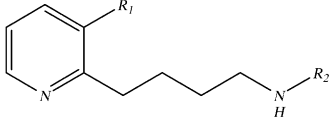
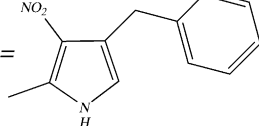
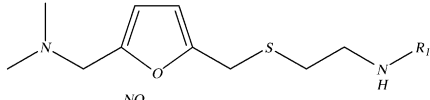
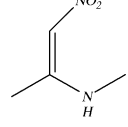
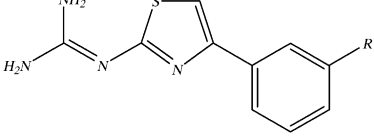
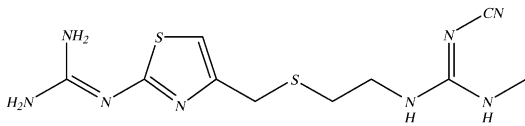
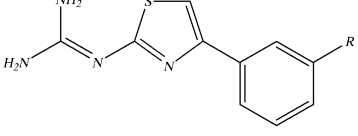
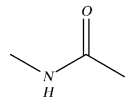
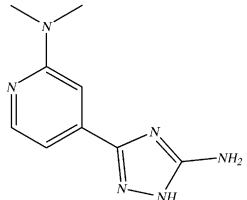
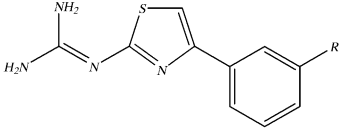
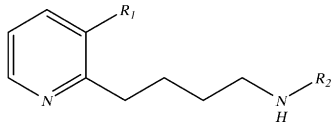
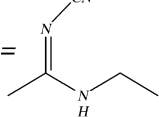
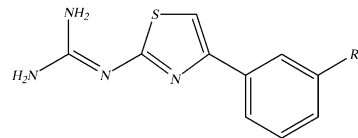
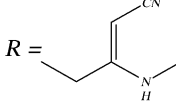
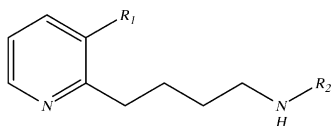
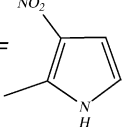
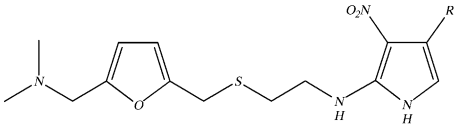
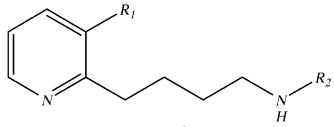
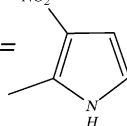
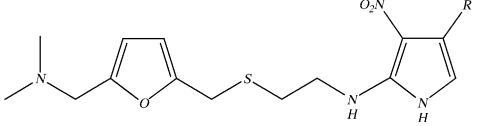
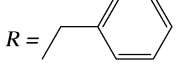
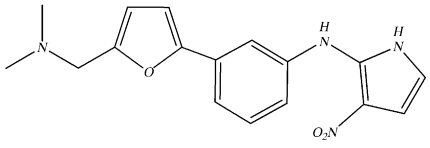
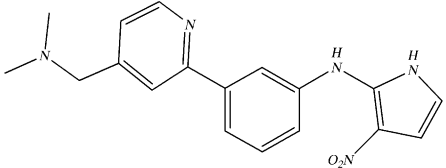
8		15	 <p>$R_1 = H, R_2 =$ </p>
9	 <p>$R_1 =$ </p>	16	 <p>$R = H$</p>
10		17	 <p>$R =$ </p>
11		18	 <p>$R = NH_2$</p>
12	 <p>$R_1 = Br, R_2 =$ </p>	19	 <p>$R =$ </p>
13	 <p>$R_1 = Br, R_2 =$ </p>	20	 <p>$R = H$</p>
14	 <p>$R_1 = H, R_2 =$ </p>	21	 <p>$R =$ </p>
22		23	

Table 1 (continued)

24	 $R =$
25	 $R =$
26	
27	 $R =$
28	 $R =$
29	 $R =$
30	 $R =$
31	

32	
33	
34	
35	
36	

Two validation sets of compounds were used in the current work. The first test set comprises 14 compounds from several literature sources (**B1–B14**, part a, b, c, d in Table 3). [11, 13; 26] The other validation set comprises 23 drugs collected by Salminen et al. [10] This data set contains acidic, basic and neutral drugs from various structural classes, and the calculations were performed using the neutral forms of the molecules.

Molecular modeling and molecular descriptors

Molecular models of compounds were constructed in the SYBYL molecular simulation package. [27] The initial structures were first minimized using molecular mechanics with the MMFF94 force field, [28] and the terminal condition was the RMS of potential energy smaller than $0.001 \text{ kcal } \text{Å}^{-1} \text{ mol}^{-1}$. For these flexible compounds, conformational analyses were performed to determine the most stable conformers.

In the current work, a total of 27 descriptors were used in the QSPR analysis. These 27 descriptors belong to three categories: spatial descriptors, structural descriptors and thermodynamic descriptors. The following indices concerned with molecular thermodynamic properties were considered: A $\log P$ proposed by Ghose and Crippen, [29] the 1-octanol desolvation free energy (Foct) and the aqueous desolvation free energy (Fh2o) derived from a hydration shell model developed by Hopfinger et al. [30] and molar refractivity proposed by Ghose and Crippen. [31]

To quantify the spatial characteristics of the compounds under study, 17 descriptors were considered, including radius of gyration, Connolly surface area, molecular volume, molecular density and 13 Jurs descriptors introduced by Stanton and Jurs. [32] The partial charges used in the determination of some Jurs descriptors were computed using the charge equilibration method proposed by Rappé and Goddard. [33] Several other important structural vari-

