Journal of Computer-Aided Molecular Design, **17:** 65–76, 2003.
KLUWER/ESCOM

Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA

Philippa R.N. Wolohan* & Robert D. Clark
Tripos, Inc., 1699 South Hanley Road, Saint Louis, Missouri 63144, USA

Received 20 August 2002; Accepted in final form 30 March 2003

Key words: ADME prediction, drug pharmacokinetics, idiotropic field orientation, molecular alignment, molecular interaction fields, SIMCA

Summary

We have developed a method that combines molecular interaction fields with soft independent modeling of class analogy (SIMCA) [1] to predict pharmacokinetic drug properties. Several additional considerations to those made in traditional QSAR are required in order to develop a successful QSPR strategy that is capable of accommodating the many complex factors that contribute to key pharmacokinetic properties such as ADME (absorption, distribution, metabolism, and excretion) and toxicology. An accurate prediction of oral bioavailability, for example, requires that absorption and first-pass hepatic elimination both be taken into consideration. To accomplish this, general properties of molecules must be related to their solubility and ability to penetrate biological membranes, and specific features must be related to their particular metabolic and toxicological profiles. Here we describe a method, which is applicable to structurally diverse data sets while utilizing as much detailed structural information as possible. We address the issue of the molecular alignment of a structurally diverse set of compounds using idiotropic field orientation (IFO), a generalization of inertial field orientation [2]. We have developed a second flavor of this method, which directly incorporates electrostatics into the molecular alignment. Both variations of IFO produce a characteristic orientation for each structure and the corresponding molecular fields can then be analyzed using SIMCA. Models are presented for human intestinal absorption, blood-brain barrier penetration and bioavailability to demonstrate ways in which this tool can be used early in the drug development process to identify leads likely to exhibit poor pharmacokinetic behavior in pre-clinical studies, and we have explored the influence of conformation and molecular field type on the statistical properties of the models obtained.

Introduction

Investment in combinatorial chemistry and high throughput screening methods has significantly increased the rate at which compounds with desired potency against a biological target can be identified. Unfortunately, the return on investment in terms of bringing new drugs to market has not increased proportionately, so that introduction of new chemical entities (NCEs) has not kept pace with expectations [3]. Most new leads that fail in pre-clinical evaluation do so because they do not exhibit the desired pharmacological activity; because they exhibit poor pharmacokinetic profiles with regard to absorption, distribution, metabolism, or excretion (ADME); or because they are toxic.

Full pharmacological characterization of every lead would be prohibitively labor intensive and time-consuming, so new methods are being sought to alleviate *in vivo* evaluation bottlenecks and decrease the attrition rate between lead optimization and the clinic. Several *in vitro* screening techniques for predicting pharmacokinetic properties exist, the most widely used being solubility assays, hepatocyte metabolism and transport across Caco-2 monolayers [4]. Most are more or less imperfect surrogates for *in vivo* evaluation however, and a whole battery of such tests would be required to adequately identify all possible

*To whom correspondence should be addressed. E-mail: pwolohan@tripos.com

pharmacokinetic problems. Hence, a combination of *in silico* methods with selected *in vitro* tests can provide a cost-effective complement to exhaustive *in vitro* screening.

Past successes in the application of computational approaches in lead identification and optimization have fueled anticipation that it may provide useful tools for predicting pharmacokinetic properties as well. Several computational methods have been reported to date, among which Lipinski's Rule of Five analysis [5] is probably the most widely used for identifying compounds unlikely to exhibit good solubility and permeability. Others have reported finding direct relationships between Polar Surface Area and passive absorption or permeability [6, 7, 8, 9, 10, 11].

Historically, 2D QSAR descriptors have been most useful when applied within congeneric series for prediction of specific biochemical properties. A number of recent studies aimed at classifying potential lead compounds as being more or less drug-like in character reported considerable success at predicting general pharmacokinetic properties using descriptors drawn from classical 2D QSAR [12, 13, 14]. Human intestinal absorption (HIA%), blood-brain barrier permeability (BBB) and human oral bioavailability have been of special interest.

3D-QSAR methods directly relate the relative placement in space of more or less generalized molecular features to biological activity. One such method is Comparative Molecular Field Analysis (CoMFA), developed by Cramer *et al.* [15, 16, 17]. The CoMFA method uses the steric and electrostatic fields surrounding ligands in conjunction with partial least squares projection onto latent structures (PLS) to relate variations in such fields to biological activities. The methodology has subsequently been extended to include hydrophobic and other fields, e.g., in the CoMSIA approach of Klebe et al. [18]. Recent publications have used CoMFA to predict ADME-Tox properties within some congeneric series [19, 20]. Segarra et al. used GRID to characterize surface interaction potentials and relate those to permeability within pharmacological classes [21]. Cruciani et al., in contrast, integrate field contours to extract information from molecular fields obtained for a range of GRID probes to generate the VolSurf descriptors, which have been usefully applied across structural series to predict pharmacokinetic properties [22, 23].

