Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates

Marc Adenot* and Roger Lahana

Synt:em, Parc Scientifique G. Besse, 30000 Nimes, France

Received September 15, 2003

The aim of this article is to present the design of a large heterogeneous CNS library (∼1700 compounds) from WDI and mapping CNS drugs using QSAR models of blood-brain barrier (BBB) permeation and P-gp substrates. The CNS library finally includes 1336 BBB-crossing drugs (BBB+), 259 molecules non-BBB-crossing (BBB−), and 91 P-gp substrates (either BBB+ or BBB−). Discriminant analysis and PLS-DA have been used to model the passive diffusion component of BBB permeation and potential physicochemical requirement of P-gp substrates. Three categories of explanatory variables (C_{diff}, BBB_{pred}, PGP_{pred}) have been suggested to express the level of permeation within a continuous scale, starting from two classes data (BBB+/BBB−), allowing that the degree to which each compound belongs to an activity class is given using a membership score. Finally, statistical data analyses have shown that some very simple descriptors are sufficient to evaluate BBB permeation in most cases, with a high rate of well-classified drugs. Moreover, a "CNS drugs" map, including P-gp substrates and accurately reflecting the in vivo behavior of drugs, is proposed as a tool for CNS drug virtual screening.

INTRODUCTION

Absorption, Distribution, Metabolism, and Excretion (ADME) predictions have been extensively studied in the past few years. The interest of predicting ADME in large virtual libraries is now of growing interest for pharmaceutical companies, in particular those focusing on the discovery and development of central nervous system medicines where crossing the blood-brain barrier is a mandatory step for drug distribution. Our own technology platform has been successfully used to develop peptidic drug vectors for brain penetration as well as new chemical entities.[1] The computing technology is now used to bypass the need for time- and resource-consuming high throughput screening of massive compound libraries and reduces the amount of medicinal chemistry required. Penetration through the blood-brain barrier (BBB) is necessary for a drug to reach its required concentration. However, experimental data on BBB permeation using experimental in vitro or in vivo models remain difficult to obtain routinely. This explains the need for reliable in silico models to screen virtual libraries of drug compounds. Some qualitative- and often intuitive-rules are currently used by medicinal chemists to predict the BBB permeation of compounds. In general, most lipophilic drugs undergo a transcellular transport through plasma membranes, but efflux transport systems at the BBB provide a protective barrier and reject certain compounds from the brain. The computational model described in this article has been designed to filter CNS drugs from large virtual libraries, taking into account simultaneously both diffusion and efflux transport components.

Drug transport from blood to organs is a continuous variable that can be thought of as the summation of flux from five different routes:

(i) **Passive diffusion** is the primary process of translocation from the blood stream to the brain for most drugs. For a given membrane or barrier characteristics (surface area, water pores, tight junction diameters, histomorphology...), the physicochemical parameters of molecules represent the major rate determinant for passive diffusion, and, for this reason, it is hitherto the most computable and measurable property which can be used as a predictor of BBB permeation. It depends on simple parameters that make sense for medicinal chemists (molecular weight, H-bond capacity, rotatable bonds, solvent-accessible surface areas...) and which are currently used for screening databases.

(ii) **Paracellular transport** is typically used as a transport pathway for small hydrophilic molecules in all organs but the brain. Brain microvessel endothelial cells are characterized by complex tight junctions, the absence of transport pores, a distinct morphological structure, and a low number of pinocytotic vesicles. In these conditions, the entry of small hydrophilic drugs from blood to the brain is very restricted unless specific transporters exist at the cell membranes.

(iii) **Active transport** or facilitated diffusion: more than 20 transporters have been identified that allow nonlipophilic molecules such as glucose, amino acids, small monocarboxylic acids, choline, vitamins, nucleosides, thyroid hormones, and peptides to enter the brain. Active transport concerns essentially known chemical series of endogenous biomolecules.

(iv) **Endocytosis** is an energy-dependent transport process triggered by a nonspecific electrostatic adsorption of polycationic substances or by receptor binding of a substrate.

(v) **Efflux Pumps.** An increasing number of transporters have been discovered recently that pump drugs out of cells. The ATP-binding cassette (ABC) transporter superfamily includes the multidrug resistant gene product 1, also known

* Corresponding author phone: +33 (0)4 66 04 22 85; fax: +33 (0)4 66 04 86 67; e-mail: madenot@syntem.com.

10.1021/ci034205d CCC: $27.50
Published on Web 01/08/2004

as P-glycoprotein (P-gp), and multidrug resistance associated proteins (MRP1 to MRP9). Discovered first, P-gp is a membrane transport protein, very abundant in the luminal membrane of the endothelial cells comprising the blood-brain barrier. Substrates of P-gp are generally chemically unrelated drugs. The efflux transport itself is a continuous variable, depending on the drug binding to P-gp, the level of expression of P-gp, the drug concentration, and a constant equilibrium. The presence of P-gp inhibitors or modulators could modify this transport component significantly. Practically, for a majority of drugs, the resulting BBB permeation can be thought as the summation of passive diffusion and efflux transport components.

The main published BBB permeation models are essentially classical regression equations. The first original publication in this field is the one from Levin[2] who showed a relation between the brain capillary permeability coefficient and $\mathrm{Log}(P(MW)^{-1/2})$ for molecules with MW < 400. The very restrictive domain of use for this rough model has led to more or less sophisticated models (from multiple regression to PLS) using new descriptors selection like the partition parameter ($\Delta\log P$)[3], polar surface area,[4−7] accessible surface area,[8] polarizability parameter, H-bonds, acidity and basicity parameters,[9,10] thermodynamics,[11,12] 3D-molecular fields derived parameters,[13,14] number of H-bonds,[15] or E-state indices.[16] A publication from Ajay,[17] who designed a large HTS-purposed CNS-library, should also be mentioned. In this work, an extensive description of a library was performed with 1D and 2D parameters, and a neural network was used to classify compounds.

In the present study, a large combination of diverse descriptors, combining molecular properties, surface areas, electronic parameters, and topology has been used to describe a large diverse library. Discriminant analysis, although a method of choice for classifying two-class data—published experimental logBB data are rare[18]—has not often been used in published papers to model BBB permeation or define CNS drug-like properties.