Table 2 Compounds comprising test set 2

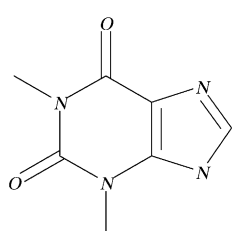
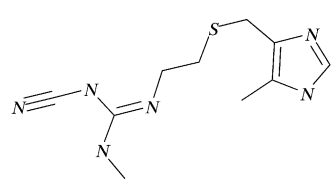
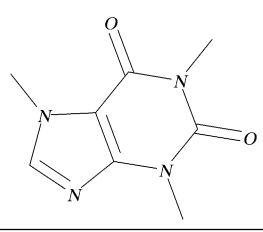
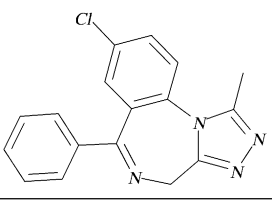
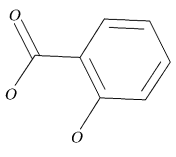
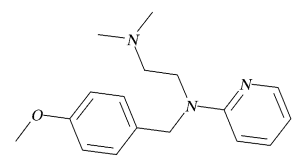
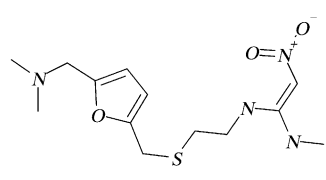
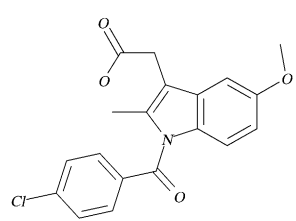
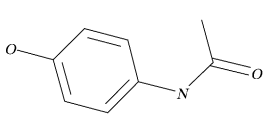
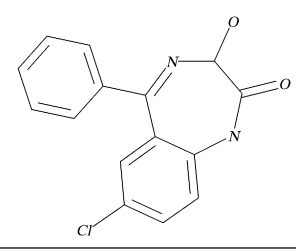
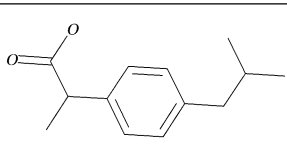
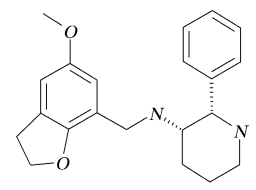
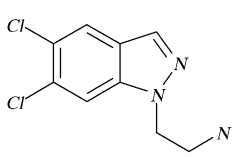
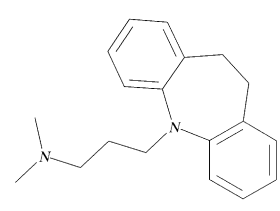
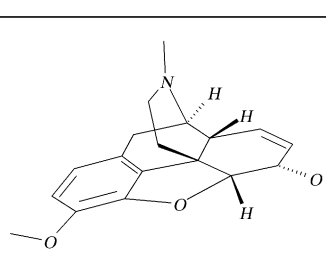
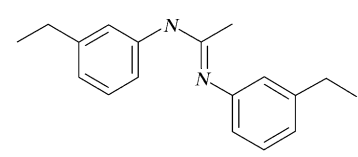
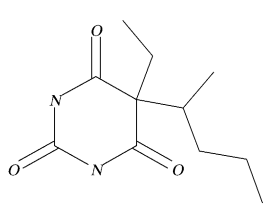
C1		C10	
C2		C11	
C3		C12	
C4		C13	
C5		C14	
C6		C15	
C7		C16	
C8		C17	
C9			

Table 2 (continued)

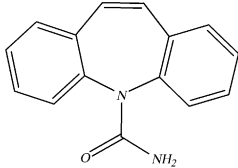
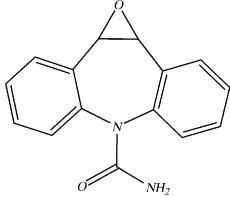
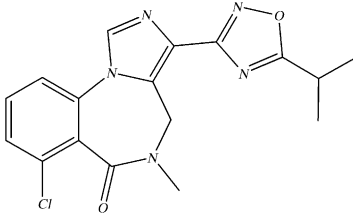
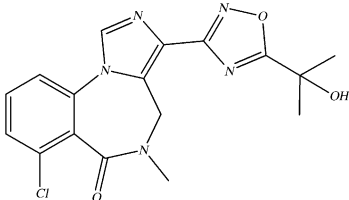
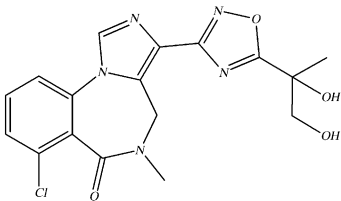
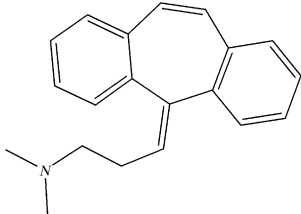
C18	
C19	
C20	
C21	
C22	
C23	

ables were also considered, including the molecular weight, the number of rotatable bonds, the number of hydrogen bond acceptors and the number of hydrogen bond donors. All descriptors used in this paper can be computed easily and rapidly, and all descriptors have obvious physico-chemical and structural meanings. Parameters that need large computations based on quantum chemistry calculations were not considered here. A list of all descriptors considered in this study is given in Table 4.

Table 3 Compounds comprising test set 1

B1	
B2	
B3	
B4	
B5	
B6	
B7	
B8	

Table 3 (continued)

B9	
B10	
B11	
B12	
B13	
B14	

Statistical analysis

The QSPR analysis program used in this paper is under development in our laboratory. [19, 20, 21] Compared with other traditional statistical methods, QSAR or QSPR based on a GA uses a population of many models and tests only the final, fully constructed models. The brief procedure of QSPR based on a GA involves five steps: creation of the initial population, crossover operation, mutation operation, comparison operation and partial reinitialization operation.