Because VolSurf parameters reflect relationships between contours for different field types within individual molecules, there is no need to align molecules of interest to one another. But this approach entails the implicit assumption that sufficient information relevant to the prediction of a specific property can be obtained from the molecular fields of the molecules in the form of isotropic descriptors. Here we have adopted a complementary approach, in which each molecule is embedded in a coordinate space deduced from its overall steric and electrostatic properties. Molecular fields obtained from steric idiotropic field orientations (IFO-CoMFA), used in conjunction with hierarchical clustering, have proven useful in sorting drugs into pharmacological classes [2].

In the present application, these fields are employed as input to soft independent modeling of class analogies (SIMCA) [1] to analyze pharmacokinetic properties. Even though the input ADME properties are tabulated as continuous variables, they are better suited to categorical analysis. This is in part because of the relatively large uncertainties in their values and in part because the decision of interest usually entails a categorical judgement made with regard to some critical threshold. The difference between 10% uptake and 90% uptake is relevant, for example, whereas the differences between 0.1 and 1% (or between 90 and 99%) usually is not.

Materials and methods

Data sets and structures

The human intestinal absorption (HIA) data set was compiled and initially analyzed by Wessel et al. [24] and has been discussed by Oprea et al., [25] among others. This data set includes 86 drug and drug-like molecules together with their experimental HIA% values gathered from literature sources. Human bioavailability data (BIO_1) for 198 compounds were drawn from those tabulated by Yoshida and Topliss [26]. A further data set (BIO_2) of 408 compounds was taken from a more extensive compilation by Sietsema [27].

The blood brain barrier (BBB) data set consists of 55 drug molecules together with the corresponding experimentally determined blood-brain partition function (logBB) values. This data set was originally compiled by Abraham et al. [28]. Experimental values of logBB range from -2.00 to $+1.00$. Within this range, compounds with logBB greater than 0.3 cross the blood brain barrier readily, while compounds with logBB less than -1.0 are only poorly distributed in the brain. Lombardo et al. [29] used free energy calculations to predict logBB by regression ($r^2 = 0.67$,

Table 1. The class ranges for each data set obtained by applying agglomerative hierarchical clustering.

Data Set[a]	HIA		BBB		BIO_1		BIO_2	
Category	Range (%)	Total no. of drug molecules (n)	LogBB	Total no. of drug molecules (n)	Range (%)	Total no. of drug molecules (n)	Range (%)	Total no. of drug molecules (n)
1	0–15	10	< −0.5	17	0–15	38	0–15	64
2	25–36	5	−0.5 to <0	15	25–36	43	25–36	67
3	45–76	14	0 to <0.5	13	45–76	52	45–76	131
4	80–100	57	0.5 to 1	10	80–100	65	80–100	146

[a] HIA[ref.23], BBB[ref.26], BIO_1[ref.25], BIO_2[ref.30].

Table 2. A summary of the models built for the HIA data set. The columns represent the number of drugs correctly classified by SIMCA for each model.

		Molecular interaction field types used in the model[a]																
Cat.	Total	S	E	S E	S_E	CS	CE	CS CE	HPH	HBD	HBA	HBD HBA	HPH HBD HBA	E HPH HBA	**S_E HPH**	CS CE HPH HBD HBA	S HPH	S HPH HBA
1	10	5	0	0	4	7	3	2	8	4	8	8	7	8	**8**	7	9	9
2	5	3	0	0	4	3	4	0	3	4	4	4	4	4	**4**	4	4	4
3	14	4	2	10	7	8	1	0	6	10	13	10	9	12	**14**	10	12	12
4	57	56	57	57	53	50	56	56	52	52	46	55	54	54	**52**	57	53	52
Total	86	68 (79%)	59 (69%)	67 (78%)	68 (79%)	68 (79%)	64 (74%)	58 (67%)	69 (80%)	70 (81%)	71 (83%)	77 (90%)	74 (86%)	78 (91%)	**78 (91%)**	78 (91%)	78 (91%)	77 (90%)

[a]Key: S = CoMFA Steric, E = CoMFA Electrostatic, S_E = Joint CoMFA, CS = CoMSIA Steric, CE = CoMSIA Electrostatic, HPH = CoMSIA Hydrophobic, HBD = CoMSIA Hydrogen Bond Donor, HBA = CoMSIA Hydrogen Bond Acceptor

SE = 0.41). It has subsequently been analyzed by several other groups, including Clark et al. [11] who used polar surface area and logP ($r^2 = 0.79$, SE = 0.354). Jørgensen et al. [30] recently published a method for predicting blood brain barrier penetration for these compounds based on the atom-type weighted water-accessible surface area ($r^2 = 0.84$).

The target properties of interest were grouped into categories within each data set by applying agglomerative hierarchical clustering as implemented in the QSAR module of SYBYL® 6.8.1. [31]. The complete linkage method was used in all cases. As indicated by the class ranges listed in Table 1, most of the data sets split into reasonably 'natural', cleanly separated classes using this technique. Note, too, that the thresholds obtained are consistent with the expected uncertainty in the underlying data (see, for example, Sietsema et al. [27]).