COMPUTATIONAL METHOD

The Need for a Drug Classification. Historically, the pharmacological activity of small molecules has been studied using QSAR approaches on a limited series of related compounds. This procedure is clearly inappropriate if the aim is to design diverse combinatorial libraries. Informative databases are needed to generate models. Specific issues related to large library design have been previously discussed by Ajay,[17] who made the first attempt to design a large library of CNS-active compounds from a list of 17 activities. Ajay used both the CMC[19] and the MDDR[20] databases, starting from more than 80 000 compounds and ending with 15 000 active and 50 000 inactive compounds. Generally speaking, the available reference chemical databases have a significant heterogeneity in the way biological activities are defined. We thus developed a specific application which can be used for all reference databases in order to homogenize their contents, to classify the activities in a coherent way, and to establish a relevant selection according to our criteria (here, CNS versus non-CNS drugs). Using such a tool, training validation or reference sets can be generated in an effective way for QSAR studies or virtual screening purposes. The database used here is the WDI [21] (version 1999, ∼62 000 compounds). An activity dictionary has been built, based on the ATC,[22] an alphanumeric system defining 5 different levels of description and currently used by the WHO. As an example, glicazide is described as follows: **A**. alimentary tract and metabolism / **A10**. drugs used in diabetes / **A10 B** oral blood glucose lowering drugs / **A10 BB** sulfonamides,-urea derivatives / **A10 BB 09** glicazide. This drug classification system enables the retrieval of all the existing chemical families related to a given therapeutic area or biological target and can be used as a basis for constructing multiscaffold models.

Database Cleaning. As mentioned by several authors (see, for example, ref 23) a number of activity classes present in the WDI are not relevant for this type of study, such as for instance, dermatological applications (anti-fungal, emollients...), local and topical applications (local anaesthetics), hospital applications (diagnostics agents, general nutrients...), and agricultural applications (insecticides, repellents, antiparasitic products). The ATC classification levels 3 and 4 are the most representative of the activities indexed in the WDI database. All molecules defined only within therapeutic activity levels 1 or 5 were removed from the reference database. Finally, 4919 molecules were kept in the WDI-_clean database.

Compound Selection Using ATC Classification System. The query was simply made by selecting the "N0" class (Nervous system) from the WDI_clean database. This selection based on the ATC code leads to 1889 molecules used in neuropsychiatry. A unique pharmacochemical class has finally been attributed for each molecule from their PT (Pharmaco/Therapeutic) field: the class names have been chosen based (i) on the neuropsychiatric activity only (for instance, "SPASMOLYTICS, CARDIANTS" is kept as "SPASMOLYTICS") (ii) on the main site of activity (central or peripheric), and (iii) on the main well characterized pharmacological effects. For instance, "SPASMOLYTICS, PARASYMPATHOLYTICS" is kept as "PARASYMPATHOLYTICS" because PARASYMPATHOLYTICS encodes the information for the spasmolytic mechanism of action. "ANTIPARKINSONIANS, PARASYMPATHOLYTICS" is kept as "PARASYMPATHOLYTICS" because PARASYMPATHOLYTICS encodes the information for the antiparkinsonian mechanism of action (in this case, only a peripherical effect on trembling).

A detailed examination of N0 drugs has shown that 706 of them are not strictly CNS-active (peripheral mechanism of action) or have an ambiguous status regarding their passage through the BBB. These molecules have been consequently removed as well as 22 N0 molecules that are known to be unable to cross the BBB, like catecholamines or ammoniums (Table 1). The remaining molecules are definitely considered as CNS+ as well as BBB+ compounds. CNS activity implies BBB permeation, but the converse is not necessarily true: some drugs that are CNS inactive may still cross the BBB and show no activity because they do not interact with any CNS targets; similarly, some drugs with an expected peripheral site of action may cross the BBB, leading to unwanted CNS side-effects. Some confusion remains in published data sets if the models are focused on CNS activity or BBB crossing, while these two behaviors are not necessarily equivalent and do not encode the same

Table 1. Molecules from the ATC N0 Class that Have Been Removed from the Set of CNS+ Compounds

class	
alkaloids	4
analeptics	13
analgesics	233
anticholinesterases	12
antihistamines	4
antiinflammatories	3
antiparkinsonians	13
antipyretics	3
dopaminergics	10
local anesthetics	16
lysergics	10
nitric acid derivatives	4
papaverine-like spasmolytics	37
parasympatholytics	54
parasympathomimetics	14
catecholamines	17
ammoniums	4

Table 2. 1336 BBB+ Compounds with or without CNS Activity that Have Been Used To Compute the Blood-Barrier Training Set Model

class	
ATC N0 Molecules: CNS+, BBB+	
amino acids	2
antiinflammatories	1
alcaloids	5
amines	18
analeptics	9
analgesics	2
anilides	8
anticholinesterases	2
anticonvulsants	69
antidepressants (except tricyclics and MAO-I)	156
antiparkinsonians	11
B1 vitamines	8
benzodiazepines	80
convulsants	1
corticosteroids	160
cyclic ureas	42
general anesthetics	40
H1-antihistaminics	1
MAO inhibitors	27
lysergics	8
morphinics	99
narcotics	12
neuroleptics (except tricyclics)	132
nootropics	37
profenes	1
psychotonics	37
salicylates	6
sedatives	73
setrons	3
sulfamides	1
tranquilizers (except benzodiazepines)	121
tricyclics	107
xanthines	5
ATC Non-N0 Molecules: CNS±, BBB+	
antibiotics	10
antivirals	9
antifungics	3
statines	10
vitamins	6

level of information. Forty-five non-N0 molecules that are known to cross the BBB were also added to the 1291 BBB+ compounds, so that the final number of compounds in the BBB+ set is 1336 (Table 2).

Identifying BBB− compounds is by far the most difficult task of such a design for the simple reason that, in most cases, this information is not available. It is not clear, for instance, if a compound which is presented as a cholagogue or antiseptic would cross the BBB or not. Two hundred sixty-nine WDI non-N0 molecules that are known not to cross the BBB have been used to define the BBB− class: amines (except some amphetamines that have already been included in BBB+ set), ammoniums IV, and antibiotics from different chemical families (most of them do not cross the BBB under normal conditions): aminosides, beta-lactamines, tetracyclines, macrolides, quinolones, intercalating agents, and miscellaneous (fusidic acid, teicoplanin...). Ninety-one non-CNS molecules that are pure P-gp substrates (not inhibitors nor modulators) have been identified from the literature[24] and have been retrieved from the WDI, stored in a PGP+ subset, and included in the BBB− set, so that the final number of compounds in this set is 360. The total number of compounds in this study is then 1696.

