

Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis

Dahua Pan, Manisha Iyer, Jianzhong Liu, Yi Li, and Anton J. Hopfinger*

Laboratory of Molecular Modeling and Design (M/C 781), College of Pharmacy,
The University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231

Received June 15, 2004

A new method, using a combination of 4D-molecular similarity measures and cluster analysis to construct optimum QSAR models, is applied to a data set of 150 chemically diverse compounds to build optimum blood-brain barrier (BBB) penetration models. The complete data set is divided into subsets based on 4D-molecular similarity measures using cluster analysis. The compounds in each cluster subset are further divided into a training set and a test set. Predictive QSAR models are constructed for each cluster subset using the corresponding training sets. These QSAR models best predict test set compounds which are assigned to the same cluster subset, based on the 4D-molecular similarity measures, from which the models are derived. The results suggest that the specific properties governing blood-brain barrier permeability may vary across chemically diverse compounds. Partitioning compounds into chemically similar classes is essential to constructing predictive blood-brain barrier penetration models embedding the corresponding key physio-chemical properties of a given chemical class.

INTRODUCTION

The ability of a central nervous system (CNS) drug to penetrate the blood-brain barrier (BBB) is a fundamental requirement for the drug to be active. Conversely, peripherally acting drugs must possess negligible BBB penetration in order to minimize undesired CNS-related side-effects. The uptake of a compound into the brain is a complex process due to the unique properties of the endothelial cells of the brain capillaries that act as both a physical barrier and a biochemical interface.^{1,2} Most drugs can penetrate the BBB by passive diffusion, and whole molecule properties such as charge, molecular volume, and partition coefficient can significantly influence transport across the BBB.³ In addition, the capability of a compound to form hydrogen bonds is another important factor governing drug BBB permeability.

The most common approaches to predict drug BBB penetration focus on modeling structurally diverse data sets by dealing with certain intramolecular properties of the molecules including lipophilicity indices, solvation, hydrogen bond parameters, and limited three-dimensional structure features.^{4,5} *Membrane-interaction QSAR (MI-QSAR) analysis* has been developed in order to better predict the transport behavior of structurally diverse organic compounds by interacting them with the phospholipids-rich regions of biological membranes.^{6,7} In MI-QSAR analysis, the major assumption is made that the phospholipids-rich regions of a cellular membrane constitute a “receptor” to penetrating solutes and permit incorporation of structural and chemical diversity into a training set. A set of membrane-solute intermolecular properties are determined and added to the “usual” set of intramolecular solute QSAR descriptors to enlarge the QSAR descriptor pool and to better provide the

information needed to incorporate chemical and structural diversity into the QSAR analysis.

There has recently been a surge in computational efforts to estimate ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of drug-like compounds, and modeling BBB penetration is one major ADMET endpoint focus.^{7–9} Efforts have particularly been made to construct a general BBB QSAR model from a large and structurally diverse training set.⁹ Unfortunately, a common mechanism of action, one of the most basic assumptions of QSAR analysis, may not be satisfied for these types of ADMET data sets. In other words, all compounds in a structurally diverse data set may not interact with the blood-brain “membrane” in the same manner. The final QSAR model may only reflect an “average” picture of drug penetration over several mechanisms. Factors governing BBB penetration for “minority” compounds may be neglected in the “average mechanism” of the general QSAR model.

Molecular similarity and cluster analyses seem to be the necessary tools to address the multiple mechanism problem. The notion of similarity is based on recognition of common features of a set of objects and consequent categorization into different subgroups. The ability of pattern recognition and classification based on likeness is a powerful tool for information handling. The concept of molecular similarity in the drug development field is, however, a complex construct which may only be meaningful with respect to a given application. Different notions of molecular similarity have been suggested and used based on molecular formula, molecular graphs, molecular skeletons, atom types and positions, conformations, van der Waals surfaces, and/or molecular fields.¹⁰ However, few molecular similarity methods include 3D conformation-dependent properties. *4D molecular similarity (4D-MS)* developed by Hopfinger and co-workers includes the thermodynamic distribution of

* Corresponding author phone: (312)996-4816; fax: (312)413-3479; e-mail: hopfingr@uic.edu.

conformer states available to a molecule in constructing a set of similarity/diversity descriptors.¹¹

Cluster analysis is the process in which groups are detected within a set of objects where members in a group are “similar” to one another with respect to some attribute while “dissimilar” to members in other groups. Cluster analysis was first used in taxonomy for species classification but has been gradually adopted by other disciplines where information about grouping is necessary. Before a cluster method can be applied, measures representing attributes of objects have to be derived as the basis for comparison among different objects. Criteria to determine what measures should be used to reflect the essential properties of an object mainly depend on the purpose of the study. Thus, selection of attributes is a subjective process. Furthermore, choosing an appropriate clustering set of criteria (method) for the task at hand also relies on the attributes of the data set and the goals one hopes to achieve. Essentially, there is no one “correct” set of clusters for a particular set of objects.¹²

Cluster analysis has been adopted by scientists working in computer-aided drug design owing to the “principle” that molecules having similar structural and physiochemical properties are likely to behave similarly in a biological system. Thus, the rational classification of a large compound library may be useful for identifying drug-likeness hit compounds.

In this study, both general two-dimensional and three-dimensional MI-QSAR descriptors are included in the BBB penetration analysis. Moreover, a novel method is reported to deal with the multiple mechanisms problem by grouping similar compounds into clusters and, consequently, constructing QSAR models for the members of each cluster.

A new method, using a combination of 4D-molecular similarity measures and cluster analysis to construct optimum QSAR models, is applied to a data set of 150 chemically diverse compounds to build optimum BBB penetration models. The complete data set is divided into subsets based on 4D-molecular similarity measures using cluster analysis. The compounds in each cluster subset are further divided into a training set and a test set. Predictive QSAR models are constructed for each cluster subset using the corresponding training sets. These QSAR models best predict test set compounds which are assigned to the same cluster subset, based on the 4D-molecular similarity measures, from which the models are derived. The results suggest that the specific properties governing blood-brain barrier permeability may vary across chemically diverse compounds. Partitioning compounds into chemically similar classes is essential to constructing predictive blood-brain barrier penetration models embedding the corresponding key physiochemical properties of a given chemical class.

MATERIALS

The 150 compound data set was constructed from data in the literature^{9,13} and is reported in Table 1. The dependent variable used in this study is the logarithm of the BBB partition coefficient

$$\text{Log BB} = \log (C_{\text{brain}}/C_{\text{blood}}) \quad (1)$$

where C_{brain} is the concentration of the test compound in the brain, and C_{blood} is the concentration in the blood.

The smallest compound in the data set is methane (B001) and the largest is indinavir (B148, $\text{C}_{36}\text{H}_{47}\text{N}_4\text{O}_4$). The obvious structural dissimilarity of compounds in this data set implies that these compounds may interact with the BBB membrane differently.

METHODS

(1) Construction of the Main Distance-Dependent Matrix (MDDM) and Corresponding Eigenvalue Calculation Using 4D-Molecular Similarity. The theory and corresponding methodology for constructing the MDDMs and computing corresponding eigenvalues for each matrix, using 4D molecular similarity (MS), are presented in detail by Duca and Hopfinger.¹⁴ The first step of a 4D-MS analysis is to generate the conformation ensemble profile (CEP) of each member of a set of molecules of interest using molecular dynamic simulation (MDS). The CEP collects the contributions from different conformational states of a molecule explicitly considering that molecular shape and flexibility can be important factors in specifying molecular properties.

After the CEPs are obtained, the MDDM for each pair of interaction pharmacophore elements (IPE) of each molecule is constructed. Up to 36 MDDMs, of the same term, or cross-term, IPE pair types, can be constructed for a molecule based on the current eight types of IPEs. The eight IPEs are listed in Table 2.

A characteristic feature of 4D-MS is that this approach can calculate both absolute and relative molecular similarity. Absolute molecular similarity only employs structural information inherent to the 3D conformations of a molecule, while the relative molecular similarity measures are alignment-dependent. This study investigates, in part, how particular alignment-independent molecular similarity features, represented by the same type IPE pairs, affect the clustering of a large diverse data set. Thus, only the theory and procedure for calculating absolute molecular similarity measures for the same type IPE pairs is presented. The MDDM element, $f_{(dij)}$, is defined as

$$f_{(dij)} = e^{(-v \langle dij \rangle)} \quad (2)$$

where v is a “universal constant” which maximizes the range in the molecular similarity measure and $\langle dij \rangle$ is the average distance between atom i and j of a given molecule over all conformations sampled in the CEP. This parameter is calculated as

$$\langle dij \rangle = \sum_k d_{ij}(k)p(k) \quad (3)$$

In eq 3, $p(k)$ is the thermodynamic probability of conformer k , d_{ij} is the corresponding distance between atoms i and j of the same type IPE pair for this particular conformer. Only distances associated with atoms of the appropriate IPE type are included in a particular MDDM.