All 2D chemical structures and chirality information used were obtained from the Merck Index [32] and converted into 3D structures on a Silicon Graphics, Inc (SGI) workstation using the SYBYL sketcher and CONCORD 4.0. [33] Atom types for the nitrogen and oxygen atoms in nitro groups were set to N.ar and O.co2, respectively, and the N–O bond types were set to aromatic. Each structure was then relaxed using the Tripos molecular mechanics force field [34]. Atomic partial charges were calculated using the GAST_HUCK option in SYBYL, which invokes an extension of the method described by Gasteiger and Marsili [35] for estimating the distribution of charge over sigma bond networks.

Molecular orientation

Passive absorption, permeability, and related properties are generally not dependent on local, anisotropic attributes, whereas active uptake and metabolism can be very sensitive to the relative distribution of more or less specific structural features in space. For active uptake and metabolism to occur specific carrier or enzymatic molecular recognition is required hence these processes are more sensitive to pharmacophoric

Table 3. The models with the highest number of drugs correctly classified for each data set using steric IFO and electrostatic IFO.

Molecular alignment			Steric IFO						Electrostatic IFO					
Data set[a]	Category	n	MIF[b]	Correct	Correct ± 1	Correct ± 2	# False positive	# False negative	MIF	Correct	Correct ± 1	Correct ± 2	# False positive	# False negative
HIA	1	10	S_E	8	8	9	2		S	10	10	10		
	2	5	&	4	4	5	1			5	5	5		
	3	14	HPH	14	14	14				12	13	14	1	1
	4	57		52	57	57		5		53	55	55		4
	Total	86		78	83	85	3	5		80	83	84	1	5
				(91%)	(97%)	(99%)	(3%)	(6%)		(93%)	(97%)	(98%)	(1%)	(6%)
BBB	1	17	S_E	15	17	17	2		S	17	17	17		
	2	15		11	14	15	3			1	13	13	15	2
	3	13		12	13	13	1			11	13	13	2	
	4	10		10	10	10				10	10	10		
	Total	55		48	54	55	6	1		51	53	55	4	0
				(87%)	(98%)	(100%)	(11%)	(2%)		(93%)	(96%)	(100%)	(7%)	
BIO_1	1	38	S_E	36	36	38	2		S_E	34	35	36	4	
	2	43		6	38	43	23	14		37	41	43	5	1
	3	52		46	48	52	2	4		44	50	52	2	6
	4	65		35	47	47		30		46	56	59		19
	Total	198		123	169	180	27	48	161		182	190	11	26
				(62%)	(85%)	(91%)	(14%)	(24%)	(81%)		(92%)	(96%)	(6%)	(13%)
BIO_2	1	64	S_E	19	34	46	45		S_E	35	40	46	29	
	2	67		43	63	69	19	5		51	64	69	14	2
	3	131		84	128	129	12	35		80	124	129	31	20
	4	146		92	128	145		54		104	125	141		42
	Total	408		238	353	389	76	94		270	353	385	74	64
				(58%)	(87%)	(95%)	(19%)	(23%)		(66%)	(87%)	(94%)	(18%)	(16%)

[a]HIA[ref.23], BBB[ref.26], BIO_1[ref.25], BIO_2[ref.30]. [b] Molecular Interaction Field Type used as descriptor for the model S_E = Joint CoMFA steric and electrostatic fields, HPH = CoMSIA hydrophobic field, S = CoMFA steric field.

features, while passive absorption and permeability are more dependent on physical properties such as molecular volume and weight. Hence it is desirable to relate both the 'bulk' properties of the molecules that contribute towards solubility and penetration, and also the specific structural features, to metabolic and toxicological profiles. The data sets considered here encompass such a wide range of structural diversity, however, that aligning structures to one another using atom- or pharmacophore-based alignments of the sort ordinarily employed in CoMFA [15, 36] is impossible. In idiotropic field orientation (IFO) the principal axes of each individual structure determines the coordinate system with respect to which, the corresponding molecular field is evaluated [37]. For steric IFO, the unweighted molecular graph is used, so as to minimize the effects of various substitutions that are generally small in terms of their orientational effect on medicinal chemistry – e.g., replacement of hydrogen by fluorine [2]. The *x*-axis is defined to coincide with the smallest moment of inertia; the *y*-axis coincides with the smallest moment perpendicular to the *x*-axis, and the *z*-axis is fixed as perpendicular to the *xy*-plane. For electrostatic IFO, each vertex in the molecular graph is weighted by the absolute value of the partial atomic charge calculated for the corresponding atom.

Principal axes do not have any intrinsic directionality, so there is a four-fold degeneracy when they are used to define a coordinate system. Here, we employed the methodology available in the Selector® module [31] of SYBYL to relieve this degeneracy, assigning the polarity of the *x* and *y* axes so as to make the dipole moment along both axes positive [38]. This is the route by which the 'bulk' electrostatic properties of the molecule contribute to the field orientation and, hence, overall similarity between fields. Since overall steric bulk and polarity are both likely to play an important role in passive diffusion, this

approach helps ensure that similar molecules will be similarly oriented, thereby enhancing the discrimination power of models constructed using the molecular fields as input data.

Molecular field evaluation

Field regions were defined for each data set so as to extend 5 Å along each coordinate axis beyond union of the van der Waals volumes of the assemblage of all molecules. Grid spacing was set to 2 Å along each axis in all cases.

For CoMFA, an sp^3-hybridized carbon atom bearing a formal charge of +1 served as the probe.