There are still some practical limitations to every BBB library selection. As mentioned above, one of the major problems in compound selection is the confusion between CNS drugs and BBB drugs. Prodrugs, like paracetamol or benorilate, have been included in the BBB+ set although they cannot cross the BBB by themselves but need a previous metabolic modification. Conversely, catecholamines (adrenaline, noradrenaline, dopamine) are naturally occurring CNS molecules that are released in situ but are too polar to enter the CNS from blood by themselves. They have been classified in the BBB− set. Moreover, they are quickly hydrolyzed in the plasma before they can reach the BBB. Chemically modified amines such as amphetamines or dopamine-derived antiparkinsonians have been considered as BBB+ compounds. Finally, whatever its potential CNS activity, a xenobiotic will reach the CNS depending on at least 3 factors: (i) the existence of a transport pathway; (ii) its chemical state (ionization state,...); and (iii) its enzymatic transformation which raises the question of the exact chemical identity of the molecule which passes through the BBB (i.e. are the "calculated" molecule and the real molecule exactly the same entity?). All compounds have been calculated in their neutral form.

Test for External Prediction. The compounds used as a test set for external predictions were obtained from the literature. They consist of 82 compounds chosen from among the 108 presented by Crivori et al.[14] and additional compounds (vitamin derivatives, oxicams, and quinolones). The compounds remaining from the Crivori set have been discarded from the selection because they were already used in the training set or were not available in the WDI database. The test set includes 20 BBB+ drugs and 62 BBB− drugs.

P-gp Substrates Selection. The original training set consists of 91 P-glycoprotein substrates (PGP+ class) selected from the literature[24] and available from the WDI database.[21] The 1595 remaining compounds are considered as nonsubstrates (PGP− class). P-gp modulators or inducers have been discarded from the selection, assuming they do not act via the same mechanism or interaction pattern than substrates. However, compounds such as verapamil that are considered both substrates and modulators have been included anyway. Some of the PGP− compounds have in fact an undetermined status in terms of their interaction with P-gp and some doubt must exist as to whether they should be considered as PGP− by default. As a consequence, the model

Table 3: P-gp Substrates and Nonsubstrates Selection for the Test Set

P-gp substrates	P-gp nonsubstrates
amiodarone	aldicarb
azidopine	atrazine
bisanthrene	carmustine
catharanthine	cysteine-methylester
cepharanthine	deoxypodophyllotoxine
dibucaine	farnesol
endosulfan	leptophos
hydroxyrubicin	melphalan
K02	paraquat
methylreserpate	phosmet
mithramycin a	reserpinic acid
mitoxanthrone	tamoxifen
morphine-6-glucuronide	triforine
nimodipine	
pafenolol	
phenoxazine	
rhodamine 123	
spiperone	
tetraphenyl-phosphonium bromide	
vindoline	

will not be representative for a subset of unknown "true" P-gp substrates, and consequently the chance of finding false negatives will be increased. The test set consists of 33 other compounds from the Seelig set,[21] including 20 P-glycoprotein substrates and 13 P-glycoprotein nonsubstrates (Table 3).

Structural Modeling and Calculation of Descriptors. All the calculated molecular parameters were grouped into 6 different sets (Table 4). 2D structures from the WDI were converted to 3D structures using the Corina[25] software. Structures were fully minimized using the Vamp[25] standard converger and the AM1 Hamiltonian type. Missing data were reported as follows: 122 missing data (including 19 PGP+ compounds) for electronic parameters, 65 for molecular surface area, and 45 for molar refractivity.

Permeation Score. Lipinski's rules[26] were aimed at detecting poor solubility and reflecting membrane permeability of drugs (molecular weight < 500, ClogP <5, H-bond donor <= 5, H-bond acceptor <= 10). Some authors[4,5,15,33] proposed the molecular polar surface area as an additional determinant for membrane permeability. These five descriptor values were automatically scaled within the range [0;5] for each compound, using predefined plateau functions. Then, the five permeation components were expressed as a unique pseudocontinuous variable C_{diff}, varying between 0 and 1 and suitable for discriminant analysis or regression

$$C_{diff} = \Sigma_{i=1,5}\, S(i)/5^2$$

where $S(i)$ is a scaling function for the ith molecular parameter. Scaling functions are characterized by their respective slopes a_i, lower threshold t_i, and upper threshold t'_i: ($a_1 = -6.7 \times 10^{-3}$, $t_1=0$, $t'_1=800$), ($a_2 = +0.4$, $t_2=-2.5$, $t'_2 = 5$), ($a_3 = -1$, $t_3=0$, $t'_3=5$),($a_4 = -0.5$, $t_4=0$, $t'_4=10$), and ($a_5 = -0.02$, $t_5=0$, $t'_5=250$) for molecular weight, ClogP, H-bond donor and acceptors, and polar surface area, respectively. The permeation score C_{diff} reflects the level of permeation directly from many molecular parameters onto a unique continuous scale.

Data Analysis. Linear Discriminant Analysis. Blood-brain barrier permeation is a continuous variable by nature which depends in particular on drug concentration, administrated dose, and histological conditions. However, quantitative data are not abundant,[5,6] and consequently, experimental BBB permeation measurements are often ambiguous and potentially unreliable. A two-class description (BBB+/BBB−) is certainly a less informative but more reliable way of describing such data and discriminant analysis is the method of choice to model BBB permeation. This technique aims at separating different classes of compounds using a pool of different explanatory variables and may be used to predict class membership, in this case BBB+ versus BBB−. It can also be used as a way to highlight the most discriminating variables ("key descriptors") in the context of a variable selection problem.

PCA/PLS. Principal Components Analysis (PCA) and Partial Least Squares (PLS) are now well recognized multivariate methods for reducing large multidimensional data tables in the most informative way. They have been extensively used for QSAR, to predict ADME properties, pharmacological activity, or toxicity. All column data were scaled to unit variance so as to give every variable the same