When the IPE types are the same, as is the case in this application, the MDDM is an upper/lower triangular matrix. Principle component analysis (PCA) is the data reduction tool to transform a MDDM into a set of eigenvalues which are normalized and sorted in numerically descending order. Normalization is performed for the unscaled eigenvalues,

Table 1. Complete BBB Data Set

Table 1 (Continued)

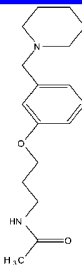
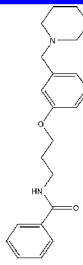
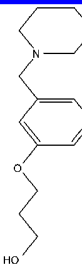
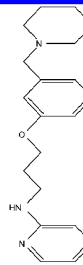
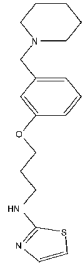
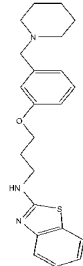
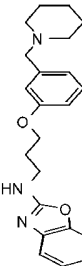
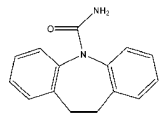
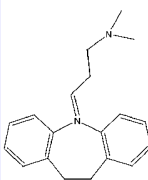
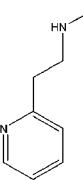
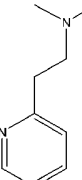
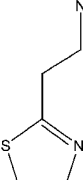
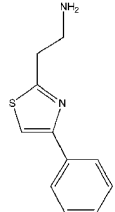
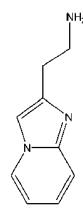
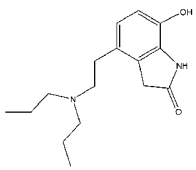
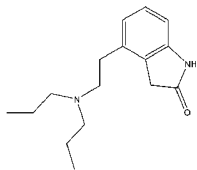
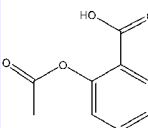
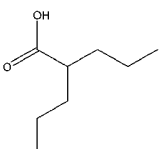
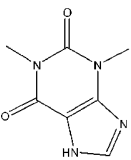
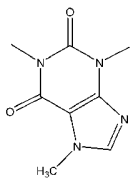
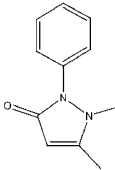
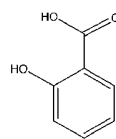
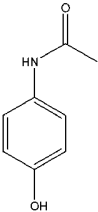
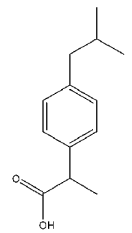
 B060	 B061	 B062	 B063
 B064	 B065	 B066	 B067
 B068	 B069	 B070	 B071
 B072	 B073	 B074	 B075
 B076	 B077	 B078	 B079
 B080	 B081	 B082	 B083

Table 1 (Continued)

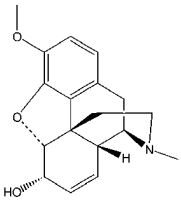
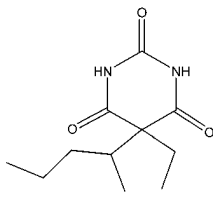
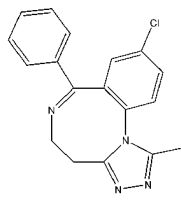
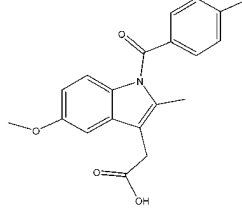
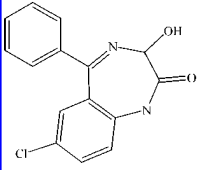
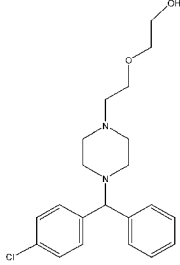
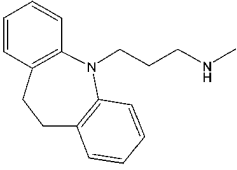
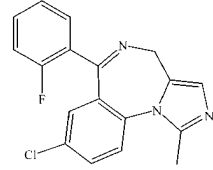
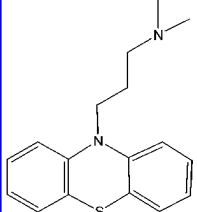
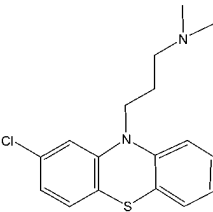
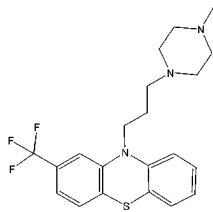
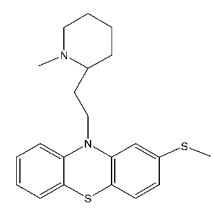
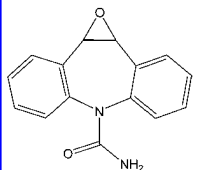
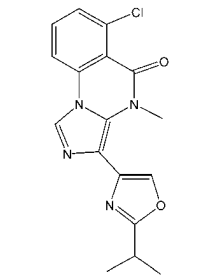
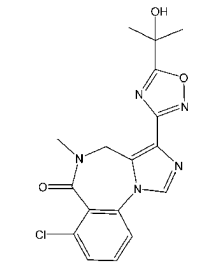
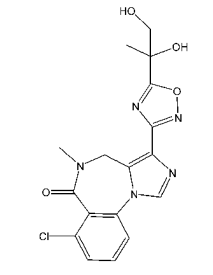
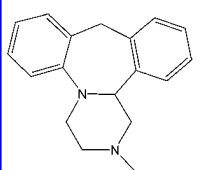
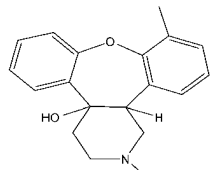
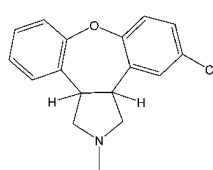
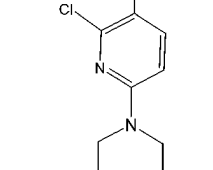
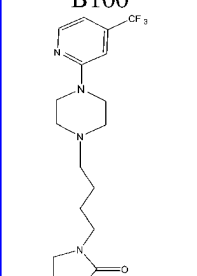
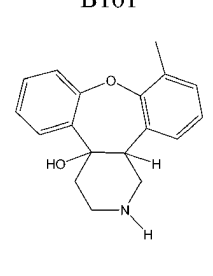
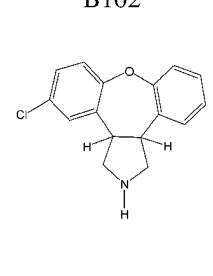
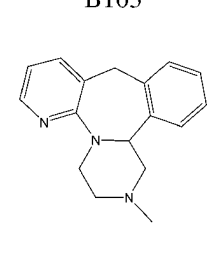
			
B084	B085	B086	B087
			
B088	B089	B090	B091
			
B092	B093	B094	B095
			
B096	B097	B098	B099
			
B100	B101	B102	B103
			
B104	B105	B106	B107

Table 1 (Continued)

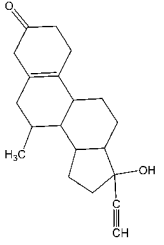
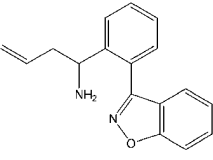
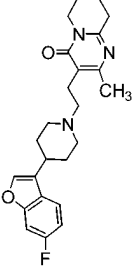
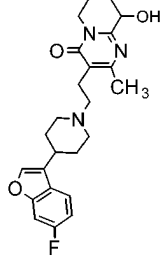
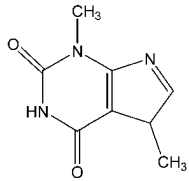
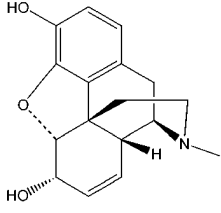
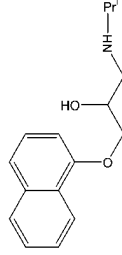
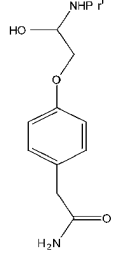
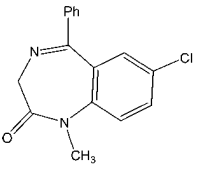
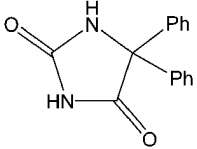
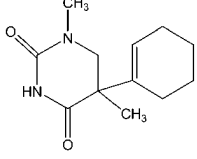
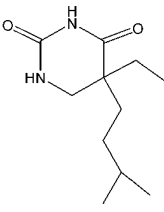
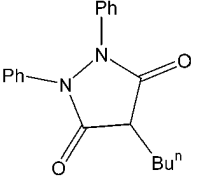
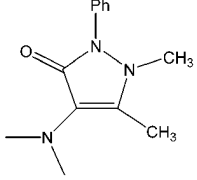
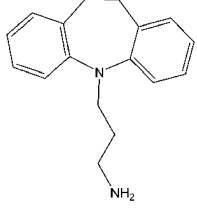
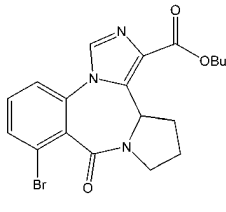
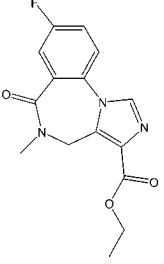
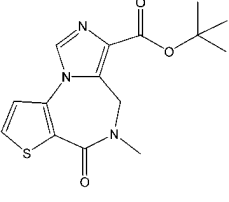
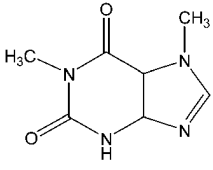
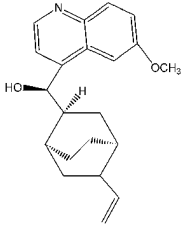
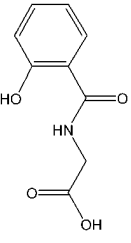
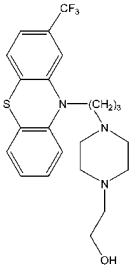
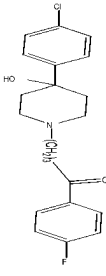
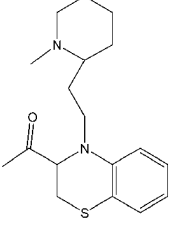
			