According to the GA, an individual should be represented as a linear string, which plays the role of the DNA for the individual, so

a series of descriptors are randomly chosen as a string. Every descriptor is expressed using two digits, one represents its serial number and another represents its function type. The initial population is generated by randomly selecting some numbers of descriptors from the training set. These individuals are then scored according to their fitness score. An elite population is used to retain the best and different individuals. Once all models in the population have been rated using the fitness score, the crossover operation is performed repeatedly. In the operation, two good models are probabilistically selected as "parents" with the likelihood of being chosen proportional to a model fitness score. A pair of children are produced by dividing both parents at a randomly chosen point and then joining the pieces together. After the crossover operation, the mutation operation may randomly alter all individuals in the new population, and the new model's fitness is determined. After the crossover and mutation operations, the newly created population and the elite population are compared. If some individuals in the newly created population are better than some in the elite population, we copy these better individuals to the elite population. When the total fitness of the elite population cannot be improved, "convergence" is achieved. After several steps of crossover and mutation operations, a partial reinitialization procedure is easily introduced into the GA by replacing the lowest 50–80% of chromosomes in the population with randomly generated ones in order to reduce the likelihood of the GA converging on a local-optimal minimum.

Upon completion, from the elite population, the models with the highest fitness score can be obtained. The models in the elite population are sorted by their fitness scores. In this study, the fitness function was defined as the multiple linear regression coefficient (r). The reliability of the models was fully tested by their leave-one-out cross-validated correlation coefficient (q) scores and the actual prediction on the test sets. Cross-validated q^2 is defined as: $q^2 = (SSY - PRESS) / SSY$, 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.

Results and discussion

The data set contains 59 compounds and 27 molecular descriptors. The four-term and five-term multiple linear correlation models were constructed. The use of more than six independent variables for this data set may not be appropriate because of possible chance correlation, so these were omitted from our models. For this data set, 200 populations were used, and the number of elite populations was defined as 100. The genetic evolutions were applied sufficiently to guarantee the convergence of the calculations. In the current work, 4,000 genetic operations were used. Moreover, a partial reinitialization procedure was applied after 100 crossover operations. After calculations, the 100 best linear models for four descriptors and three descriptors could be obtained, respectively.

Usage of descriptors

During the process of genetic evolution, the use of descriptors in the elite population changes clearly. Figure 1 shows the change of descriptor use as the evolution proceeds in the elite populations with four descriptors. The continuing change in the use suggests that further evolution of the population may be warranted. In Fig. 1, it can be seen that after convergence the frequencies appearing in these models in the elite population are quite different.

Table 4 The descriptors used in the QSPR analysis

PPSA-1	Sum of the solvent-accessible surface areas of all positively charged atoms
PNSA-1	Sum of the solvent-accessible surface areas of all negatively charged atoms
PPSA-2	Partial positive solvent-accessible surface area multiplied by the total positive charge
PNSA-2	Partial negative solvent-accessible surface area multiplied by the negative charge
DPSA-1	Partial positive solvent-accessible surface area minus partial negative solvent-accessible surface area (DPSA-1)
DPSA-2	Total charge weighted positive solvent-accessible surface area minus total charge weighted negative solvent-accessible surface area
RPCG	Relative positive charge: charge of most positive atom divided by the total positive charge
RNCG	Charge of most negative atom divided by the total negative charge
TASA	Sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2
RPSA	Total polar surface area divided by the total molecular solvent-accessible surface area
SASA	Total molecular solvent-accessible surface area
TPSA	Sum of solvent-accessible surface areas of atoms with absolute value of partial charges greater or equal than 0.2
RASA	Total hydrophobic surface area divided by the total molecular solvent-accessible surface area
RadOfGyration	Radius of gyration
Area	Molecular surface area
Vm	Molecular volume
Density	Density
$A \log P$	Ghose and Crippen octanol/water partition coefficient
Foct	Desolvation free energy for octanol
Fh2o	Desolvation free energy for water
Apol	Sum of atomic polarizabilities
MolRef	Ghose and Crippen molar refractivity
Rotlbonds	Number of rotatable bonds
Hbond acceptor	Number of hydrogen bond acceptors
Hbond donor	Number of hydrogen bond donors
MW	Molecular weight
PMI-mag	Principal moment of inertia

For the QSPR analysis based GA, the appearance reflects a feature's utility in many different combinations towards building high-score models. Thus, observation of the descriptor use as it changes is an effective way to watch the evolution of the elite population, to estimate when the population has converged and to judge the relative utility of different descriptors quickly. The convergence was met after about 1,700 operations.

In these descriptors, the frequencies of PNSA-2 and $A \log P$ in the models are the highest, significantly higher than the other descriptors. The descriptor PNSA-2 is used in all models, and the descriptor $A \log P$ is used in about 99% of the models. Due to the high appearance frequencies of $A \log P$ and PNSA-2, it seems that these two descriptors play strong roles in the proposed QSPR models. $A \log P$ represents the octanol/water partition coefficient. The positive value of $A \log P$ for this term indicates that a high octanol/water partition coefficient contributes to strong brain–blood permeation ability. In fact, for almost all linear correlation models predicting log BB described in the literature, the octanol/water partition coefficient was very important. Previously, several attempts were reported to predict BB transport with the octanol/water partition coefficient. [3] Our data indicate that, although this descriptor was validated to be very important, it does not show effective linear correlation with log BB for the 59 structurally diverse compounds in the training set (see Eq. 1).

$$\log BB = -0.61 + 0.24 \times A \log P (r = 0.23F = 3.11) \quad (1)$$

The descriptor PNSA-2 was found in all QSPR models. This descriptor demonstrates some linear correlation with log BB (see Eq. 2).

$$\log BB = -0.44 + 0.0014 \times PNSA - 2 (r = 0.68F = 40.03) \quad (2)$$

It is interesting to find that the partitioning of compounds between the blood and brain compartments can be described effectively by a combination of $A \log P$ and PNSA-2 (see Eq. 3).

$$\log BB = -0.31 + 0.44 \times A \log P + 0.0018 \times PNSA - 2 (r = 0.80q = 0.77F = 47.34) \quad (3)$$

PNSA-2 is defined as partial negative SASA multiplied by the negative charge. In previous work, the total polar surface area (TPSA) was used in the QSPR studies of blood–brain partitioning. [34] It seems that the descriptor PNSA-2 is quite similar to the descriptors PNSA-1 and TPSA. However, in contrast to PNSA-1 and TPSA, partial charges are considered by PNSA-2. For the descriptors PNSA-1 and TPSA, whether partial charges are significantly different for different atoms or not, the molecular surface areas for different atoms are weighted equally. But in fact, partial charges are directly connected with the electrostatic interactions. We believe that the descriptor PNSA-2, which considers the partial charges, is better than PNSA-1 or TPSA for describing the electrostatic interactions between the drugs and the lipid bilayer during BBB penetration.