Table 4. Molecular Descriptors Selections

variable selection 1	ADME screen	no. of atoms − no. of halogens − no. of heteroatoms − molecular weight − HBond_acceptors − HBond_donors − CLOGP (Sybyl) − rotatable bonds
variable selection 2	molec attributes (geometry)	molecular volume − inertia moment size 1 − inertia moment size 2 − inertia moment size 3 − inertia moment length 1 − inertia moment length 2 − inertia moment length 3 − ellipsoidal volume − molecular refractivity − total surface area (Sybyl)
variable selection 3	topology	Kier Chi0, Kier Chi1, Kier Chi2, Kappa1 index, Kappa2 index, Kappa3 index, Kalpha1 index, Kalpha2 index, Kalpha3 index, shape flexibility, Randic index, Balaban index, Wiener index
variable selection 4	VAMP electronic parameters	total dipole − dipole component X − dipole component Y − dipole component Z − polarization XX, XY, XZ, YY, YZ and ZZ
variable selection 5	VAMP energy parameters	total energy − electronic energy − heat of formation − LUMO − HOMO
variable selection 6	Sybyl surface areas	PSA: polar surface area TSA: total surface area PPSA1, partial positive surface area PPSA2, total charge weighted PPSA PPSA3, atomic charge weighted PPSA PNSA1, partial negative surface area PNSA2, total charge weighted PNSA PNSA3, atomic charge weighted PNSA DPSA = PPSA − PNSA, differential PSA FPSA = PPSA/TSA, fractional positive SA FNSA = PNSA/ TSA, fractional negative SA

Table 5. P-Glycoprotein Substrates Models: Discriminant Analysis[a]

variable selection number	step	variable entering	total fraction well classified	fraction of PGP− well classified	fraction of PGP+ well classified
1	1	HB donors	0.82	0.83	0.53
1	2	number of atoms	0.83	0.84	0.51
1	3	CLOGP	0.84	0.86	0.53
1	4	number of halogen atoms	0.85	0.86	0.56
1	5	number of heteroatoms	0.85	0.87	0.56
2	1	inertia moment 1 size	0.90	0.93	0.38
2	2	inertia moment 3 length	0.91	0.94	0.41
2	3	inertia moment 3 size	0.91	0.94	0.41
2	4	inertia moment 2 size	0.91	0.95	0.38
3	1	Wiener index	0.93	0.96	0.32
4	1	polarization XX	0.72	0.73	0.63
4	2	polarization ZZ	0.74	0.75	0.63
4	3	polarization YY	0.74	0.75	0.65
4	4	polarization YZ	0.75	0.76	0.67
4	5	total dipole	0.76	0.76	0.63
4	6	polarization XY	0.76	0.77	0.63
5	1	electronic energy	0.75	0.76	0.53
5	2	total energy	0.80	0.82	0.42
5	3	LUMO	0.80	0.82	0.36
5	4	HOMO	0.80	0.82	0.42
5	5	heat of formation	0.82	0.83	0.46
6	1	PPSA2	0.90	0.93	0.38
6	2	PNSA2	0.91	0.94	0.40
6	3	PNSA1	0.91	0.94	0.38
6	4	FPSA3	0.91	0.94	0.37
6	5	PSA	0.91	0.94	0.38
other	1	PLS-DA PGP_{pred}	0.94 (*0.79*)	0.96 (*0.92*)	0.74 (*0.70-0.80*)

[a] Cross-validation method: leaving-out each of 3 groups in turn. Values in parentheses are for the test set.

influence in the data analysis. PCA can be used to reduce a large number of variables to a smaller number of components by transformation into an orthogonal set of linear combinations maximizing the amount of total variance from the original set. PLS is a numerical method aimed at finding the linear relationship between an activity matrix and a numerical descriptors matrix. PLS modeling consists of simultaneous projection of both the pharmacological and chemical spaces onto low-dimensional hyper planes. The model is calculated so as to simultaneously minimize the residuals and yield latent variables which are optimally correlated. The statistical significance of the PLS model is determined by cross-validation.

PLS-DA: PLS-DA is a practical way to transform a *class*-model into a pseudo*continuous variable*-model. The derivation of a PLS equation from initial X-block and Y-block leads to a new quantitative univariate Y-variable varying between two indices of class (0 for BBB− and 1 for BBB+). The predicted variable can be thought of as a predictor of blood-brain barrier permeation on a scale from 0 to 1. Discriminant analysis can be used to determine the threshold of a PLS-predicted variable that discriminates optimally between the two original classes. The predicted variable value weights predictions in each class, compounds with a predicted Y around the threshold value being more uncertainly predicted than those with a predicted Y value close to the class indices. With PLS-DA, r^2 and q^2 criteria are meaningless despite the existence of a regression equation. The validation criteria are those of the discriminant analysis (rate of prediction, confidence estimate). The same techniques are applied to P-gp predictions considering a two-class model (PGP−/ PGP+). The predicted values, i.e. PGP_{pred} and BBB_{pred} for P-gp substrates and blood-brain barrier permeation modeling, respectively, allows that the degree to which each compound belongs to an activity class is given using a membership score.

All statistical analyses were performed using the Tsar software.[25,27]

RESULTS AND DISCUSSION

P-gp Substrate Models. Linear Discriminant Analysis. Ninety-one P-gp substrates were included in the analysis using various descriptor selections. The objective was to delineate the minimum descriptor selection set that is sufficient to discriminate between PGP+ and PGP− classes. PGP+ compounds have a sparse distribution in the descriptor spaces, so that the two classes appear to be closely embedded. This explains a rather low fraction of well-classified PGP+ compounds (Table 5), that does not exceed 67% using polarization components.

PLS-Discriminant Analysis. A PLS-DA model using all the 67 molecular parameters in Table 4 has shown that the predicted PGP value (PGP_{pred}) alone is able to classify correctly up to 74% of PGP+ compounds and 96% of PGP−. Under these conditions, using PGP_{pred} as a discriminant variable, a threshold value of 0.16 was fixed to delineate PGP+ and PGP− compounds. Excluding molecular surface areas from the calculation has a significant effect on the classification rate causing the fraction of well-classified PGP+ compounds to fall down to 65%. However, the exclusion of electronic parameters has only a minor effect on the classification rate (72% of well-classified PGP+ drugs instead of 74%) (Table 5) and eliminates most of the missing data due to calculation failures over complex structures. The PGP_{pred} threshold value is unchanged. A PGP_{pred} distribution

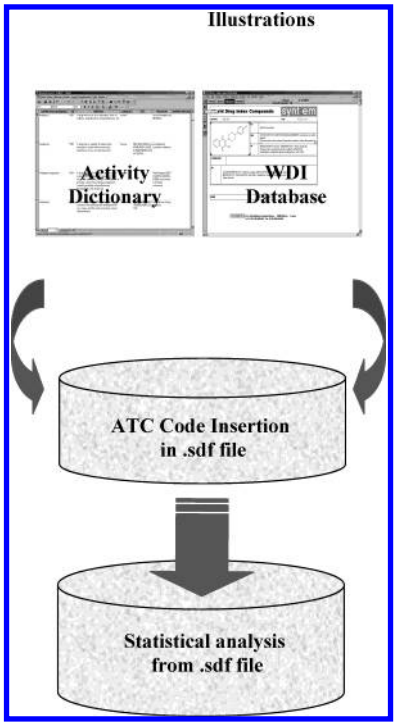

Figure 1. Work flow: The activity dictionary covers the description of 50%WDI activity terms and is updatable. Each activity is defined by its definition, synonyms, ATC code, key words, application (human, animal, or vegetable). The WDI database has been cleaned in order to keep only molecules with a INN or USAN name and a defined structure. The molecules have one or more activities. Salts, complexes, and peptides have been discarded.