B108	B109	B110	B111
			
B112	B113	B114	B115
			
B116	B117	B118	B119
			
B120	B121	B122	B123
			
B124	B125	B126	B127
			
B128	B129	B130	B131

Table 1 (Continued)

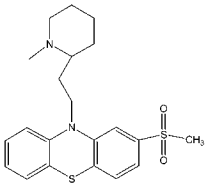
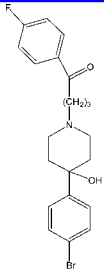
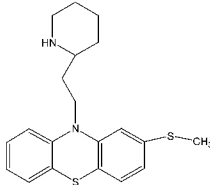
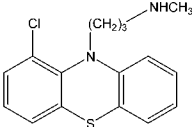
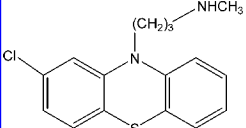
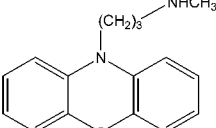
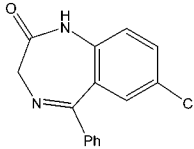
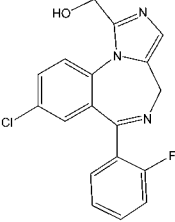
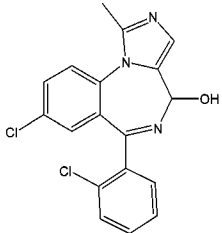
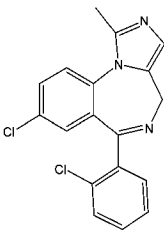
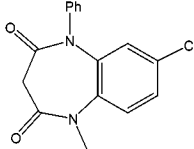
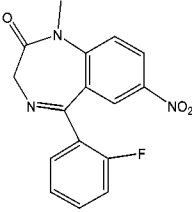
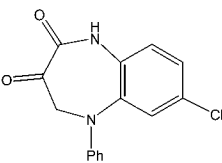
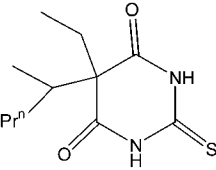
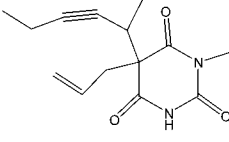
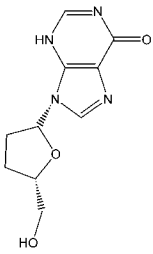
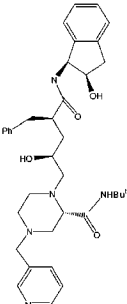
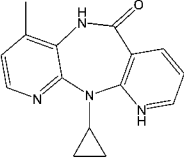
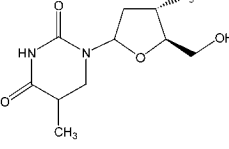
			
B132	B133	B134	B135
			
B136	B137	B138	B139
			
B140	B141	B142	B143
			
B144	B145	B146	B147
			
B148	B149	B150	

Table 1 (Continued)

compound	name	logBB	compound	name	logBB
B001	methane	0.04	B041	5 ^a	-1.06
B002	pentane	0.76	B042	clonidine (6) ^a	0.11
B003	hexane	0.8	B043	mepyramine (7) ^a	0.49
B004	2-methylpentane	0.97	B044	imipramine (8) ^a	1.06
B005	3-methylpentane	1.01	B045	rantidine (9) ^a	-1.23
B006	2,2-dimethylbutane	1.04	B046	tiotidine (10) ^a	-0.82
B007	heptane	0.81	B047	11 ^a	-1.17
B008	3-methylhexane	0.9	B048	12 ^a	-2.15
B009	cyclopropane	0	B049	13 ^a	-0.67
B010	cyclohexane	0.92	B050	14 ^a	-0.66
B011	methylcyclopentane	0.93	B051	15 ^a	-0.12
B012	dichloromethane	-0.11	B052	16 ^a	-0.18
B013	trichloromethane	0.29	B053	17 ^a	-1.15
B014	1,1,1-trichloroethane	0.4	B054	18 ^a	-1.57
B015	trichloroethylene	0.34	B055	19 ^a	-1.54
B016	1,1,1-trifluoro-2-chloroethane	0.08	B056	20 ^a	-1.12
B017	halothane	0.35	B057	21 ^a	-0.73
B018	teflurane	0.27	B058	22 ^a	-0.27
B019	diethyl ether	0	B059	23 ^a	-0.28
B020	divinyl ether	0.11	B060	24 ^a	-0.46
B021	methoxyflurane	0.25	B061	25 ^a	-0.24
B022	isoflurane	0.42	B062	26 ^a	-0.02
B023	enflurane	0.24	B063	27 ^a	0.69
B024	fluroxene	0.13	B064	28 ^a	0.44
B025	propanone	-0.15	B065	29 ^a	0.14
B026	butanone	-0.08	B066	30 ^a	0.22
B027	ethanol	-0.16	B067	31 ^a	0
B028	propan-1-ol	-0.16	B068	36 ^a	0.89
B029	propan-2-ol	-0.15	B069	Y-G14	-0.3
B030	2-methylpropan-1-ol	-0.17	B070	Y-G15	-0.06
B031	benzene	0.37	B071	Y-G16	-0.42
B032	toluene	0.37	B072	Y-G19	-1.3
B033	ethylbenzene	0.2	B073	Y-G20	-1.4
B034	<i>p</i> -xylene	0.31	B074	SKF 89124	-0.43
B035	<i>m</i> -xylene	0.29	B075	SKF 101468	0.25
B036	<i>o</i> -xylene	0.37	B076	acetylsalicylic acid	-0.5
B037	cimetidine (1) ^a	-1.42	B077	valproic acid	-0.22
B038	2 ^a	-0.04	B078	theophylline	-0.29
B039	3 ^a	-2	B079	caffeine	-0.05
B040	4 ^a	-1.3	B080	antipyrene	-0.1
B081	salicylic acid	-1.1	B116	diazepam	0.52
B082	acetaminophen	-0.31	B117	phenytoin	-0.04
B083	ibuprofen	-0.18	B118	hexobarbital	0.1
B084	codeine	0.55	B119	amobarbital	0.04
B085	pentobarbital	0.12	B120	phenylbutazone	-0.52
B086	alprazolam	0.04	B121	aminopyrine	0
B087	indomethacin	-1.26	B122	desmethydesipramine	1.06
B088	oxazepam	0.61	B123	bretazenil	-0.09
B089	hydroxyzine	0.39	B124	flumanezil	-0.29
B090	desipramine	1.2	B125	RO19-4603	-0.25
B091	midazolam	0.36	B126	paraxanthine	0.06
B092	promazine	1.23	B127	quinidine	-0.46
B093	chloropromazine	1.06	B128	salicylic acid	-0.44
B094	trifluoperazine	1.44	B129	fluphenazine	1.51
B095	thioridazine	0.24	B130	haloperidol	1.34
B096	32 ^a	-0.34	B131	mesoridazine	-0.36
B097	33 ^a	-0.3	B132	sulforidazine	0.18
B098	34 ^a	-1.34	B133	bromperidol	1.38
B099	35 ^a	-1.82	B134	northioridazine	0.75
B100	mianserin	0.99	B135	nor-1-chlorpromazine	1.37
B101	Org4428	0.82	B136	nor-2-chlorpromazine	0.97
B102	Org5222	1.03	B137	desmonomethylpromazine	0.59
B103	Org16962	1.64	B138	desmethyldiazepam	0.5
B104	Org13011	0.16	B139	1-hydroxymidazolam	-0.07
B105	Org32104	0.52	B140	4-hydroxymidazolam	-0.3
B106	Org30526	0.39	B141	triazolam	0.74
B107	mirtazapine	0.53	B142	clobazam	0.35
B108	tibolone	0.4	B143	flunitrazepam	0.06
B109	Org34167	0	B144	desmethylocobazam	0.36
B110	risperidone	-0.02	B145	thiopental	-0.14
B111	risperidone-9-OH	-0.67	B146	methohexital	-0.06
B112	theobromine	-0.28	B147	didanosine	-1.3
B113	morphine	-0.16	B148	indinavir	-0.74
B114	propranolol	0.64	B149	nevirapine	0
B115	atenolol	-1.42	B150	zidovudine	-0.72