Beside these two parameters, the frequency of the descriptor RadOfGyration seems relatively high, which is distributed among about 23% the models. But the remainder of the descriptors shown in Fig. 1, and those not shown, are not well distinguished from each other by their use. This suggests that the information represented

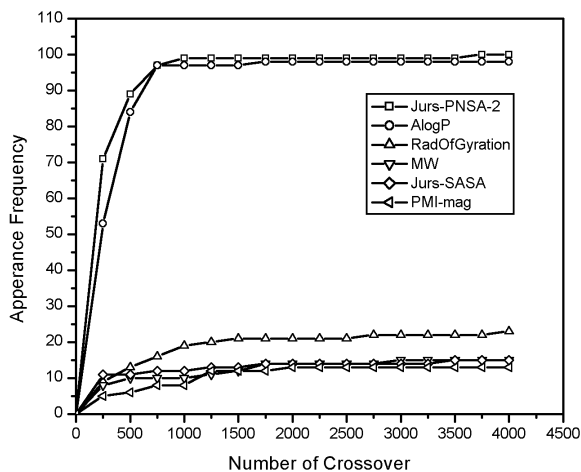


Fig. 1 Change in descriptor use for the elite population as the evolution proceeds with four descriptors

by some descriptors can be mostly embodied in the descriptors $A \log P$, PNSA-2 and RadOfGyration. For example, previous work has shown that descriptors relating to hydrogen bonding were necessary. [11, 12] However, our calculations show that the descriptor of Hbond acceptor occurs in 11 equations, and the descriptor of Hbond do-

nor can only be found in three. In fact, Hbond acceptor shows a close linear correlation with PNSA-2 ($r=-0.89$), which means that the Hbond acceptor descriptor may be replaced by the descriptor PNSA-2. The correlation between Hbond acceptor and PNSA-2 is not difficult to interpret. We know that an atom that is a hydrogen bond acceptor is often an oxygen or nitrogen atom with relatively large negative charges. If a molecule bears many hydrogen bond acceptors, it should also have a large negative PNSA-2 value. This suggests that the information in the data set is often duplicated over many of the descriptors. We cannot say that the hydrogen capability of a molecule is not important in BBB permeation. We can only say that the hydrogen bonding property of a molecule may be absorbed in the other descriptors, which may be described better by another or several other descriptors.

QSPR models from GA optimization

After calculations, 100 models with three descriptors and 100 models with four descriptors were obtained. The top 16 QSPR models with the highest fitness scores are shown in Table 5. For a QSPR model, the quality of a model cannot be estimated simply by its multiple correlation regression coefficient. Thus, the quality of the models, as indi-

Table 5 The top 16 QSPR models generated using the training set

5	$\log BB = -0.62 - 0.11 \times \text{Rotlbonds} + 0.32 \times A \log P + 0.0024 \times \text{Jurs-PNSA-2} + 0.35 \times \text{RadOfGyration}$ n=59 r=0.870 q=0.837 F=42.135 SD=0.408
6	$\log BB = -0.56 + 0.35 \times A \log P + 0.0033 \times \text{Jurs-PNSA-2} + 0.60 \times \text{RadOfGyration} - 0.0037 \times \text{Jurs-PPSA-1}$ n=59 r=0.865 q=0.837 F=40.407 SD=0.415
7	$\log BB = -0.67 + 0.33 \times A \log P + 0.0035 \times \text{Jurs-PNSA-2} + 0.46 \times \text{RadOfGyration} - 0.0020 \times \text{Jurs-DPSA-1}$ n=59 r=0.865 q=0.836 F=40.304 SD=0.415
8	$\log BB = -0.68 + 0.28 \times A \log P + 0.0025 \times \text{Jurs-PNSA-2} - 0.10 \times \text{Rotlbonds} + 0.0032 \times \text{Jurs-SASA}$ n=59 r=0.863 q=0.823 F=39.433 SD=0.419
9	$\log BB = -1.03 + 0.24 \times \text{RadOfGyration} + 0.31 \times A \log P + 0.0030 \times \text{Jurs-PNSA-2} + 0.55 \times \text{Density}$ n=59 r=0.861 q=0.689 F=38.797 SD=0.421
10	$\log BB = -0.32 + 0.27 \times A \log P + 0.0024 \times \text{Jurs-PNSA-2} + 0.21 \times \text{RadOfGyration} + 0.016 \times \text{Foct}$ n=59 r=0.860 q=0.827 F=38.514 SD=0.422
11	$\log BB = -0.38 - 0.11 \times \text{Rotlbonds} + 0.27 \times A \log P + 0.0024 \times \text{Jurs-PNSA-2} + 0.0042 \times \text{Area}$ n=59 r=0.860 q=0.817 F=38.437 SD=0.423
12	$\log BB = -0.29 + 0.28 \times A \log P + 0.0047 \times \text{Vm} + 0.0024 \times \text{Jurs-PNSA-2} - 0.087 \times \text{Rotlbonds}$ n=59 r=0.858 q=0.809 F=37.849 SD=0.432
13	$\log BB = -0.087 + 0.018 \times \text{Foct} + 0.012 \times \text{MolRef} + 0.23 \times A \log P + 0.0025 \times \text{Jurs-PNSA-2}$ n=59 r=0.858 q=0.819 F=37.841 SD=0.425
14	$\log BB = -0.72 + 0.0026 \times \text{Jurs-PNSA-2} + 0.38 \times \text{RadOfGyration} - 0.00042 \times \text{Jurs-PPSA-2} + 0.32 \times A \log P$ n=59 r=0.736 q=0.823 F=37.685 SD=0.426
15	$\log BB = -0.71 + 0.0027 \times \text{Jurs-PNSA-2} + 0.25 \times \text{RadOfGyration} + 0.869 \times \text{Jurs-RPCG} + 0.31 \times A \log P$ n=59 r=0.858 q=0.826 F=37.667 SD=0.426
16	$\log BB = -0.56 + 0.0032 \times \text{Jurs-PNSA-2} + 0.18 \times \text{RadOfGyration} + 0.13 \times \text{Hbond acceptor} + 0.35 \times A \log P$ n=59 r=0.858 q=0.825 F=37.619 SD=0.426
17	$\log BB = -0.22 - 0.088 \times \text{Rotlbonds} + 0.27 \times A \log P + 0.0016 \times \text{Jurs-PNSA-2} + 0.0018 \times \text{Jurs-TASA}$ n=59 r=0.858 q=0.812 F=37.341 SD=0.427
18	$\log BB = -0.18 + 0.25 \times A \log P + 0.0028 \times \text{Jurs-PNSA-2} + 0.0040 \times \text{MW} + 0.015 \times \text{Foct}$ n=59 r=0.857 q=0.814 F=37.224 SD=0.428
19	$\log BB = -0.83 + 0.0015 \times \text{Jurs-TASA} + 0.0024 \times \text{Jurs-PNSA-2} + 0.60 \times \text{Density} + 0.26 \times A \log P$ n=59 r=0.856 q=0.820 F=36.984 SD=0.429
20	$\log BB = -0.69 - 0.0031 \times \text{Jurs-PPSA-1} + 0.0034 \times \text{Jurs-PNSA-2} + 0.0053 \times \text{Jurs-SASA} + 0.29 \times A \log P$ n=59 r=0.856 q=0.820 F=36.889 SD=0.429