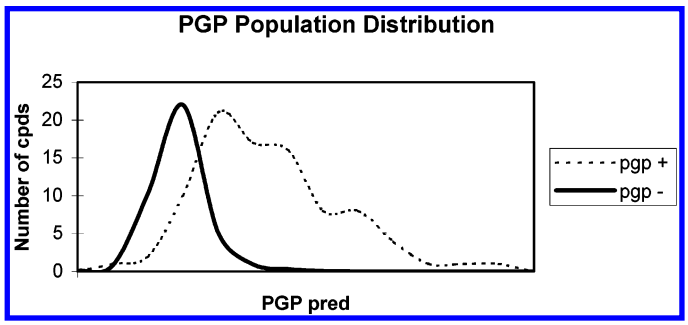

Figure 2. P-gp population distribution analysis between P-gp substrates and P-gp nonsubstrates among 1696 WDI drug compounds. PGP_{pred} stands for the PGP predicted value (a high PGP_{pred} value reflects a better chance for the molecule to be a P-gp substrate). For clarity, the number of PGP− compounds have been scaled by a factor 0.025 for PGP− compounds.

analysis (Figure 2) shows that PGP+ and PGP− classes are characterized by two relatively distinct distributions: the P-gp nonsubstrates distribution has a relatively narrow shape profile centered on $PGP_{pred} = 0.04$ (± 0.09) and varies in the range [−1.62; 0.68]. In contrast, the P-gp substrate distribution is a flatter shape profile centered on $PGP_{pred} = 0.29$ (± 0.19) and varies in the range [−0.18; 0.87]. A positive skewness indicates that PGP+ distribution is skewed toward higher PGP_{pred} values. This PLS model, including all descriptors except electronic parameters, was finally retained and applied to predict the 33 compounds from the test set. The level of "well-classified" was 70% for PGP+ compounds and 92% for PGP− at a threshold = 0.16. Note that given the existence of borderline points, the rate of well classified PGP+ can be improved up to 80% when just slightly lowering the threshold from 0.16 to 0.12 in the test set. While any treshold can be considered as a fuzzy value in a screening perspective, we have considered that the rate of well classified PGP+ lies between 0.7 and 0.8.

Comparison with Other Published P-gp Permeation Models. Some authors have attempted to define structure−activity relationships or pharmacophores for P-gp substrates.[32] In its multiple-pharmacophore model, Penzotti[28] highlights the amphiphilic nature of P-gp substrates. The rate of well classified compounds is 96% (79% in test set) for PGP− and 64% (53% in test set) for PGP+. Seelig[24] reports the presence of two or three electron donor groups with specific spatial separation. No classification or predictive models were presented. Ecker[29] describes a simple regression (r=0.93) between EC50 values for P-gp inhibition and hydrogen bond acceptor strength in a propafenone-related series. In a PLS-based study, Osterberg[30] reports the influence of molecular surface area, polarizability, and hydrogen bonding for P-gp associated ATPase activity taken from Litman.[31] The reported model shows values of r^2 between 0.72 and 0.78.

BBB Permeation Models. Linear Discriminant Analysis. To a first approximation, P-gp substrates have been explicitly included in the BBB model. A discriminant analysis based on all 67 molecular parameters in Table 4 has shown that up to 90% of the compounds are well classified, using such simple descriptors as the atom counts (heteroatoms, halogens, hydrogen bond donors, and acceptors) or differential polar surface area. The results from discriminant analysis, including P-gp substrates, are given in Table 6. BBB− compounds are characterized by a larger scattering within the parameter space and a lower well-classified fraction than BBB+ compounds, mainly due to the presence of PGP+ compounds. The score plot for principal components 1 and 2, using variable subselections 1 and 6 only, shows that PC1 discriminates between BBB+ and BBB− compounds (Figure 3). PGP+ compounds (not shown) are spread all over the plot, leading to a lowering of the BBB− classification rate when they are included in the BBB− set. For this reason, the P-gp substrates have been removed from the calculation. As a consequence, the analysis highlights variable selections 1 and 6 again but shows a significant increase, from 78% to 87%, in the number of well-classified BBB− compounds. The results from discriminant analysis excluding P-gp substrates are given in Table 7. It can be seen that taking the number of heteroatoms alone leads to the correct classification of 92% of the compounds. Most of the BBB− compounds have a number of heteroatoms larger than 8, while most of the BBB+ compounds have a number of heteroatoms lower than 9 (Figure 4). The fact that the number of heteroatoms well classifies most of the compounds is not surprising, as it is correlated to the polar surface area which, in turn, is known to reflect the brain uptake.[4−6] In this work, heteroatoms include oxygen and nitrogen as well as phosphorus, sulfur, and halogens. This explains why the heteroatoms cutoff is not in accordance with the rule proposed by Norinder[34] which states that molecules with less than five nitrogens and/or oxygens atoms have a high chance of entering the brain. Finally, the permeation score C_{diff} alone performs slightly better than the number of heteroatoms and performs as well as variable selection 1. Its use is recommended for obtaining a ranked list of compounds.

Table 6. Blood-Brain Barrier Permeation Models: Discriminant Analysis Including PGP Substrates[a]

variable selection number	step	variable entering	total fraction well classified	fraction of BBB− well classified	fraction of BBB+ well classified
1	1	number of heteroatoms	0.90	0.72	0.94
1	2	number of HB donors	0.91	0.74	0.96
1	3	number of halogen atoms	0.92	0.78	0.96
1	4	number of HB acceptors	0.92	0.78	0.96
2	1	inertia moment 1 size	0.81	0.42	0.91
2	2	inertia moment 3 length	0.84	0.53	0.92
2	3	inertia moment 2 size	0.84	0.53	0.92
2	4	inertia moment 3 size	0.85	0.56	0.92
3	1	Wiener index	0.85	0.40	0.97
4	1	total dipole	0.71	0.61	0.73
4	2	dipole X component	0.74	0.62	0.77
4	3	polarization XY	0.75	0.62	0.77
5	1	total energy	0.75	0.71	0.75
5	2	heat of formation	0.75	0.71	0.75
5	3	LUMO	0.75	0.74	0.75
6	1	DPSA2	0.87	0.58	0.94
6	2	DPSA1	0.90	0.66	0.97
other	1	C_{diff}	0.89	0.78	0.92
other	1	PLS-DA BBB_{pred}	0.91	0.73	0.97
other	1	LogBB (Clark)	0.79	0.63	0.84
other	1	LogBB (Van de Waterbeemd)	0.84	0.88	0.69

[a] Cross-validation method: leaving-out each of 3 groups in turn.