^a No chemical names are given to these compounds. The presented numbers are taken from literature reports.^{9,13}

Table 2. Interaction Pharmacophore Elements, IPEs, Used in the 4D-MS Analyses

definition	symbol	IPE code
all atoms in the molecule	any	0
nonpolar atoms	np	1
polar atoms with positive charge	p+	2
polar atoms with negative charge	p-	3
hydrogen bond acceptor atoms	hba	4
hydrogen bond donor atoms	hbd	5
aromatic atoms	a	6
non-hydrogen atoms	hs	7

$\epsilon'_i(\alpha)_{u,u}$, relative to the rank of the MDDM matrix, $\text{rank}(\alpha)$, as given by eq 4.

$$\{\epsilon'_i(\alpha)\}_{u,u} = \{\epsilon_i(\alpha)\}_{u,u} / \text{rank}(\alpha) \quad (4)$$

This set of normalized eigenvalues, $\{\epsilon_i(\alpha)\}_{u,u}$, is used as a fingerprint for molecule α with respect to IPE type u . Molecular similarity and/or dissimilarity are thus estimated based on the set of fingerprints across the IPEs.

The composite difference between corresponding eigenvalues of molecules α and β , $D_{\alpha\beta}$, is defined as the molecular dissimilarity and is given as

$$D_{\alpha\beta} = \sum_i |\epsilon(\alpha)_i - \epsilon(\beta)_i| \quad (5)$$

The molecular similarity, $S_{\alpha\beta}$, is defined by eq 6

$$S_{\alpha\beta} = (1 - D_{\alpha\beta}) (1 - \varphi) \quad (6)$$

where

$$\varphi = |\text{rank}(\alpha) - \text{rank}(\beta)| / (\text{rank}(\alpha) + \text{rank}(\beta)) \quad (7)$$

The rank of a MDDM matrix is essentially the number of atoms of a specific IPE type present in the corresponding molecule. Therefore, the φ term in eq 6 serves to reincorporate molecular size information. All eigenvalues used in computing molecular similarity are normalized values. Thus, the molecular similarity measure is a number between 1 and 0, where a value closer to 1 indicates higher commonality between a pair of compounds with respect to the particular IPE being considered.

Absolute molecular similarity MDDMs for all pairs of atoms with the same IPE types are computed for each molecule in the data set. Eigenvalues of the MDDM matrices are then employed as the 4D-fingerprints to represent a molecule with respect to a particular IPE type. Thus, for a molecule α , eight MDDMs are constructed and eight sets of eigenvalues (4D-fingerprints) computed that correspond, individually, to the eight IPE types listed in Table 2. A threshold cutoff value for eigenvalues is applied, and those normalized eigenvalues below the threshold cutoff value are disregarded. For this study, the threshold was set at 0.002.

(2) Subset Dividing. The set of eigenvalues, the 4D-fingerprint, for a given IPE type, is used as the numerical representation of a molecule for that IPE type. Eight sets of 4D-fingerprints, one corresponding to each IPE type, are used for the total numerical descriptor representation of a molecule. Partition around the medoid (PAM) is employed to partition the complete data set into subsets based on the eigenvalues of a particular IPE type.¹⁵ PAM is performed using the S-plus statistical package.¹⁶ PAM first selects k

centrally located objects called medoids. Medoids are actual objects in the data set, and they are representative of their respective clusters. After the medoids are determined, each object is assigned to the nearest medoid such that the total dissimilarity of all objects to their nearest medoid is minimal. PAM provides a novel graphical display, the silhouette plot, which allows the rational selection of the appropriate number of clusters. A silhouette value, $s(i)$, which largely measures the certainty that an object is correctly assigned to a cluster, is calculated for each object. Given a PAM analysis has been performed, and data points grouped in different cluster in order to compute the $s(i)$ of object i in cluster A , the average dissimilarity of i to all other objects j in A , $a(i)$, is first calculated

$$a(i) = \sum_j d(i, A_j) / \text{number of objects in } A \quad (8)$$

where $d(i, A_j)$ is the distance of each object j in A from i .

The lowest corresponding dissimilarity of i to any other clusters, $b(i)$, is calculated as well. If i is most similar to objects in cluster B , rather than A , then

$$b(i) = \sum_j d(i, B_j) / \text{number of objects in } B \quad (9)$$

The silhouette value for object i is thus defined as

$$s(i) = (b(i) - a(i)) / \max \{a(i), b(i)\} \quad (10)$$

where $\max \{a(i), b(i)\}$ is the maximum value between $a(i)$ and $b(i)$.

The $s(i)$ values are between 0 and 1. A value closer to 1 indicates a higher probability of appropriate object assignment. The average silhouette value of all objects indicates the overall "goodness" of clustering across the data set. The appropriate clustering is achieved by selecting the clustering solution with the highest average silhouette value.

(4) Defining the Training and Test Sets for the Subset Clusters. To investigate how clustering a large structurally diverse data set into smaller groupings may influence QSAR model construction, the compounds in a subset are further divided into training and test sets. The underlying idea for this type of focused data separation is that both the training and test sets should occupy the same "biological" and chemical spaces. The complete data set is first categorized into four biological groups with respect to their permeability measure: high, medium, fair, and low BBB penetration. In addition to the biological activity space representation, a classic clustering method, *k-mean*,¹⁷ is employed, using S-plus,¹⁶ to further group compounds in each biological category into four subgroups, with respect to molecular structure, based on the 4D-fingerprints. The use of *K-means* clustering is to realize the division of chemical space. The selection of the clustering method and/or number of groups to be created can vary depending on the data set. *K-means* clustering is a well-developed method and is implemented in many statistical software packages, providing easy access, and, thus, is chosen as an appropriate method for subset division. Finally, one-third of the compounds in each of the subgroups created from *k-means* clustering of each subset are selected to form a test set, and the remaining compounds compose the corresponding training set.

(5) Descriptor Selection. A MI-QSAR analysis^{6,7} is employed to generate a set of geometric and thermodynamic properties of the interaction of a solute with the phospholipids in a cellular membrane. These descriptors are then associated with the event of solute BBB penetration. The methodology has been discussed in previously published papers.^{6,7} A brief description of the method, as it is applied in this study, is given to provide a general understanding of the rationale for this approach to ADMET property prediction.

The application of MI-QSAR analysis on the data set follows the procedure reported by Iyer et al.⁷ All the (solute) molecules in the BBB penetration data set were built using HyperChem 5.01. Partial atomic charges of the molecules were computed using the semiempirical AM1 method in Hyperchem.¹⁸ Dimyristoylphosphatidylcholine (DMPC) was selected as the model phospholipid in this study.

An ensemble of 25 DMPC molecules (5×5×1) was constructed as the model membrane monolayer. The center DMPC molecule was removed, and a solute was inserted in the space created by the “missing” DMPC molecule. This solute is placed at three trial positions, upper, middle, and bottom, of the membrane monolayer, with respect to polar surface of the membrane model, to represent membrane transport during the BBB penetration event.

MDS of the solute molecules in the free-state is first carried out. An initial MDS on the model membrane, without a solute molecule, is performed to permit structural relaxation over the model membrane monolayer. In addition, three corresponding MDS for solute-membrane complexes, with respect to the different trial solute positions, upper, middle, and bottom, are conducted. All MDS are performed using the MOLSIM package.¹⁹ The simulation temperature is held constant at 311 K. An MDS of 10 ps sampling time with time-step intervals of 0.001 ps is performed for a total sampling of 10 000 conformations of the solute, the DMPC monolayer, and each solute-membrane complex. Periodic boundary conditions ($a = 40 \text{ \AA}$, $b = 40 \text{ \AA}$, $c = 80 \text{ \AA}$) are applied to the MDS of the model membrane and the solute-membrane complexes.

Most of the “classic” intramolecular descriptors of the free-state solute molecules were also included using the Cerius 2,²⁰ Daylight,²¹ and MI-QSAR software packages.^{22,23} The descriptors used in the trial descriptor pool are given in Table 3.