Table 6 Experimental and predicted log BB values for test set 1

No.	log BB (Exp.)	log BB (Pred.)		
		1 ^a	2 ^b	3 ^c
B1 ^d	-1.30	0.42	0.27	0.33
B2 ^d	-1.40	0.12	0.03	0.09
B3	-0.43	-0.83	-0.48	-0.42
B4	0.25	-0.28	-0.04	0.05
B5	-0.30	-0.03	0.00	0.019
B6	-0.06	0.12	0.13	0.16
B7	-0.42	-0.12	-0.09	-0.07
B8	-0.16	0.51	0.34	0.51
B9	0.00	0.40	0.22	0.37
B10	-0.34	0.00	-0.11	0.04
B11	-0.30	-0.48	-0.68	-0.58
B12	-1.34	-1.28	-1.34	-1.30
B13	-1.82	-2.09	-1.94	-1.95
B14	0.89	1.07	0.83	1.04
SSE ^e		1.50	0.84	1.16
rmse		0.35	0.26	0.31

cated by SD , F , q , was tested statistically. In the equations in Table 5, n is the number of compounds used in the fit, SD is the standard error of mean, and F is the overall F -statistics. We find that, compared with the two-descriptor model (Eq. 3), the quantities of the four-descriptor models listed in Table 5 are improved significantly. It can thus be concluded that the log BB values can be chiefly explained by the two descriptors A log P and PNSA-2, but the addition of the other descriptors can also introduce obvious effects to the QSPR models obtained.

For the analysis of multiple linear correlation, the data material must be reduced to fewer and less correlated variables. Cross-correlated descriptors would prevent the QSAR model disclosing the actual relationship between the biological activity and the descriptors. Thus, it is important to check the uniqueness of the descriptors in a model. In the current work, the independence of the descriptors was checked by calculating the correlation matrix of the parameters in the final models in Table 5. In the regression equation, $r^2=0.8$ was considered to be the threshold that would be required for variables to substitute for each other. [35] After correlation analysis, some descriptors were detected to be highly correlated. These cannot be considered independent. Finally, Eqs. (6), (8), (11), (12), (13), (14), (18) and (20) in Table 5 were eliminated from the viewpoint of statistics.

The selection of the definitive model was carried out on the basis of prediction for the compounds comprising the validation test. Firstly, the compounds in the test set were predicted using the QSPR model with the best fitness score (Eq. 5 in Table 5). The first validation set includes the BB ratio of eight H1-receptor histamine antagonist/agonist (**B1–B8**) and six miscellaneous CNS agents (**B9–B14**). The observed and predicted log BB are shown in Table 6. As may be seen from Table 6, the predictions for compounds **B3** to **B14** in the test set are very good, whereas the log BB values for compounds **B1** and **B2** are strongly overestimated by this model. It