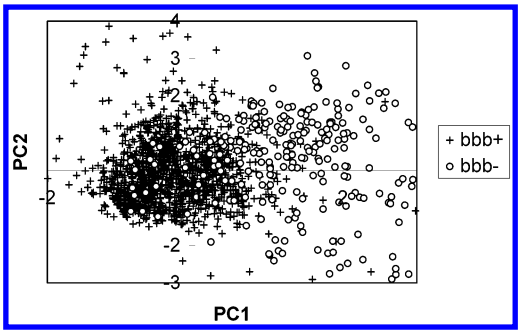

Figure 3. PCA plot for all compounds, except P-glycoprotein substrates, using variable subselection 1 (ADME) and 6 (surface areas). The three first principal components respectively explain 59%, 82%, and 93% of the total (cumulative) variance.

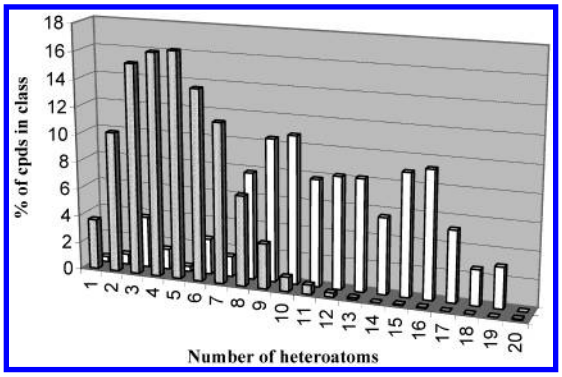

Figure 4. Distribution of BBB+ compounds (in gray) and BBB− compounds (in white) as a function of the number of heteroatoms in the molecule. A large majority of BBB+ compounds have no more than 9 heteroatoms in their structure.

PLS-Discriminant Analysis. A PLS-DA model using all the 67 variables available in Table 4 has shown that BBB_{pred} alone is able to classify correctly up to 98% of the compounds (Table 8). Under these conditions, using BBB_{pred} as a discriminant variable, a threshold value of 0.62 was fixed to delineate BBB+ and BBB− compounds. However, 201 compounds were excluded due to the absence of electronic parameters for complex structures. To avoid this disadvantage, alternative PLS-DA models were derived from other variable selections, which excluded electronic parameters. Variable selections 1 and 6, previously selected by discriminant analysis, were finally retained for this model. One thousand six hundred five compounds and 25 variables were included. The discriminant threshold is $BBB_{pred} = 0.63$. The rate of classification (Table 8) is slightly lower than in the "all-parameters" model, but it has the advantage that it can be used for molecules with missing electronic parameters. BBB_{pred} and C_{diff} are two ways of expressing the level of permeation within a continuous scale. These two variables are correlated (r^2=0.84), but performances are slightly in favor of BBB_{pred} with respect to C_{diff}. External predictions were processed over the test set. The rate of well-classified compounds was highly satisfactory (>90%) in both classes, even when retaining PGP+ compounds for predictions.

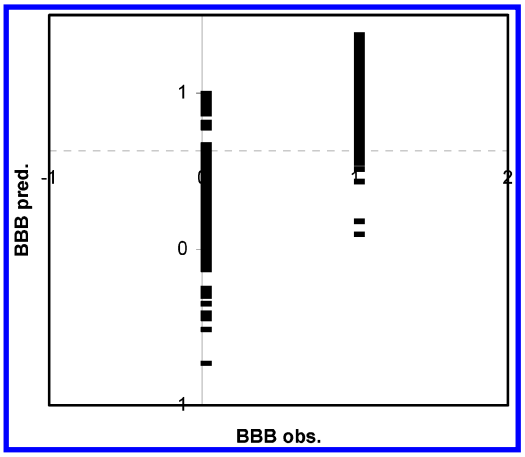

Figure 5. Observed BBB permeation versus Predicted BBB Permeation using the PLS-DA Model. A threshold value of 0.63 has been fixed to delineate BBB+ and BBB− compounds. Note that BBB_{obs} is a binary variable (0 for BBB− and 1 for BBB+) while BBB_{pred} is expressed on a continuous scale.

Comparison with Other Published BBB Permeation Models. Ajay[17] published rates of classification, but these

Table 7. Blood-Brain Barrier Permeation Models: Discriminant Analysis Excluding PGP Substrates[a]

variable selection number	step	variable entering	total fraction well classified	fraction of BBB− well classified	fraction of BBB+ well classified
1	1	number of heteroatoms	0.92	0.81	0.94
1	2	number of HB donors	0.94	0.80	0.97
1	3	number of halogen atoms	0.94	0.83	0.97
1	4	number of atoms	0.95	0.86	0.97
1	5	molecular weight	0.95	0.87	0.97
2	1	inertia moment 1 size	0.81	0.49	0.88
2	2	inertia moment 3 length	0.84	0.58	0.89
2	3	Inertia moment 2 size	0.84	0.60	0.89
2	4	inertia moment 3 size	0.86	0.65	0.89
2	5	inertia moment 1 length	0.86	0.65	0.90
3	1	Wiener index	0.85	0.46	0.93
4	1	total dipole	0.73	0.63	0.74
4	2	dipole X component	0.77	0.64	0.79
4	3	polarization XY	0.78	0.65	0.80
4	4	dipole Y component	0.78	0.65	0.80
5	1	total energy	0.75	0.74	0.75
5	2	electronic energy	0.76	0.83	0.75
6	1	PSA	0.89	0.73	0.92
6	2	FPSA2	0.93	0.80	0.96
6	3	PPSA1	0.94	0.96	0.82
6	4	FPSA3	0.95	0.83	0.97
6	5	FNSA3	0.95	0.82	0.97
6	6	FNSA2	0.95	0.83	0.97
6	7	DPSA1	0.95	0.85	0.97
6	8	PNSA1	0.95	0.85	0.97
other	1	C_{diff}	0.93 (*0.94*)	0.84 (*0.95*)	0.95 (*0.90*)
other	1	PLS-DA BBB_{pred}	0.97 (*0.91*)	0.89 (*0.92*)	0.99 (*0.90*)
other	1	LogBB (Clark)	0.87 (*0.90*)	0.71 (*0.93*)	0.91 (*0.80*)
other	1	LogBB (Van de Waterbeemd)	0.90 (*0.75*)	0.76 (*0.75*)	0.93 (*0.75*)

[a] Cross-validation method: leaving-out each of 3 groups in turn. Values in parentheses are for the test set.