(6) Model Construction and Evaluation for the Training Set Compounds Clustered into Subsets. QSAR models for both the complete and partitioned BBB penetration data sets are built and optimized using GFA,²⁴ which is a multidimensional optimization procedure based on genetic algorithms. Only linear representations of each of the descriptor values were included in the model construction process. Both r^2 and the cross-validated correlation coefficient, q^2 , are employed to judge the quality of the resultant QSAR models.

(7) Test Set Prediction. The optimized QSAR model obtained for a given subset is used to predict the BBB penetration values of the corresponding test set compounds and for compounds in the other subsets. The idea is to verify that test compounds in the same subset, from which the models are derived, are better predicted than those test compounds from other subsets. This procedure is a way to validate the partitioning process. To evaluate the predictive

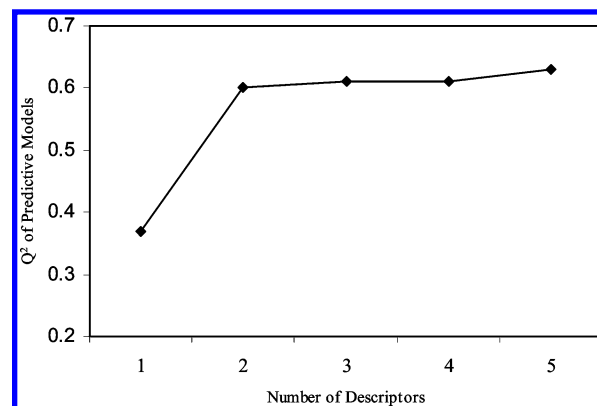


Figure 1. A plot of q^2 vs the number of descriptor terms in the corresponding BBB QSAR models for the complete data set.

ability of the QSAR models, the r^2 of test set prediction, termed $r^2_{\text{prediction}}$, is calculated as

$$r^2_{\text{prediction}} = 1 - \frac{\sum (y_{i(\text{pred})} - y_{i(\text{obs})})^2}{\sum (y_{i(\text{obs})} - k)^2} \quad (11)$$

where k is the average of the observed test set values and $y_{i(\text{pred})}$ and $y_{i(\text{obs})}$ are the predicted and observed activity values for the i^{th} test set compound, respectively. A r^2 of prediction of greater than zero indicates that the model predicts better than if the average activity value over the test set compounds was chosen for each of the predicted activity values. Conversely, a value less than zero indicates that using the average predictive activity value of the test set compounds gives a better prediction for any of the test set compounds than that computed from the model being used to make the predictions.

RESULTS

The optimized QSAR model for the complete 150 compound data set has two descriptor terms and an r^2 of 0.79 and a q^2 of 0.60. The model is given by eq 12.

$$\text{LogBB} = 0.064 + 0.20\text{ClogP} - 0.01\text{TPSA}$$

$$N = 150, r^2 = 0.69, q^2 = 0.60 \quad (12)$$

This general model, which is similar to the model reported by Clark for a smaller data set,⁴ captures two well-known factors that influence drug BBB permeability, namely the lipophilicity [ClogP] of a solute and its polar surface area [TPSA]. However, the moderate q^2 value suggests that this model might only perform marginally in predicting the logBB of compounds outside the training set. Efforts were also made to obtain models with additional descriptor terms in order to enhance predictivity. The q^2 values of the optimized QSAR models plotted against number of descriptors are presented in Figure 1. As is shown in Figure 1, q^2 converges for models with more than two descriptor terms. Given the small number of descriptors needed to optimize q^2 and the moderate optimized value of q^2 , it is of interest to further probe the complete data set in order to better model BBB penetration.

To explore the possibility that compounds in the data set may interact with the BBB membrane differently, 4D molecular similarity, which contains three-dimensional structural information, was employed to generate 4D-fingerprints as the numerical representations of a molecule. 4D-fingerprints of 10 compounds, based on hydrogen bond donor

Table 3. Descriptors Used in BBB Model Construction

Part A. The General Intramolecular Solute Descriptors		
functional families	an example of the functional family	description of the example
fragment constants descriptors	HA	hydrogen bond acceptor
conformational descriptors	energy	energy of the current selected conformation
electronic descriptors	charge	sum of partial charges
topological descriptors ^b	Kappa-M	Kier & Hall topological descriptors
molecular shape analysis (MSA)	DIFFV	difference volume
spatial descriptors	area	molecular surface area
structural descriptors	MW	molecular weight
thermodynamic descriptors	Hf	heat of formation
	TPSA ^{a, b}	topological polar surface area
Part B. The Intermolecular (Membrane-Solute) Interaction Descriptors Used in BBB Model Construction		
the membrane-solute descriptors –symbols	description of the membrane-solute descriptors	
<F(total)>	average total free energy of interaction of the solute and membrane	
<E(total)>	average total interaction energy of the solute and membrane	
E _{INTER} (total)	interaction energy between the solute and the membrane at the total intermolecular system minimum potential energy	
E _{XY} (Z)	Z = 1,4-nonbonded, general van der Waal, electrostatic, hydrogen bonding, torsion and combinations thereof energies at the total intermolecular system minimum potential energy	
	X, Y can be the solute, S, and/or membrane, M	
ΔE _{XY} (Z)	change in the Z = 1,4-nonbonded, general van der Waal, electrostatic, hydrogen bonding, torsion and combinations thereof energies due to the uptake of the solute to the total intermolecular system minimum potential energy.	
	X, Y can be the solute, S, and/or membrane, M	
E _{TT} (Z)	Z = 1,4-nonbonded, general van der Waal, electrostatic, hydrogen bonding, torsion and combinations thereof energies of the total [solute and membrane model] intermolecular minimum potential energy	
ΔE _{TT} (Z) ^{&}	change in the Z = 1,4-nonbonded, general van der Waal, electrostatic, hydrogen bonding and combinations thereof of the total [solute and membrane model] intermolecular minimum potential energy	
ΔS	change in entropy of the membrane due to the uptake of the solute	
S	absolute entropy of the solute-membrane system	
Δρ	change in density of the model membrane due to the permeating solute	
<d>	average depth of the solute molecule from the membrane surface	
Part C. None MI-QSAR Intermolecular Dissolution and Solvation Descriptors of The Solute		
dissolution and solvation – solute descriptors –symbols	description of the dissolution/solvation –solute descriptors	
F(H2O)	the aqueous solvation free energy	
F(OCT)	the 1-octanol solvation free energy	
Log(P) ^{&}	the 1-octanol/water partition coefficient	
E(coh) ^{&}	the cohesive packing energy of the solute molecules	
T _M	the hypothetical crystal-melt transition temperature of the solute	
T _G	the hypothetical glass transition temperature of the solute	

^a No functional family is identified for TPSA. ^b Descriptor or member(s) of the functional family is identified to be important in BBB-QSAR models.

similarity, IPE type 4, are shown in Table 4 as an example of 4D-fingerprinting. The threshold cutoff value for the eigenvalues [fingerprints] is set at 0.002. Eigenvalues below the cutoff value are disregarded. Thus, only four eigenvalues are listed for the compounds in Table 4.

PAM is subsequently performed on the complete data set for subset division based on the 4D-fingerprints. The average silhouette values of partitioning the 150 compound data set into different numbers of subsets with respect to the 4D-fingerprints of different IPE types are listed in Table 5.

The silhouette value is a unique criterion of PAM to determine what is the optimum number of subgroups present

in a particular data set for a particular set of grouping features (the 4D-fingerprints). The highest average silhouette value indicates the best way to partition a given data set. The results presented in Table 5 indicate the optimum number of subgroups for the 150 compounds data set is two for most IPE types. The only subgrouping outlier is that based on the aromatic IPE type 6 (aromatic atoms), where nearly identical silhouette values are found for divisions into two, three, and four subsets. However, the silhouette values only suggest an appropriate statistical means to separate a large data set. The legitimacy of data division may still rely on identifying the best predictive models. To further explore the best IPE

Table 4. Four Largest 4D-Fingerprints, Eigenvalues, of Ten Compounds for IPE Type 4

compound	eigenvalue 1	eigenvalue 2	eigenvalue 3	eigenvalue 4
1	0	0	0	0
10	0	0	0	0
20	0	0	0	0
30	0	0	0	0
40	0.704	0.150	0.145	0
50	0.632	0.165	0.112	0.088
60	0.477	0.302	0.220	0
70	0.662	0.337	0	0
80	0.563	0.297	0.138	0
90	0.640	0.359	0	0
100	0.650	0.349	0	0

type for subgroup dividing, QSAR models are constructed for each of the two subsets determined for each IPE types. To be able to best compare the new QSAR models built from the subsets of various sizes to the general model given by eq 12, optimum two-term QSAR models for each subset were constructed, and corresponding r^2 and q^2 values of each QSAR model are listed in Table 6.