Table 7 Experimental and predicted log BB values for test set 2

No.	Exp.	Pred.		
		1 ^a	2 ^b	3 ^c
C1	-0.29	-0.51	-0.55	-0.50
C2	-0.06	-0.31	-0.40	-0.33
C3	-0.10	0.02	-0.06	-0.00
C4	-1.23	-1.38	-0.98	-1.01
C5	-0.31	-0.53	-0.41	-0.43
C6	-0.18	0.39	0.35	0.40
C7	0.11	-0.40	-0.35	-0.25
C8	0.55	-0.14	-0.21	0.01
C9	0.12	-0.83	-0.68	-0.66
C10	-1.42	-1.27	-0.97	-0.92
C11	0.04	0.68	0.37	0.56
C12	0.50	0.03	0.16	0.27
C13	-1.26	-1.66	-1.41	-1.42
C14	0.61	-0.57	-0.60	-0.52
C15	0.39	-0.21	-0.22	-0.04
C16	1.3	0.66	0.54	0.73
C17	1.2	1.17	0.98	1.01
C18	0.36	0.31	0.05	0.22
C19	-0.70	-1.73	-1.25	-1.29
C20	1.23	0.38	0.35	0.37
C21	1.06	0.18	0.17	0.07
C22	0.24	0.02	-0.04	0.14
C23	-0.52	-0.70	-0.27	-0.32
SSE		7.78	6.98	6.11
rmse		0.58	0.55	0.52

should be noted, however, that these two compounds were also overestimated in their BB ratio by the linear free-energy equations reported by Abraham et al., [36] the PLS model reported by Luco [16] and the three-descriptor linear model proposed by Feher et al., [34] and may therefore be considered as outliers. For the remaining 12 compounds, the RMS prediction error using Eq. (5) in Table 5 is 0.29, which is at the level of the experimental error in the BB determinations.

The second validation test included 23 structurally diverse compounds, and their log BB values were taken from Salminen et al. [10] The observed and predicted log BB data are shown in Table 7. Inspection of these results shows that the linear model performs reasonable well, and only two compounds **C14** and **C19** were strongly underestimated, and may be considered as outliers. The RMSE value calculated for the 23 validation compounds is 0.58, which means that the RMSE value for the reduced data set (excluding **C14** and **C19**) is 0.50.

We know that the experimental data of log BB are often highly heterogeneous and of fairly poor quality. It is questionable whether a model with a significantly higher number of parameters would not simply fit the model to errors in the experimental data. The data set used in the current work is too limited. Moreover, in Table 5, the fitness scores among these models really do not show large differences. It is thus very difficult for us to give a decisive conclusion as to which model is the best based on the fitness score and the predictions for the limited test compounds. We think that selection of a single model and the discarding of the remaining models may not be the

Table 8 Experimental and computed log BB values

Compound	Exp. log BB	1 ^a		2 ^b	
		Calc. log BB	Residue	Calc. Log BB	Residue
1	-1.42	-1.11	0.31	-0.78	0.64
2	-0.04	-0.32	-0.28	-0.38	-0.34
3	-2.00	-1.25	0.75	-1.15	0.85
4	-1.30	-1.00	0.30	-0.83	0.47
5	-1.06	-1.43	-0.37	-1.24	-0.18
6	0.11	-0.18	-0.29	-0.21	-0.32
7	0.49	0.03	-0.52	0.17	-0.32
8	0.83	0.69	-0.14	0.57	-0.26
9	-1.23	-1.60	-0.37	-1.18	0.05
10	-0.83	-1.95	-1.12	-1.66	-0.83
11	-1.17	0.10	1.27	-0.05	1.12
12	-2.15	-1.14	1.01	-0.81	1.34
13	-0.67	-0.95	-0.28	-0.70	-0.03
14	-0.66	-0.85	-0.19	-0.62	-0.04
15	-0.12	-0.69	-0.57	-0.52	-0.40
16	-0.18	-0.13	0.05	-0.24	-0.06
17	-1.15	-0.52	0.63	-0.61	0.54
18	-1.57	-1.22	0.35	-1.14	0.43
19	-1.54	-1.43	0.11	-1.31	0.23
20	-1.12	-1.29	-0.17	-1.02	0.10
21	-0.73	-1.11	-0.38	-0.93	-0.20
22	-0.27	-0.37	-0.10	-0.43	-0.16
23	-0.28	-0.25	0.03	-0.28	0.00
24	-0.46	-0.33	0.13	-0.29	0.17
25	-0.24	-0.05	0.19	-0.04	0.20
26	-0.02	0.06	0.08	0.14	0.16
27	0.69	0.33	-0.36	0.37	-0.32
28	0.44	-0.09	-0.53	-0.05	-0.49
29	0.14	0.20	0.06	0.04	-0.10
30	0.22	0.12	-0.10	-0.03	-0.25
31	-1.88	-1.20	0.68	-1.11	0.77
32	-0.50	-0.86	-0.36	-0.61	-0.11
33	-0.22	-0.02	0.20	0.13	0.35
34	-1.10	-0.34	0.76	-0.24	0.86
35	-0.31	-0.44	-0.13	-0.36	-0.05
36	-1.70	-1.18	0.52	-0.95	0.75
Butanone	-0.08	0.11	0.19	0.08	0.16
Benzene	0.37	0.61	0.24	0.45	0.08
3-Methylpentane	1.01	0.76	-0.35	0.62	-0.39
3-Methylhexane	0.90	0.86	-0.04	0.73	-0.17
2-Propanol	-0.15	-0.09	0.06	-0.08	0.07
2-Methylpropanol	-0.17	0.00	0.17	0.03	0.20
2-Methylpentane	0.97	0.78	-0.19	0.63	-0.34
2,2-Dimethylbutane	1.04	0.83	-0.21	0.64	-0.40
1,1,1-Trifluoro-2-chloroethane	0.08	-0.18	-0.26	-0.14	-0.22
1,1,1-Trichloroethane	0.40	0.12	-0.28	0.14	-0.26
Diethyl ether	0.00	0.09	0.09	0.09	0.09
Ethanol	-0.16	-0.27	-0.11	-0.20	-0.04
Fluoroxene	0.13	-0.10	-0.23	-0.05	-0.18
Halothane	0.35	-0.31	-0.66	-0.21	-0.56
Heptane	0.81	0.90	0.09	0.78	0.03
Hexane	0.80	0.79	-0.01	0.66	-0.14
Methane	0.04	0.03	-0.01	0.02	-0.02
Methylcyclopentane	0.93	0.78	-0.15	0.55	-0.38
Pentane	0.76	0.68	-0.08	0.54	-0.22
Propanol	-0.16	-0.17	-0.01	-0.09	0.07
Propanone	-0.15	-0.07	0.08	-0.08	0.07
Toluene	0.37	0.82	0.45	0.59	0.22
Trichloroethene	0.34	0.22	-0.12	0.25	-0.09