Table 8. Blood-Brain Barrier Permeation Alternative Models: PLS-DA Results[a]

parameters selection	PGP+ set	total fraction well classified	fraction of BBB− well classified	fraction of BBB+ well classified
all	excl	0.98 (*0.84*)	0.90 (*0.81*)	0.99 (*0.90*)
all but electronic parameters	excl	0.97 (*0.91*)	0.88 (*0.92*)	0.99 (*0.90*)
ADME and surfaces	excl	0.97 (*0.91*)	0.89 (*0.92*)	0.99 (*0.90*)
all	incl	0.95 (*0.91*)	0.85 (*0.92*)	0.97 (*0.90*)
all but electronic parameters	incl	0.95 (*0.91*)	0.84 (*0.92*)	0.98 (*0.90*)
ADME and surfaces	incl	0.94 (*0.93*)	0.80 (*0.91*)	0.98 (*0.95*)

[a] The analyzed variable is PLS-DA BBB_{pred} value. Values in parentheses are for the test set.

results focused on CNS activity rather than BBB permeation stricto sensu. As an example, Ajay's rates of prediction are 92% for CNS+ and 71% for CNS− in a $n = 275$ compounds set. Basak[18] found rates of prediction of 100% for CNS+ and 91% for CNS−, based on topological indices only, using 28 compounds.

CNS Mapping. All compounds are characterized by both a passive diffusion component (BBB_{pred} or C_{diff}) and a P-gp efflux component (PGP_{pred}). Combinations of these two components can be used to map CNS compounds (Figures 7 and 8). This mapping clearly delineates three distinct areas: PGP+ drugs (with a PGP > 0.14), BBB+ drugs with C(P-gp) < 0.3 and C(diff) > 0.4, BBB− drugs with C(P-gp) < 0.3 and C(diff) < 0.4. The borders of these areas are fuzzy: score values just give a probabilistic weight to the prediction. Surprisingly, focusing on some P-gp substrates that have been predicted with low C(P-gp) values, like amitriptyline, chlorpromazine, morphine, and disulfiram, shows that all these compounds are CNS drugs anyway despite the fact that they are also P-gp substrates in vitro. It should be noted that some drugs apparently impermeable, such as cyclosporine A, become highly permeable when the P-gp is saturated. It is reassuring that the mapping of CNS drugs makes them appear as such and underlines the practical interest of such a mapping to predict the behavior of drugs under "real" conditions.

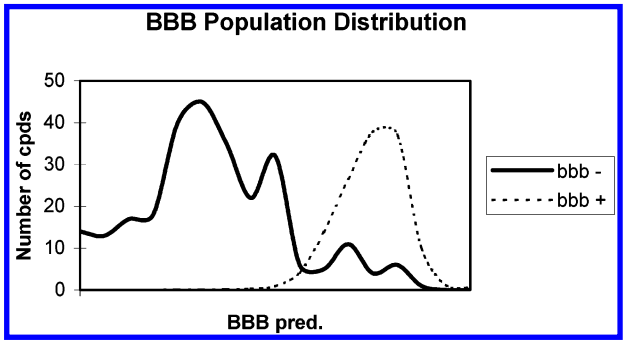

Figure 6. BBB population distribution analysis between BBB+ and BBB− compounds among 1696 WDI molecules. BBB_{pred} stands for the BBB predicted value (a high BBB_{pred} value reflects better chance for the molecule to cross the BBB). For clarity, the number of BBB+ compounds have been scaled by a factor 0.1.

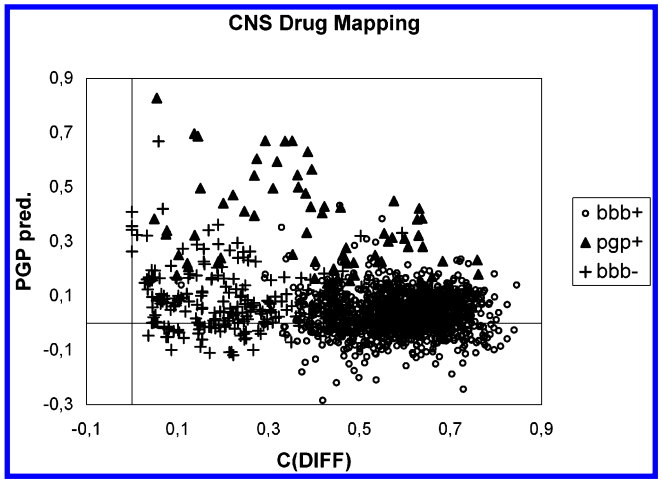

Figure 7. CNS drug mapping using PGP_{pred} and $C_{diff.}$.

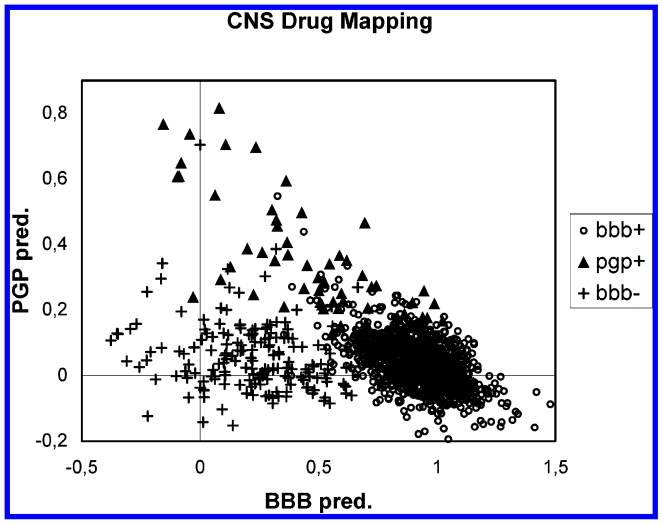

Figure 8. CNS drug mapping using PGP_{pred} and BBB_{pred}.