The steric field reflects van der Waals (dispersion-like) interactions, while the Coulombic term represents the electrostatic interactions. A distance dependent dielectric expression $\varepsilon = \varepsilon_0 R_{ij}$ was used, where R_{ij} is the distance between two charges q_i and q_j, and $\varepsilon_0 = 1.0$. Field values were truncated at 30 kcal/mol for the steric field energies and at ± 30.0 kcal/mol for the electrostatic field energies. By default, the steric and electrostatic fields are considered jointly, suppressing electrostatics at lattice points where the steric cutoff is exceeded. These electrostatic energy values fall inside a molecule and so generally do not contribute to intermolecular interactions. Such suppression is lost when the two fields are generated as separate columns in a SYBYL molecular spreadsheet (MSS). Both contingent and independent steric and electrostatic fields were considered here.

Steric and electrostatic fields of the type used in Comparative Molecular Similarity Index Analysis (CoMSIA) [18, 39] were also considered. In CoMSIA, Gaussian functions of the form $f(r) = e^{-\alpha r^2}$ replace the Lennard-Jones and Coulombic potentials used in CoMFA. In this case, electrostatic potentials are not suppressed at points within a molecule's steric envelope, and cutoff values are not used. The Gaussian functions are 'softer' and place relatively more weight on interactions close to the molecular surface. These two CoMFA-like field types are augmented by hydrophobic fields and by hydrogen bond donor and acceptor fields.

We considered the three types of CoMFA field (sterics and electrostatics calculated separately or jointly) and the five 'flavors' of CoMSIA field (steric, electrostatic, hydrophobic, donor and acceptor) alone and in combination to see which could best be used in conjunction with both types of IFO to categorize structural properties relevant to drug absorption. Different 'flavors' of steric field were not considered in combination, nor were the different flavors of electrostatic field, because of concern that introducing that much descriptor redundancy might distort our analyses.

Analysis

All three of the data sets considered here involve nominally continuous response variables – percentage intestinal absorption, logBB or bioavailability – but the errors in determination are typically large relative to the possible range of response values, so they are really more qualitative than quantitative in nature. Moreover, discrimination beyond the categorical level (e.g., good, okay, fair, or poor) is not really relevant for ADME properties in the context of drug development. SIMCA is a good choice for modeling such categorical data, especially when a large number of intercorrelated descriptors are involved. In this approach, principal components analysis (PCA) is used to identify the number of underlying independent latent variables characterizing each class and to extract the linear combination of the original variables corresponding to each. Each observation is then projected onto the hyperplane defined by those latent variables. The projected values, together with the residual variance in the descriptor space, defines a region in the original descriptor space into which additional observations from that class are expected to fall. Prediction consists of calculating the distance to the hyperplane for each category; if the projection falls significantly outside the range delimited by the projections for the training set, the distance to the boundary defined thereby is used instead.

As has been noted by Hunt, [40] the implementation of SIMCA in earlier releases of the SYBYL QSAR module did not conform fully to the method originally described [1] by Wold. In particular, the hyperplanes obtained were not bounded within classes. The implementation found in SYBYL 6.8.1, which was used in the present work, includes such bounding and provides scaling by residual variance for the 'distance' to each category. The 'zeroth component' behavior [40] found when the original descriptors are not significantly correlated – i.e., when only the centroid of a category is meaningful – is logically consistent and so has been retained despite its absence from Wold's original description. That behavior is not relevant to the work described here, because all categories examined produced at least one significant component.

Table 4. An example of test set results for alignment methods applied to the HIA data set.

Training set	Reduced HIA model (S_IFO)		Reduced HIA Model (E_IFO)		Test set drug	Literature category	Predicted category (S_IFO)	Predicted category (E_IFO)
Category	Total	Correct	Total	Correct	Acrivastine	4	3	4
1	8	8	10	8	Bumetanide	4	4	4
2	3	1	4	3	Etoposide	3	4	1
3	13	13	11	9	Norfloxacin	2	3	4
4	45	45	49	46	Predenisoline	4	3	4
Total	69	67	74	66	Valproic acid	4	3	4

Block scaling between fields ('CoMFA Standard') was used throughout. A column filter of 2.0 kcal/mol was employed to drop field values at lattice points for which the variance across the data set is too small, thereby improving computational efficiency and reducing noise. The maximum number of principal components allowed was capped at six to avoid overfitting.

Results

HIA data set

The HIA data set is heavily biased towards drugs with high oral absorption, so it is not surprising that clustering results in uneven ranges for each class (Table 1). A variety of factors can affect the assessment of our experimental properties such as biological factors, errors in reports and imprecise clinical or analytical methods. Therefore, we felt it was within reason to categorize the accuracy of our predictions within ranges of approximately 25% HIA.