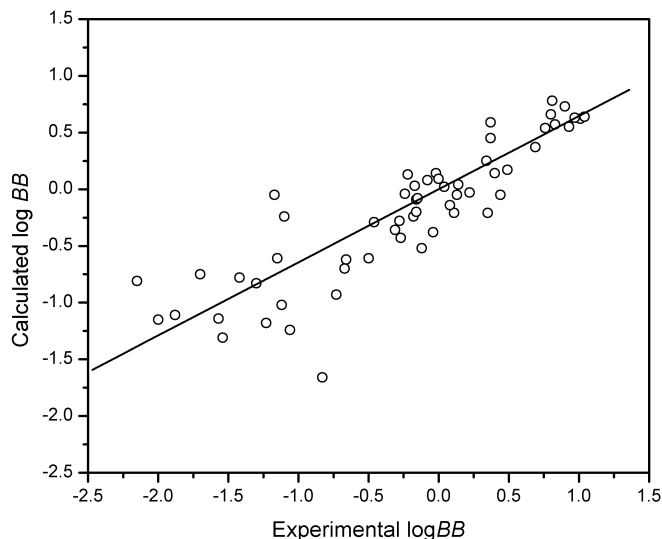


Fig. 2 Comparison of experimental log BB with calculated log BB for the compounds in the training set using multiple models

most advantageous course. It has been proposed that the outputs of the multiple models can be averaged to gain the most reliable results. [22] We therefore averaged the results with the best eight four-descriptor models (bold equations in Table 5). The top model predicted the training set with $r=0.87$; by averaging its output with the eight best models, the correlation coefficient climbs to 0.88. The calculated log BB values for the compounds in the training set averaged from the output with the eight best models are shown in Table 8, and the relationship between the observed and calculated log BB values is illustrated in Fig. 2. The test compounds were also predicted by averaging the predicted values with the eight best models (see Tables 6 and 7 and Fig. 3). The averaged predicted log BB values for **B1** and **B2** are also highly overestimated, but the RMSE value for the predicted log BB values of **B3** to **B14** using multiple models is decreased to 0.26. The prediction for compounds **C1** to **C23** by averaging the multiple models is also significantly better than that by using a single model. The RMSE value for **C1** to **C23** using multiple models is 0.55. Using a single model, compound **C14** is highly overestimated, which may be determined as an outlier, but the predicted value for **C19** is quite reasonable. It is interesting to find that compound **C14** was also strongly underestimated using the PLS model proposed by Luco [16]

Here, it should be noted that the QSPR models after GA optimizations do not consider any molecules in the data set as outliers. Obviously, removing outliers from the training set will greatly improve the apparent performance of the model. If we define the compounds whose error of prediction is larger than 1.0, for the values using the top eight models, two outliers (compounds **11** and **12**) are identified. After removing compounds **11** and **12** from the correlation, the models in Table 5 could be improved significantly. For example, for Eq. 5 in Table 5, after re-

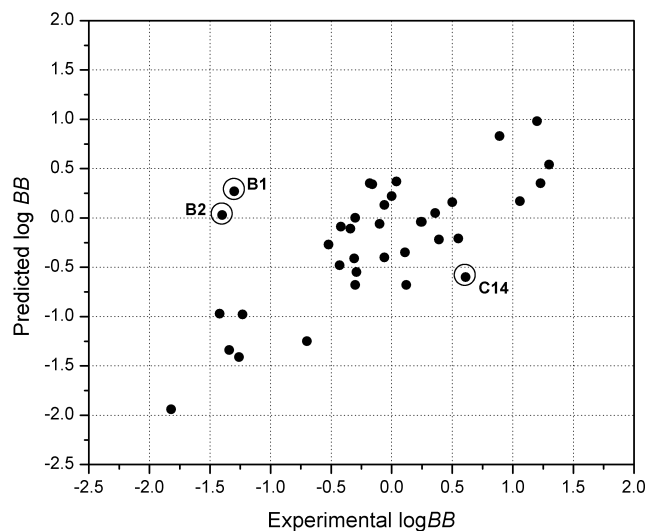


Fig. 3 Comparison of experimental log BB with calculated log BB for the compounds in the test set using multiple models. (Three compounds with predicted error larger than 1.0 are marked with circles)

moving compounds **11** and **12**, r of the model increases from 0.87 to 0.93, while q increases from 0.84 to 0.89.

$$\begin{aligned} \log BB = & -0.29 + 0.43 \times A \log P + 0.0017 \\ & \times \text{PNSA} - 2 - 0.0017 \times \text{Rotlbonds} \\ & + 0.0051 \times \text{RadOfGyration} \end{aligned} \quad (4)$$

$(r = 0.93q = 0.89F = 58.49SD = 0.26)$

In the earlier literature, researchers tend to eliminate the compounds with poor prediction from the training set in order to gain better statistical significance. In the work of Lombardo et al., two outliers (**3** and **12**) were removed from the 57-compound set. [13] In the work of Luco, compounds **2**, **10** and **12** were identified as outliers and removed from the training set. [16] In the current work, the predicted errors for compounds **3** and **10** using multiple models are not larger than 1.0. If we consider these two outliers, the resulting four-descriptor model (Eq. 4) is better than any other models reported using the similar training sets. However, we believe that any model using specific descriptors based on a limited data set may bear some predicted tendency for some types of compounds. Thus, 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 either test set. The top eight models in Table 5 were rebuilt using the training set after removing compounds **11** and **12**, and the test compounds were predicted again using these eight new models. Tables 6 and 7 show the predicted values after averaging the results using the new models. It can be seen that, for the compounds **B3** to **B14**, the RMSE value is 0.31, which is worse than the previous prediction. In contrast, for the compounds **C1** to **C23**, the RMSE value is 0.52, which is a little better than the

Table 9 Comparison of several statistical models

	Current work		Reference (Luco [16])	Reference (Feher et al. [34])
	1 ^a	2 ^b		
Training set				
<i>n</i>	59	57	58	61
<i>r</i>	0.87	0.93	0.92	0.85
<i>q</i>	0.84	0.89	0.87	0.83
rmse	0.41	0.35	0.40	0.42
Test set 1 (<i>n</i> =12)				
<i>r</i>	0.94	0.94	0.92	0.97
rmse	0.26	0.31	0.54	0.24
Test set 2				
<i>n</i>	23	23	25	25
<i>r</i>	0.80	0.80	0.79	0.76
rmse	0.55	0.52	0.79	0.52

previous predictions. However, overall, the new model after removing outliers does not have a large effect on predictions for the test compounds.