Different subsets, with correspondingly dissimilar QSAR models, are realized for different IPE types. The goal is to seek the most appropriate IPE type in order to be able to further analyze how subset dividing may improve the resultant QSAR models. An initial focus is given to IPE types 2, 5, and 6 that yield the optimum subsets as judged by the silhouette values reported in Table 5. Unfortunately, the QSAR models built for the subsets of IPE types 2, 5, and 6 are either less significant, or only slightly better, than the general QSAR model, eq 12, built for the complete data set. However, for IPE types 3 and 4, clustering is reasonably significant in term of the silhouette values. Moreover, additional information is seemingly extracted from the first subsets that is reflected by the large increase in the r^2 and q^2 of the corresponding QSAR models, as compared to eq 12. For the first subset of IPE type 3, r^2 increases from 0.69 to 0.88, and q^2 from 0.60 to 0.85. For IPE type 4, r^2 increases to 0.85 and q^2 to 0.83 for the first subset. Thus, priority is given to investigating separation of the data with regard to IPE types 3 and 4.

The best models for subsets based on IPE type 3 similarity are presented below:

Subset 1:

$$\text{Log BB} = 0.24 + 0.27\text{ClogP} + 0.00073\text{Echg}$$

$$N = 26, r^2 = 0.88, q^2 = 0.85 \quad (13)$$

Subset 2:

$$\text{Log BB} = 0.14 - 0.013\text{TPSA} + 0.20\text{ClogP}$$

$$N = 124, r^2 = 0.60, q^2 = 0.57 \quad (14)$$

The best models for subsets based on IPE type 4 similarity are also presented:

Subset 1:

$$\text{Log BB} = 0.10 + 0.26\text{ClogP} + 0.00052\text{Echg}$$

$$N = 37, r^2 = 0.85, q^2 = 0.83 \quad (15)$$

Subset 2:

$$\text{Log BB} = 0.28 - 0.014\text{TPSA} + 0.18\text{ClogP}$$

$$N = 113, r^2 = 0.64, q^2 = 0.61 \quad (16)$$

IPE type 3, a polar atom with partial negative charge, and IPE type 4, a hydrogen bond acceptor, represent similar functional atom types. Thus, it is reasonable that subset division based on these two IPE types would yield similar results. The QSAR models given by eqs 13 and 14 are essentially the same as those of eqs 15 and 16, in terms of both descriptors and statistical quality of the models.

Subset 1 of IPE type 3 contains 26 compounds, namely compound B001–B011, B019, B020, B025–B036, and B068. Subset 1 of IPE type 4 includes compounds B001–B036 and B068. Thus, IPE type 3 captures most of the same compounds found in using IPE type 4. Since IPE type 4 includes more compounds in subset 1 than IPE type 3, and IPE type 4 is a more specific physiochemical description of an atom property, IPE type 4 is selected as the optimum IPE type for further subset investigations.

Table 5. Average Silhouette Values of a PAM Analysis Indicating the “Quality” of Subset Division with Respect to IPE Types

no. of subsets	IPE types							
	0 av silhouette value	1 av silhouette value	2 av silhouette value	3 av silhouette value	4 av silhouette value	5 av silhouette value	6 av silhouette value	7 av silhouette value
2	0.61	0.51	0.91	0.76	0.78	0.93	0.85	0.59
3	0.42	0.47	0.79	0.52	0.62	0.88	0.88	0.43
4	0.4	0.36	0.79	0.4	0.62	0.87	0.89	0.40

Table 6. Quality of the Two-Term Models, Based on r^2 and q^2 , for Two Subsets of the Complete Data Set with Respect to IPE Types.

IPE types subsets	0		1		2		3	
	subset 1	subset 2	subset 1	subset 2	subset 1	subset 2	subset 1	subset 2
no. of compds	46	104	62	88	93	57	26	124
r^2	0.73	0.66	0.66	0.65	0.44	0.51	0.88	0.60
q^2	0.68	0.61	0.61	0.62	0.40	0.48	0.85	0.57
IPE types subsets	4		5		6		7	
	subset 1	subset 2	subset 1	subset 2	subset 1	subset 2	subset 1	subset 2
no. of compds	37	113	104	46	52	98	102	48
r^2	0.85	0.60	0.47	0.45	0.73	0.64	0.48	0.54
q^2	0.83	0.57	0.44	0.36	0.71	0.62	0.45	0.50

Table 7. Training and Test Sets for Subset 1 Based on IPE Type 4 4D-Fingerprints^a

compd	logBB	compd	logBB
BB_030	-0.17	BB_035	0.29
BB_027	-0.16	BB_034	0.31
BB_028	-0.16	BB_015	0.34
BB_025	-0.15	BB_017	0.35
BB_029	-0.15	BB_031	0.37
BB_012	-0.11	BB_032	0.37
BB_026	-0.08	BB_036	0.37
BB_009	0	BB_014	0.4
BB_019	0	BB_022	0.42
BB_001	0.04	BB_002	0.76
BB_016	0.08	BB_003	0.8
BB_020	0.11	BB_007	0.81
BB_024	0.13	BB_068	0.89
BB_033	0.2	BB_008	0.9
BB_023	0.24	BB_010	0.92
BB_021	0.25	BB_011	0.93
BB_018	0.27	BB_004	0.97
BB_013	0.29	BB_005	1.01
		BB_006	1.04

^a The test set compounds are in bold. Every third compound is selected as a test set compound.

A common feature of compounds in subset 1 based on IPE type 4 similarity is that all these compounds have less than two heteroatoms, excluding halogens. Thus, using IPE type 4 (hydrogen bond acceptor), subset 1 compounds essentially have identical 4D-fingerprints and cannot be further divided into chemical diverse groups. Every third compound, as ranked by logBB value, in subset 1 is selected as a member of the corresponding test set. The remaining two-third compounds form the training set. The training and test sets for subset 1 are given in Table 7. The test set compounds are listed in bold.

A significant QSAR model, virtually identical to eq 15, is obtained for the training set in subset 1 and presented as below

$$\text{Log BB} = 0.26 + 0.26\text{ClogP} + 0.00077\text{Echg}$$

$$N = 24, r^2 = 0.85, q^2 = 0.83 \quad (17)$$

where Echg is electrostatic interaction energy for the solute-membrane complex.

Using the same strategy employed for subset 1, compounds in subset 2 are first divided into four logBB classes of high ($1 < \log\text{BB}$), medium ($0 < \log\text{BB} < 1$), fair ($-1 < \log\text{BB} < 0$), and low logBB ($\log\text{BB} < -1$). Categorizing compounds in subset two into four activity subgroups provides a convenient ranking of compound BBB permeability. Concentrations are higher in the brain than in blood for compounds in the high and medium activity groups, while it is lower for compounds in the fair and low activity groups. Subset 2 contains 113 structurally diverse compounds, making it feasible to group structurally similar compounds into common classes. The *k-means* method of partitioning is employed to further divide compounds in each logBB class into four sub-subgroups based on 4D-fingerprints. Thus, sixteen sub-subgroups are formed. One-third of the compounds in each of these sub-subgroups are chosen to construct the corresponding test set for each subgroup. If one "sub-subgroup" contains less than three compounds, one compound is included in the test set. The combination of

Table 8. Training and Test Sets for Subset 2 Based on IPE Type 4 Similarity and logBB Values

logBB space representation							
high		medium		fair		low	
compd	log BB	compd	log BB	compd	log BB	compd	log BB
Chemical Space Representation Based on 4D-Fingerprints							
Group 1							
BB133	1.38	BB067	0	BB120	-0.52	BB039	-2
		BB109	0	BB076	-0.5	BB099	-1.82
		BB121	0	BB128	-0.44	BB054	-1.57
		BB086	0.04	BB096	-0.34	BB055	-1.54
		BB119	0.04	BB140	-0.3	BB098	-1.34
		BB126	0.06	BB078	-0.29	BB040	-1.3
		BB118	0.1	BB112	-0.28	BB147	-1.3
		BB042	0.11	BB052	-0.18	BB045	-1.23
		BB085	0.12	BB145	-0.14	BB047	-1.17
		BB075	0.25	BB139	-0.07	BB056	-1.12
		BB142	0.35	BB146	-0.06	BB041	-1.06
		BB091	0.36	BB079	-0.05		
		BB144	0.36	BB038	-0.04		
		BB138	0.5	BB117	-0.04		
		BB105	0.52				
		BB116	0.52				
		BB107	0.53				
		BB088	0.61				
		BB114	0.64				
		BB141	0.74				
		BB101	0.82				
Group 2							
BB090	1.2	BB095	0.24	BB060	-0.46	BB072	-1.3
BB130	1.34	BB106	0.39	BB127	-0.46	BB073	-1.4
		BB137	0.59	BB074	-0.43		
		BB134	0.75	BB131	-0.36		
		BB136	0.97	BB082	-0.31		
		BB100	0.99	BB061	-0.24		
				BB113	-0.16		
				BB062	-0.02		
Group 3							
BB094	1.44	BB108	0.4	BB071	-0.42	BB048	-2.15
BB129	1.51			BB069	-0.3	BB037	-1.42
				BB077	-0.22	BB115	-1.42
				BB083	-0.18	BB087	-1.26
				BB080	-0.1	BB053	-1.15
				BB070	-0.06		
Group 4							
BB102	1.03	BB043	0.49	BB046	-0.82	BB081	-1.1
BB044	1.06	BB063	0.69	BB148	-0.74		
BB093	1.06	BB064	0.44	BB057	-0.73		
BB122	1.06	BB065	0.14	BB150	-0.72		
BB092	1.23	BB066	0.22	BB049	-0.67		
BB135	1.37	BB084	0.55	BB111	-0.67		
BB103	1.64	BB089	0.39	BB050	-0.66		
		BB104	0.16	BB097	-0.3		
		BB134	0.18	BB124	-0.29		
		BB145	0.06	BB059	-0.28		
		BB151	0	BB058	-0.27		
				BB125	-0.25		
				BB051	-0.12		
				BB123	-0.09		
				BB110	-0.02		

test set compounds from all sub-subgroups form the test set for subset 2. Compounds not selected as test compounds are used as training set compounds for model construction. The resulting training and test sets derived from subset 2 are presented in Table 8