To assess the value of different molecular fields at modeling the experimental properties of our data sets each of the different CoMFA and CoMSIA fields were computed using a single region file. SIMCA analyses were carried out using each of the CoMFA steric, electrostatic and the joint steric and electrostatic fields in turn as the descriptor column (Table 2). Separate analyses were also carried out using the CoMSIA steric, electrostatic, hydrophobic, hydrogen-bond donor and hydrogen bond acceptor fields and using combinations of the fields as the descriptors. Table 2 shows the results obtained for SIMCA models based on the individual fields and on combinations of the fields that produced the best models. This procedure was first carried out with the data set aligned using steric IFO and then repeated using electrostatic IFO. Using steric IFO, four different combinations of molecular field types correctly classified 91% of the drugs in the data set. Interestingly, all four of these models included the CoMSIA hydrophobic field as a descriptor. The model that consisted of a combination of the hydrophobic field with the joint CoMFA steric and electrostatic fields was selected for further evaluation because it does a better job of modeling the more poorly absorbed compounds, which are typically more difficult to predict correctly than are highly absorbed compounds. For this model, SIMCA classifies 78 drugs correctly, with five more falling to within one category of the correct classification (Table 3).

To assess the quality of these models it is useful to compare the results with a random class assignment case. In a four-category model the probability of randomly assigning a drug to its correct experimental class is 25%. Therefore, our model classifies the compounds by a factor of 3.6 over random, or if we average the percentage correct within categories then the value obtained (88%) represents an enhancement of 3.5. If we are willing to except a classification of within $\pm$ 1 of the experimental category as acceptable, [13] the random probability of a correct assignment increases to 62.5% and for within $\pm$ 2 to 87.5%.

Another assessment we used to measure the quality of the models was the false positive / false negative distribution. Keeping the false negative rate low is generally more important than is minimizing the false positive rate, at least in early evaluation phases of lead selection. Only five drugs were false negatives, i.e., the model incorrectly places them in poorly absorbed categories. The CoMFA steric field alone proved to be the best descriptor for the data set aligned using the electrostatic IFO method.

The distance of each molecule to the hyperplane for each category can be presented in a Cooman plot.

This gives a graphical indication of the discrimination of the model between the classes and can also be used to identify outliers in individual categories. Zidovudine was identified as an outlier in category four. As we describe in this paper, our observation that a compound with a known active transport mechanism had the greatest distance of any one compound from its category provided us with a means to identify other such actively transported compounds. However, we should note here that because the charge distribution within Zidovudine's azide group is poorly modeled the quality of this model might be compromised.

The effect of varying the maximum number of components allowed was examined using the model selected above. In some cases this did improve the fit of the model. In general, however, the number of principal components was kept at or below 6, which is to the high end of the acceptable range of PCs for a typical PLS model. However, with SIMCA models, tightly fitting data within a category can help to better discriminate between categories that have a 'fuzzy' separation and thereby increase the predictive power of the model.

Model validation

The principal components for each category are limited in length and therefore define a hypervolume in which each category member lies. Category membership for the test point then depends upon whether the point falls within one of these defined volumes.

Therefore, another experiment was run to see if any of the missclassified compounds were contributing significantly to correctly predicting other compounds or were simply adding noise to the model. A reduced model was built using only those compounds that were predicted in the correct class, which left us with a data set consisting of 78 compounds. This model was used to predict the 8 molecules that were missclassified. The omitted compounds were predicted in exactly the same (incorrect) category as previously reported. In addition, the reduced model incorrectly predicted 3 of the 78 compounds. A third model of the remaining 75 compounds successfully classified 3 out of the 11 omitted compounds. One of these 3 successfully classified compounds belong to the set of 8 molecules that was missclassified in the full model and the other two belong to the 3 that were incorrectly predicted in the first reduced model. This finding that a further reduced model successfully classifies some compounds that were previously classified incorrectly offers evidence that our SIMCA model is not guilty of over-fitting the data.

Four test sets were selected from the HIA data with the remaining drugs used as the training set in each case. For the four test sets an average of 89% of drugs were predicted within one category of their correct absorption class. An example of the test set results for each of the alignment methods is presented in Table 4.

Sensitivity to conformation

We also explored the sensitivity of the idiotropic alignment to molecular conformation. The program Confort [41] was used to generate up to ten diverse conformers of each drug. All the rotatable bonds were searched up to a maximum of 30, which generated a total of 630 conformers for the HIA data set. Confort uses the Tripos Force Field to calculate molecular mechanics energies for the conformers. A SIMCA model based on the lowest energy conformer of each drug only classified 64 molecules correctly (Table 5a). Another model was created by taking the highest class predicted for any single conformer as the prediction for each drug. Here we made the assumption that if at least one conformer of a particular drug is well absorbed, then the drug in question will be well absorbed. The model performed better when more conformers were generated: when 1–5 conformers were generated 58% were classified correctly and when 6–10 conformers were generated 69% were correct (Table 5b). We conclude that a conformational analysis of the molecules provided no benefit over using the 3D-structures built using CONCORD rules. CONCORD is a rule-based system for generating low-energy conformers, so it produces characteristic conformations that are generally similar for similar compounds. This reduces the sensitivity of the method to conformational variation. We employed Confort for conformer generation specifically since it provides a systematic method for conformer generation, however we acknowledge that the use of a different conformational analysis tool [42] may have resulted in a different conclusion.

Charged vs. neutral species

For each of the models described above, a new model was built in which each molecule in the data set was protonated or deprotonated based on their literature pK_a values. A model built for the modified structures using the CoMFA steric field as the descriptor column (ca. Table 6) predicted 63 drugs correctly, compared

Table 5a. Conformational sensitivity of the method for classification of the drugs in the HIA data set.