CONCLUSIONS

Prediction of BBB permeation directly from molecular structures is of great importance in the development of CNS agents, where a high penetration is needed, or in minimizing CNS-related side-effects of drugs with a peripherical mechanism of action. Lipinski[26] has shown that four simple current parameters are good predictors of drug absorption in a large panel of drug-like chemical series. In this study, we have also demonstrated that such a simple descriptor as the number of heteroatoms is sufficient to predict BBB permeation with a high rate of classification. There are many published regression equations describing the continuous variable LogBB, all taken from the original sets of Young,[3] Abraham,[9] or Lombardo.[11] There are many cases where variables such as PSA or ClogP are inaccurate or relying on particular algorithms and parameters, if not impossible to calculate. From a practical standpoint, this makes most of these equations difficult to use in virtual screening applications. Using the unique descriptor C_{diff}, derived from Lipinski rules, or PLS-derived variables such as BBB_{pred} has led also to a high rate of discrimination based on a large number of chemically diverse compounds. Moreover, the inclusion of P-glycoprotein transport as a component of BBB permeation allows a clear mapping of drugs in terms of their behavior toward blood-brain barrier permeation.

Supporting Information Available: 91 P-glycoprotein substrates. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floch, R.; Buelow, R. Computer-Assisted Rational Design of Immunosuppressive Compounds. *Nat. Biotechnol.* **1998**, *16*, 748−752.

(2) Levin, V. A. Relationship of Octanol/Water Partition Coefficient and molecular Weight to Rat Brain Capillary Permeability. *J. Med. Chem.* **1980**, *23*, 682−684.

(3) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. Development of a New physicochemical Model for Brain Penetration and Its Application to the design of Centrally Acting H2 Receptor Histamine Antagonists. *J. Med. Chem.* **1988**, *31*, 656−671.

(4) Van de Waterbeemd, H.; Kansy, M. Hydrogen-Bonding Capacity and Brain Penetration. *Chimia* **1992**, *46*, 299−303.

(5) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J. P. Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.

(6) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Pharm. Sci.* **1999**, *88*, 8, 815−821.

(7) Liu, R.; Sun, H.; So, S. Development of Quantitative Structure−Property relationship Models for early ADME Evakluation in drug Discovery. 2. Blood-Brain Barrier Penetration. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1623−1632.

(8) Jorgensen, F. S.; Jensen, L. H.; Capion, D.; Christensen, I. T. Prediction of Blood-Brain Barrier Penetration. In *Rational Approaches to drug Design*; Höltje H.-D., Sippl, W., Eds.; 2001; pp 281−285.

(9) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen Bonding. 33. Factors that Influence the Distribution of Solutes Between Blood and Brain. *J. Pharm. Sci.* **1994**, *83*, 1257−1268.

(10) Van de Waterbeemd, H.; Camenish, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. Estimation of Blood-Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H-bonding descriptors. *J. Drugs Target.* **1998**, *6*, 2, 151−165.

(11) Lombardo, F.; Blake, J. F.; Curatolo, W. Computation of Brain-Blood Partitioning of Organic Solutes via Free-Energy Calculations. *J. Med. Chem.* **1996**, *39*, 4750−4755.

(12) Keseru, G. M.; Molnar, L. High-Througput Prediction of Blood-Brain Partitioning: A Thermodynamic Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 120−128.

(13) Norinder, U.; Sjöberg, P.; Osterberg, T. Theoritical Calculations and Prediction of brain-blood partitionning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.* **1998**, *88*, 815−821.

(14) Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting Blood-Brain Permeation from Three-Dimensional Molecular Structure. *J. Med. Chem.* **2000**, *43*, 2204−2216.

(15) Osterberg, T.; Norinder, U. Prediction of Polar Surface Area and Drug Transport Processes using Simple Parameters and PLS Statistics. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1408−1411.

(16) Rose, K.; Hall, L. H.; Kier, L. B. Modeling Blood_Brain Barrier Partitioning Using the Electrotopological State. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651−666.

(17) Ajay, Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(18) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood_Brain Transport of drugs: A Computational approach. *Pharm. Res.* **1996**, *13*, 5, 775−778.

(19) Comprehensive Medicinal Chemistry Release 94.1 available from MDL Information Systems Inc., San Leandro, CA 94577.

(20) MACCS-II Drug Data report available from MDL Information Systems Inc., San Leandro, CA 94577.

(21) WDI, Daylight Chemical Information Systems, Inc., 27401 Los Altos, #360, Mission Viejo, CA 92691, U.S.A. E-mail: info@daylight.com.

(22) *Guidelines for ATC classification and DDD assignment*. WHO Collaborating Centre for Drug Statistics Methodology: Oslo, Norway 2001; ISBN 82 90312 32 6 (www.whocc.no/atcddd/).

(23) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(24) Seelig, A. A General Pattern for Substrate Recognition by P-Glycoprotein, *Eur. J. Biochem.* **1998**, *251*, 252−261.

(25) Accelrys, 9685 Scranton Road, San Diego, California, U.S.A. http://www.accelrys.com.

(26) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1−3), 3−25.
(27) Lahana, R.; Grassy, G. In QSAR: Quantitative Structure−Activity Relationships in Drug Design: Proceedings of the 7th European Symposium on QSAR; Interlaken, Switzerland, 5−9 September 1988, Fauchère, J. L. Ed.; Liss: New York.
(28) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*(9), 1737−1740.
(29) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The Importance of a Nitrogen Atom in Modulators of Multidrug Resistance. *Mol. Pharmacol.* **1999**, *56*, 791−796.
(30) Osterberg, T.; Norinder, U. Theoritical Calculation and Prediction of P-Glycoprotein-Interacting Drugs using MolSurf Parametrization and PLS Statistics. *Eur. J. Pharm. Sci.* **2000**, *10*, 295−303.
(31) Litman, T.; Zuthen, T.; Skovsgaard, T.; Stein, W. Structure−Activity Relationships of P-Glycoprotein Interacting Drugs: Kinetic Characterization of their Effects on ATP-ase Activity. *Biochim. Biophys. Acta* **1997**, *1361*, 159−168.
(32) Stouch, T. R.; Gudmundsson, O. Progress in Understanding the Structure−Activity Relationships of P-Glycoprotein. *Adv. Drug Delivery Rev.* **2002**, *54*(3), 315−328.
(33) Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*(27), 5382−5392.
(34) Norinder, U.; Haeberlein, M. Computational approaches to the prediction of blood-brain distribution. *Adv. Drug Delivery Rev.* **2002**, *54*(3), 291−313.

CI034205D