Subset 1 largely contains simple, nondrug-like molecules. Thus, the operational objective of developing a predictive ADMET tool to evaluate the BBB penetration profiles of drug-like molecules could be negated if subset 2 is not further

Table 9. Subset 2'^a

compd	logBB	compd	logBB	compd	logBB
BB048	-2.15	BB113	-0.16	BB055	-1.54
BB039	-2	BB145	-0.14	BB115	-1.42
BB099	-1.82	BB051	-0.12	BB045	-1.23
BB054	-1.57	BB123	-0.09	BB041	-1.06
BB037	-1.42	BB139	-0.07	BB150	-0.72
BB098	-1.34	BB146	-0.06	BB131	-0.36
BB040	-1.3	BB038	-0.04	BB096	-0.34
BB147	-1.3	BB117	-0.04	BB097	-0.3
BB087	-1.26	BB067	0	BB125	-0.25
BB047	-1.17	BB086	0.04	BB052	-0.18
BB053	-1.15	BB119	0.04	BB079	-0.05
BB056	-1.12	BB126	0.06	BB062	-0.02
BB046	-0.82	BB143	0.06	BB110	-0.02
BB148	-0.74	BB118	0.1	BB121	0
BB057	-0.73	BB042	0.11	BB149	0
BB049	-0.67	BB132	0.18	BB085	0.12
BB111	-0.67	BB066	0.22	BB065	0.14
BB050	-0.66	BB075	0.25	BB104	0.16
BB120	-0.52	BB144	0.36	BB142	0.35
BB076	-0.5	BB089	0.39	BB112	0.64
BB060	-0.46	BB064	0.44	BB101	0.82
BB127	-0.46	BB043	0.49	BB130	1.34
BB128	-0.44	BB138	0.5	BB133	1.38
BB074	-0.43	BB105	0.52	BB129	1.51
BB082	-0.31	BB116	0.52	BB103	1.64
BB140	-0.3	BB107	0.53		
BB078	-0.29	BB084	0.55		
BB124	-0.29	BB088	0.61		
BB059	-0.28	BB063	0.69		
BB112	-0.28	BB141	0.74		
BB058	-0.27	BB094	1.44		
BB061	-0.24				

^a The test compounds are listed in bold.

considered. Therefore, probing subset 2, based on IPE type 4 similarity, should be a meaningful step to provide insight into the different possible mechanisms of interaction between drug-like molecules and the BBB.

PAM analysis is again used to further divide the complete data set with the hope that subset 2 can be meaningfully broken into a set of clusters. As shown in Table 5, when the original data set is partitioned into three and four subgroups, the silhouette values in both cases reduced to 0.62 from 0.78. However, a silhouette value of 0.62 indicates that the separation is still significant. Only a silhouette value less than 0.3 would suggest a poorly resolved partitioning has been performed.¹⁵ When the data set is divided into four subsets, one subset contains only 14 compounds. A 14 compound training set may be too small to build a robust statistical model. Thus, the partitioning of the complete data set into three subsets, each containing an adequate number of compounds for QSAR analysis, is preferable to forming four subsets.

Subset 1 in the "three-group-data set" is identical to subset 1 in the "two-group-data set" reported in Table 7. To keep a consistent labeling system for all three subsets, this subset is renamed as subset 1', and the optimum QSAR model is termed model 1'. That is, subset 1' and model 1' are identical to subset 1 and model 1, eq 17, respectively.

Subset 2 from the "two-group-data set" is separated into two smaller subsets. These two new subsets are termed subset 2' and subset 3' and are given in Tables 9 and 10, respectively. Test set compounds chosen in the previous "two-group-data set" analysis are retained for these two new

Table 10. Subset 3'^a

compd	logBB	compd	logBB	compd	logBB
BB073	-1.4	BB134	0.75	BB070	-0.06
BB071	-0.42	BB136	0.97	BB072	-1.3
BB069	-0.3	BB102	1.03	BB077	-0.22
BB083	-0.18	BB044	1.06	BB081	-1.1
BB080	-0.1	BB093	1.06	BB100	0.99
BB109	0	BB090	1.2	BB108	0.4
BB095	0.24	BB092	1.23	BB122	1.06
BB091	0.36	BB135	1.37	BB137	0.59
BB106	0.39				

^a The test compounds are listed in bold.

subsets. QSAR models were constructed for subsets 2' and 3' and are given by eqs 18 and 19.

The optimum QSAR model for subset 2' is

$$\text{Log BB} = 0.66 - 0.01\text{PSA} + 0.25\text{ClogP} - 0.025\text{S}_{\text{sF}} - 0.11\text{Kappa3}$$

$$N = 63, r^2 = 0.69, q^2 = 0.66 \quad (18)$$

The QSAR model for subset 2' contains two topological descriptors, S_{sF} and Kappa3. S_{sF} is the electrotopological state index for a fluorine atom with one single bond. Kappa3 is generally considered an index which reflects the molecular shape of a molecule.

The best QSAR model for subset 3'

$$\text{Log BB} = -0.58 - 0.031\text{PSA} + 0.069\text{EcoH} - 0.095\Delta E_{\text{TT}}(\text{vdw})$$

$$N = 17, r^2 = 0.80, q^2 = 0.72 \quad (19)$$

In subset 3', ClogP is not significant in explaining drug BBB permeability. Instead, two other descriptors terms are found in its place. EcoH is the cohesive energy of the solute, that is, a measure of how well a compound self-aggregates. $\Delta E_{\text{TT}}(\text{vdw})$ is an explicit MI-QSAR descriptor of the van der Waals energy difference between the solute-membrane complex and the free states of the solute molecule and the membrane monolayer.

To better evaluate the merit of performing the subset division, a QSAR model of the complete set is needed as a reference. The combination of training sets from subset 1' and subset 2' and 3' forms an overall training set of 104 compounds with the remaining 46 compounds composing the test set for the complete 150 compound data set. The optimum model for this overall training set has also been built:

$$\text{Log BB} = 0.05 - 0.011\text{TPSA} + 0.19\text{ClogP}$$

$$N = 104, r^2 = 0.69, q^2 = 0.64 \quad (20)$$

To determine if additional information is captured by the subset models, as compared to the general model, correlations between the residuals of fit are determined for the corresponding training set compounds using the subset models and the general model. The results are presented in Table 11.

The high correlations between the residuals of fit of both model 1' and model 2' with the general model suggests that the general model captures most of the same information as

Table 11. Correlation of the Residuals of Fit between Subset Models and the General Model

	model 1'	model 2'	model 3'
general model	0.79	0.84	-0.37

Table 12. $r^2_{\text{prediction}}$ Of Predictions of the Test Sets Using the Best QSAR Models for Subset 1', Subset 2', and Subset 3' and the General Model, Eq 20^a

	test set 1'	test set 2'	test set 3'
model 1'	0.76	0.30	0
model 2'	-1.01	0.79	0.31
model 3'	-2.10	0.19	0.92
general model	0.56	0.47	0.56

^a Model 1' is derived from subset 1', model 2' from subset 2', and model 3' from subset 3'.

models 1' and 2'. There is very little correlation between the residuals of fit of model 3' and the general model indicating new/additional information is provided by model 3'. Moreover, the high correlations of model 1' and 2' with the general model only indicate that the general model possesses the similar overall trend of prediction as these subset models. However, the magnitude of the residual of fit is not reflected by the correlation values. Thus, a direct estimation of the accuracy of prediction, $r^2_{\text{prediction}}$ (see eq 11), for the test set is employed to further compare models. Predictions using all three subset models in the "three-group-data set", and using the general model, eq 20, have been performed on all three test sets and the results are given in Table 12.