Category	All conformers		Lowest energy conformer	
Category	Total	Correct	Total	Correct
1	92	50	10	2
2	50	27	5	2
3	95	35	14	4
4	453	373	57	56
Total	690	485 (70%)	86	64 (74%)

Table 5b. Variation in model performance with number of conformers generated.

# Conformers generated	Total	Correct	Correct ± 1	Flexible vs. rigid
1	12	7	9	
2	6	3	3	
3				11 58%
4	1	1	1	
5				
6				
7	1	1	1	
8	2	2	2	46 (69%)
9	1			
10	63	43	52	
Total	86	57 (66%)	68 (79%)	

to 68 for the model in which all the molecules were evaluated in their neutral state. The average percentage correct within categories dropped to only 36%. Also, none of the molecules in category 1 or 2 (below 50% HIA) were classified correctly by the charged model. No molecules were predicted to belong to category 1, one molecule to category 2 and eight to category 3; the majority were predicted as well absorbed. The poor performance of the models in which charge is accounted for can be attributed to the fact that counter ions or solvent molecules will normally shield charged species. Therefore, *in vacuo*, a neutral representation of the molecules may be closer to the structural form of the drugs that is actually absorbed.

Active transport

The experimental absorption values that Wessel et al. gathered from the literature include several drugs for which an active transport mechanism has been identified [8]. We did not exclude the compounds known to be actively transported from our model because our aim is to develop a model that would implicitly account for such effects and be relatively insensitive to noise introduced by such compounds. Prior to biological testing it would be difficult to exclude such molecules. We found evidence in the literature that the following drugs are actively transported: amoxcillin, cephalexin, cefatrizine, cefuroxime axetil, captopril, enalapril, lisinopril, methotrexate, salicylic acid and zidovudine while, doxorubicin, gentamycin and etopside are known efflux substrates [8, 31]. Remarkably the SIMCA models for passive drug absorption correctly classified all of these compounds. However, in the Cooman plot several of the above compounds displayed the greatest distance of any one compound from their category. This presented us with an opportunity to develop a method of identifying actively transported compounds. The scores for the first three principal components for each category were clustered using hierarchical cluster analysis. Scatter graphs of the first thee components for each category are shown in Figure 1. In Figure 1 the colors are based on a hierarchical clustering of the scores. These graphs can be used to identify other drug molecules close to the known actively transported molecules.

Note that active transport does not exclude the possibility of passive absorption. This may in part account for the fact that our model correctly classifies the actively transported compounds. This possibility was tested by excluding compounds with a molecular weight greater than 500 under the assumption that large molecules are the least likely to be passively absorbed. This removed cefuroxime axetil, doxorubicin and etoposide from the model, which improved the model statistics whereas removing all of the drugs with mechanisms of absorption other than passive resulted in a model that was poorly predictive. Other models were also investigated, including creation of separate classes for different transport mechanisms; varying the number of categories considered and exclusion of different classes of drugs. None of these produced any significant improvement. In most of these cases however, the number of representative molecules in some classes was too low to produce a useful model anyway.