It is of interest to evaluate other models suggested in the literature using a similar data set. The properties of our four-descriptor model are compared to those of the three-PCA-component PLS model proposed by Luco [16] and the three-descriptor model proposed by Feher et al. [34] (see Table 9). Not considering outliers, our four-descriptor model has a similar performance for the training set. However, after removing the outliers, our four-descriptor model shows better performance than the other two. The predictions for the test set 1 using multiple four-descriptor models developed in the current work are only slightly worse than the results reported by Luco, [16] while much better than the results provided by Feher et al.'s model. [34] For the test set 2, our predictions are similar to those reported by Luco, but also much better than the results reported by Feher et al. However, it should be noted that the model provided by Luco needs numerous physicochemical descriptors and PLS analysis, which cannot be easily and rapidly calculated for an arbitrary compound. The four-descriptor models provided here are simple and can be applied in high-throughput screening easily.

Conclusions

This study has shown that QSPR analysis based on a GA can afford us a group of linear models with high statistical significance. The change of descriptor use as the evolution proceeds demonstrates that the octane/water partition coefficient and partial negative SASA multiplied by the negative charge are crucial to brain-blood barrier permeability. Moreover, the predictions using multiple QSPR models from GA optimization gave quite good results in spite of the diversity of structures, which is better than the predictions using the best single model. The linear models developed in the current work are easily cal-

culated and suitable for the rapid prediction of the log BB values in a high-throughput fashion.

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

References

- Guyton AC, Hall JE (1991) Textbook of medical physiology, 8th edn. WB Saunders, Philadelphia, Pa. p 683
- Tamai I, Tsuji A (1996) Adv Drug Delivery Rev 19:401–424
- Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffiths R, Jones M, Rana KK, Saunders D, Smith LR, Sore NE, Wilks TJ (1998) J Med Chem 31:656–671
- Eddy EP, Maleef BE, Hart TK, Smith PL (1997) Adv Drug Deliv Rev 23:185–198
- Reichel A, Begley DJ (1998) Pharm Res 15:1270–1274
- Gupta SP (1989) Chem Rev 89:1765–1800
- Hansch C, Bjorkroth JP, Leo AJ (1987) Pharm Sci 76:683–687
- Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffiths R, Jones M, Rana KK, Saunders D, Smith IR, Sore NE, Wilks TJ (1988) J Med Chem 31:656–671
- Kaliskan R, Markuszewski M (1996) Int J Pharm 145:9–16
- Salminen T, Pulli A, Taskinen JJ (1997) Pharm Biomed Anal 15:469–477
- Abraham MH, Chadha HS, Mitchell RC (1994) J Pharm Sci 83:1257–1268
- Waterbeemd H, Camenisch G, Folkers G, Chretien JR, Raevsky OA (1998) J Drug Targeting 6:151–165
- Lombardo F, Blake JF, Curatolo WJ (1996) J Med Chem 39:4750–4755
- Norinder U, Sjoberg P, Osterberg TJ (1998) Pharm Sci 87:952–959
- Sjoberg P (1997) In van der Waterbeemd H, Testa B, Folkers G (eds) Computer-assisted lead finding and optimization. Verlage Helvetica Chimica Acta, Basel, p 83
- Luco JM (1999) J Chem Inf Comput Sci 39:396–404
- Cruciani G, Crivori P, Carrupt PA, Testa B (2000) J Mol Struct (THEO) 503:17–30
- Kaznessis YN, Snow ME, Blankley CJ (2001) J Comput Aid Mol Des 15:697–708
- Hou TJ, Wang JM, Li YY, Xu XJ (1998) Chin Chem Lett 9:651–654
- Hou TJ, Wang JM, Xu XJ (1999) Chemometr Intell Lab 45:303–310
- Hou TJ, Wang JM, Liao N, Xu XJ (1999) J Chem Inf Comp Sci 39:775–781
- Rogers D, Hopfinger AJ (1994) J Chem Inf Comp Sci 34:854–866
- Kubinyi H (1994) Quant Struct–Act Rel 13:285–294
- Kubinyi H (1994) Quant Struct–Act Rel 13:393–401
- Greenwood J (1995) New Concepts of a Blood–Brain Barrier. Plenum, New York, p 251
- Calder JAD, Ganellin R (1994) Drug Des Discov 11:259–268
- SYBYL 6.5 User Guide (1999). Tripos, St. Louis, Mo. USA
- Halgren TA (1996) J Comput Chem 17:490–519
- Ghose AK, Crippen GM (1986) J Comput Chem 7:565–577
- Yalkowsky SH, Sinkula AA, Valvani YC (1980) Physical chemistry properties of drugs. Marcel Dekker, New York
- Ghose AK, Crippen GM (1987) J Chem Inf Comput Sci 27:21–35
- Stanton DT, Jurs PC (1990) Anal Chem 62:2323–2329
- Rappé AK, Goddard WA (1991) J Phys Chem 95:3358–3364
- Feher M, Sorial E, Schmidt JM (2000) Int J Pharm 201:239–247
- Martin YC (1978) Quantitative drug design: a critical introduction. Marcel Dekker, New York, p 194
- Abraham MH, Chada HS, Mitchell RC (1995) Drug Des Discov 13:123–131