Test set compounds are most accurately predicted using QSAR models derived from the training set of the same subset based on IPE type 4 (hydrogen bond acceptor) molecular similarity. Model 1' predicts test set 1' with an $r^2_{\text{prediction}}$ value of 0.76. Both model 2' and model 3' predict test set 1' more poorly than using the average value of the test set since the $r^2_{\text{prediction}}$ values are negative. Test set 2' is best predicted by model 2' with an $r^2_{\text{prediction}}$ of 0.79. However, this test set is only estimated fairly by models 1' and 3'. The most appropriate predictive model for test set 3' is model 3'. The general model predicts each of the three test sets with a similar significance but each test set less well than the individual models. The $r^2_{\text{prediction}}$ values for the general model are 0.56, 0.47, and 0.56, respectively, for the three test sets. It would appear that the general model possesses the capability to predict all test set compounds with an overall reasonable accuracy by compromising its ability to predict an individual set particularly well.

DISCUSSION

The "traditional" molecular similarity concept uses the entire structure of a molecule as the basis for estimating molecular similarity. 4D-MS permits molecular similarity comparisons based on specific pharmacophore features of a molecule termed IPE types. This breakdown of a molecule into IPEs is particularly advantageous when the objective of the study is to evaluate factors governing a biological system where the complete structure of a molecule is not involved in expressing biological action. For instance, if the ability to form a hydrogen bond is dominant for activity determination, comparison of two molecules based on whole

molecule structure can be deceptive. 4D-MS allows multiple measures of molecular similarity to be made based on eight IPE types while expands the application of molecular similarity to identify common features in diverse chemistries. Moreover, 4D-MS provides a general set of numerical fingerprints, the eigenvalues from the MDDM, of a molecule.

The most direct application of molecular similarity is to identify, describe, and group similar molecules from a data set. However, molecular similarity is still a subjective notion that may only be validated by success with the task at hand. Only one similarity measure, based on IPE type 4 (hydrogen bond acceptor), is shown to be significant in separating this particular data set into subsets with respect to constructing more specific and predictive QSAR models for each subset as compared to the best QSAR model for the complete data set.

The clustering strategy employed in this study provides a quantitative and convenient way to divide a large diverse data set into smaller focused subsets. After dividing the original data set into three subsets using PAM, more significant QSAR models can be built for the subsets. The value of r^2 is 0.69 and q^2 is 0.60 for the general QSAR model. r^2 increases to 0.85 and q^2 increases to 0.83 for subset 1', while r^2 and q^2 increase to 0.80 and 0.72, respectively, for subset 3' after PAM data dividing. For subset 2', r^2 and q^2 remain comparable to those of the general model. More importantly, a criterion to assess the predictive power of a model, $r^2_{\text{prediction}}$, is developed for all of the models. The subset models better predict test set compounds that fall within the subset than the general models or the models derived from other subsets. The collection of subset models can be viewed as a consensus model that can be used as the replacement of the less reliable general model. These results imply that when predicting BBB penetration of a new compound, it is best to assign the new compound to its most appropriate subset and correspondingly use the QSAR model specific to that subset to estimate the new compound BBB permeability.

Subset 1' is composed of simple, relatively nondrug-like molecules. The log BB values are closely correlated to ClogP and the total electrostatic interaction energy of the solute with the membrane when the solute is located at its optimum position in the membrane. The octanol-water partition coefficient has long been recognized as a significant factor in solute BBB penetration. The second descriptor in the model, Echg, indicates that the electrostatic interaction between the solute molecule and the membrane is another important factor governing solute (drug) BBB penetration. When a solute possesses charge sites, and/or is polarized, the interactions between the solute and the polar head region of the membrane can impede movement of the solute in the membrane, thus decreasing BBB penetration.

ClogP is also found as a significant descriptor in the model for subset 2'. However, Echg is not seen to be important for BBB penetration of compounds in this subset. Instead, three alternate descriptors, TPSA, S_sF, and Kappa3, are presented in the QSAR. TPSA is the total surface area of polar atoms in a molecule, which is well recognized as being indicative of the ability of a molecule to form hydrogen bonds. TPSA may also capture some of the same information as Echg.

The remaining two QSAR descriptors are Kier and Hall's topological terms. S_sF is the electrotopological index for a

fluorine atom and Kappa3 is a molecular shape index. Unfortunately, the abstract concept of topological indices provides little mechanistic information on the correlation between drug BBB penetration and readily recognized molecular features.

Despite the molecular similarity among compounds in subsets 2' and 3', both of which belong to subset 2, QSAR models obtained for these two subsets are different, with only one common descriptor, TPSA. Two additional descriptors are selected in model optimization. Ecoh is the cohesive energy of the solute. Another descriptor in model 2' is ΔE_{TT} (vdw), the change in the van der Waal energy of the solute-membrane system due to the uptake of the solute from the free state to the position corresponding to the lowest energy state of the model solute-membrane complex.

Models for subsets 1', 2', and 3' capture a range of factors affecting solute BBB penetration. Moreover, individual factors seem to influence only certain classes of compounds. The largest subset, subset 2', has 80 compounds, while the other two smaller subsets, subset 1 and subset 3', represent "minorities". The QSAR models obtained for subset 2' are most similar to the general model of the complete data set. In contrast, models for the "minority" subsets, eqs 17 and 19, are relatively diverse from the general model. It seems that the factors governing the penetration of "minor" compounds are "shielded" by the factors governing the major class, that is those compounds of subset 2'. Thus, the "minority" descriptor do not survive in a general model built for a large diverse set.

Overall, it appears to be advantageous to build consensus models, consisting of individual QSARs for each compound class, for a large data set. Such consensus models should provide the most accurate prediction and the best mechanistic probing of the endpoint biological process.

ACKNOWLEDGMENT

Partial funding for this study was provided by the National Institute of Health Grant P01-GM 62195. Resources of Laboratory of Molecular Modeling and Design at UIC and The ChemBats21 Group, Inc. were used in performing this work. An unrestricted financial gift from the Procter and Gamble Company is also gratefully acknowledge.

REFERENCES AND NOTES

- (1) Goldstein, G. W.; Betz, A. L. The blood-brain barrier. *Sci. Am.* **1986**, 255, 74–83.
- (2) Pardridge, W. M. CNS drug based on principle of blood-brain barrier transport. *J. Neurochem.* **1998**, 70, 1781–1792.
- (3) Mouritsen, O. G.; Jorgensen, K. A new look at lipid-membrane structure in relation to drug research. *Pharm. Res.* **1998**, 15, 1507–1519.
- (4) Clark, D. E. Rapid calculation of polar molecular surface and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **1999**, 88, 815–821.
- (5) Crivori, P.; Cruciani, G.; Carrupt, P.; Testa, B. Predicting blood-brain barrier permeation using three-dimensional molecular structure. *J. Med. Chem.* **2000**, 43, 2204–2216.
- (6) Kulkarni, A.; Han, Y.; Hopfinger, A. J. Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2002**, 42(2), 331–342.
- (7) Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. J. Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res.* **2002**, 19, 1611–1621.
- (8) Keseru, G. M.; Molnar, L. High-throughput prediction of blood-brain partitioning: A thermodynamic approach. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 120–128.
- (9) Platts, J. A.; Abraham, M. H.; Zhao, Y.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and prediction of a large blood-brain distribution data set—an LFER study. *Eur. J. Med. Chem.* **2001**, 36, 719–730.
- (10) Langer, T. Molecular similarity characterization using CoMFA. *Perspect. Drug Discov. Des.* **1998**, 12(14), 215–231.
- (11) 4D-QSAR Molecular Similarity Program. Version 1.0, The ChemBats21 Group Inc., Lake Forest, IL, 2001.
- (12) Barnard, J. M.; Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 644–649.
- (13) Norinder, U.; Sjöberg, P.; Osterberg, T. Theoretical calculation and prediction of brain-blood partitioning of organic solutes using Molsurf parametrization and PLS statistics. *J. Pharm. Sci.* **1998**, 87, 952–959.
- (14) Duca, J. S.; Hopfinger, A. J. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1367–1387.
- (15) Kaufman, L.; Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*; Wiley: New York, 1990.
- (16) S-plus 6 for Windows. Insightful Corp, Seattle, WA, 2001.
- (17) Hartigan, J. A.; Wong, M. A. A k-means clustering algorithm. *Appl. Statistics* **1979**, 28, 100–108.
- (18) HyperChem Program Release 5.01 for Windows, Hypercube, Inc., 1996.
- (19) Doherty, D. C. MOLSIM User's Guide. The ChemBats21 Group, Inc., Lake Forest, IL, 1997.
- (20) Cerius 2, Molecular Simulation Package. Ver 3.0. MSI, San Diego, CA, 1997.
- (21) ClogP Daylight Chemical Information Software, Ver 4.51. Daylight Chemical Information Inc., Los Altos, CA, 1998.
- (22) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, 43, 3714–1717.
- (23) MI-QSAR Program. Version 1.0, The ChemBats21 Group Inc., Lake Forest, IL, 2000.
- (24) Rogers, D. G.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 854–866.

CI0498057