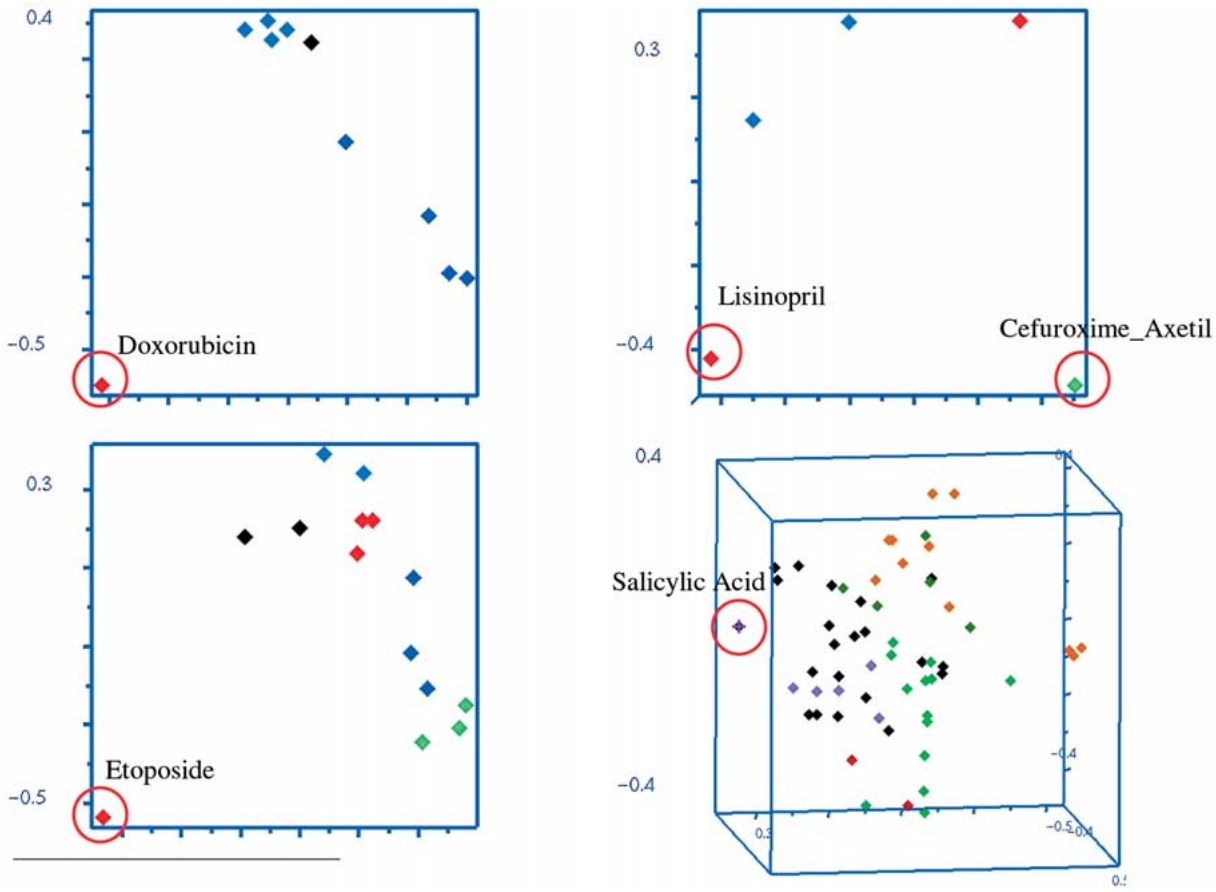

Figure 1. Scatter graphs of the first components for each category. Colors are assigned on hierarchical clustering of the scores: Top left, Category 1; Top right, Category 2; Bottom left, Category 3; Bottom right, Category 4.

Table 6. The HIA data set was protonated or deprotonated according to literature pK_a values. Results are for a SIMCA model of the charged data set using the CoMFA steric field as the descriptor and steric IFO alignment.

Category	Full model			Acids			Bases			Neutral		
	Total	Correct	Correct ± 1	Total	Correct	Correct ± 1	Total	Correct	Correct ± 1	Total	Correct	Correct ± 1
1	10	0	1	6	0	0	1	0	0	3	0	1
2	5	0	0	3	0	0	1	0	0	1	0	0
3	14	6	14	5	3	5	4	1	4	5	2	5
4	57	57	57	14	14	14	22	22	22	21	21	21
Total	86	63 (73%)	72 (84%)	28	17	19	28	23	26	30	23	27

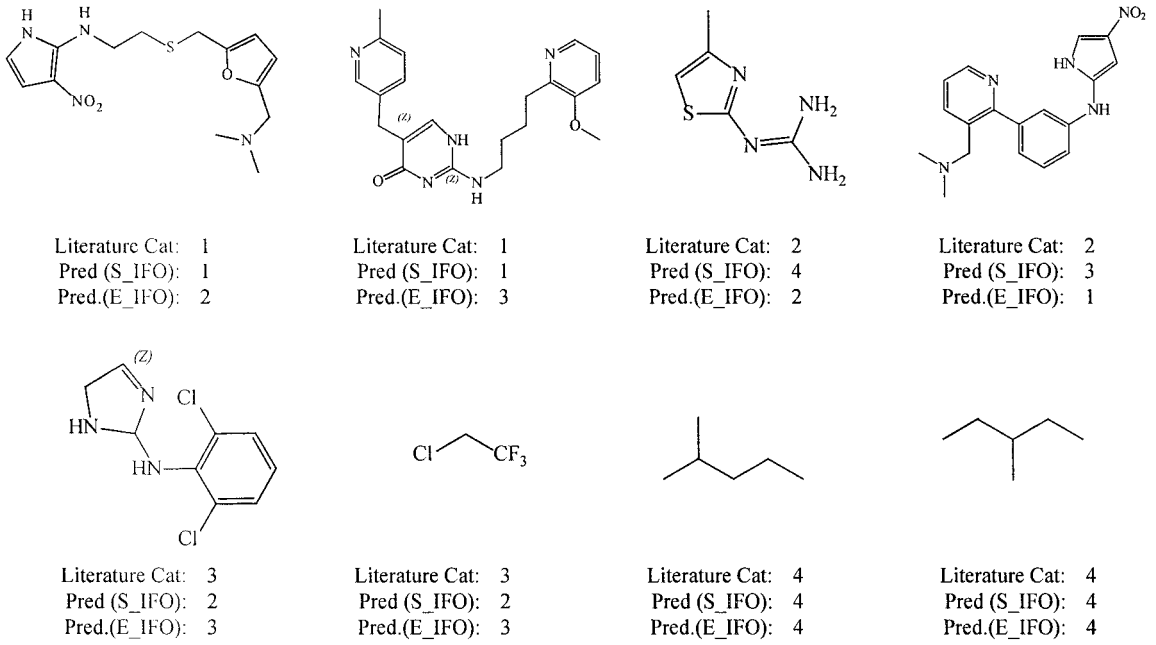

Figure 2. Test set results for alignment methods applied to the BBB data set. (The training sets for these predictions are presented in Table 5.)

Blood – brain barrier permeation

In the case of the Blood Brain Barrier data set (Table 1) the joint CoMFA steric and electrostatic fields using the steric IFO alignment produced the best model. This model placed 48 out 55 drugs in the correct category and only one compound was a false negative. 98% of the compounds fell within ± 1 category of the experiment values (Table 3). Four test sets were randomly selected from the BBB data with the remaining drugs used as the training set in each case. For the four test sets an average of 81% of drugs were predicted within one category of their correct absorption class. An example of the test set results for each of the alignment methods is presented in Figure 2 and the corresponding reduced models in Table 7.

Table 7. Training set for alignment methods applied to the BBB data set.

Training set	Reduced BBB model (S_IFO)		Reduced BBB model (E_IFO)	
Category	Total	Correct	Total	Correct
1	15	15	15	15
2	13	6	13	7
3	11	11	11	10
4	8	8	8	8
Total	47	40	47	40

Human bioavailability

The models for two bioavailability data sets aligned using steric IFO produced fairly similar models. BIO_1 model classified 62% of the compounds correctly (Table 3) and the Sietsema data set (BIO_2), which contains twice as many drugs, classified 58% of the compounds correctly. When the data sets were aligned using electrostatic IFO the models improved significantly: BIO_1 (81% correct), BIO_2 (66% correct). The reason we chose to build separate models for the two data sets was to highlight that having a larger data set does not guarantee that quality of the data is equivalent, especially for these types of data sets which comprise data compiled from a wide variety of literature sources. The quality of the data is difficult to assess quantitatively but it seems reasonable to suspect that assembling data from a higher number of sources will increase the variation in the data quality in an area where raw data are scarce.

When the data sets are oriented using electrostatics, the steric field alone produces the best model for the HIA model and the BBB model. For the bioavailability data sets, on the other hand the joint CoMFA steric and electrostatic fields produced the best model. From this observation we conclude that the electro-

static IFO alignment captures the coarse features of a drug molecules electrostatic properties well enough that the steric field alone is adequate for modeling HIA and BBB permeation. In contrast, the fine details of drug molecules' electrostatic features need to be captured for modeling the bioavailability data. Including the joint CoMFA steric and electrostatic interaction energies allows us to create an anisotropic model of these properties.

Not surprisingly BIO_1 and BIO_2 models correctly predicted fewer drugs. Predictions made for a test set using a reduced BIO_1 model correctly classified 8 out of 20 drugs and a reduced BIO_2 model only correctly assigned 9 out of 41 test set drugs. Again, the performance of the models improves considerably if we consider a classification of within $\pm$ 1 of the experimental category as acceptable. The BIO_1 model correctly assigned 16 drugs to within $\pm$ 1 of the experimental category and the BIO_2 model assigned 31 drugs. However, considering the higher degree of complexity involved, the observed success of the bioavailability models obtained here is very encouraging. That it is possible for a drug molecule to be completely absorbed yet be entirely destroyed or removed by first-pass elimination or metabolism makes the task of modeling bioavailability much more complicated than that of modeling absorption or permeability.

Discussion

The majority of commercially successful drugs are administered orally because patients and physicians both favor this method of drug delivery. But this route of administration imposes a whole series of constraints on a compound which are, usually, only tangentially related to effectiveness at the ultimate biochemical target site. The extent of intestinal uptake of a drug is governed by many factors, including rates of disintegration and dissolution as well as solubility and uptake [43]. The first step in attaining high bioavailability, then, is to achieve good oral absorption. Once in the tissue a drug is considered absorbed. The American Pharmaceutical Association defines bioavailability as the rate and extent to which an active drug substance is absorbed and becomes available to the general circulation. For a drug to become 'bioavailable' it must reach the general circulation intact, passing through the GI tract tissue and the liver. First-pass metabolism or elimination in any of these tissues may destroy a portion of the drug that was absorbed and therefore reduce the drug's bioavailability. Once in the circulation, blood-brain barrier (BBB) penetration is essential for drugs targeted at the central nervous system (CNS). Hence molecules that are CNS-active should exhibit features that enable them to cross the BBB, although permeation of the BBB does not guarantee CNS-activity. For drugs targeted to other areas however, passage through the BBB may lead to unwanted side effects.

In this study we have developed and demonstrated novel computational approaches for the efficient and accurate prediction of ADME properties. The goal was to construct a simple and direct method involving the exploitation of molecular interaction fields to model the complex interactions that constitute the ADME/Tox characteristics of drug-like compounds. The high correct classification rate found suggests that this method can be a fast and reliable way to predict bioavailability. 3D-QSAR techniques such as CoMFA described in this paper provide a fast empirical method of analyzing and predicting drug action. Such computational models may be useful as *in-silico* pre-filters of lead compounds in a HTS environment and as a research tool for identifying and improving the pharmacokinetic profiles of drug candidates.

References

1. Wold, S., Sjöström, M. Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In: Chemometrics: Theory and Applications, Kowalski, B.R., (Ed.), ACS Symposium Series, 1977, 52, 243–282.
2. Clark, R.D., Ferguson, A.M., Cramer, R.D., Persp. Drug Discov. Design. (1998), 9/10/11, 213.
3. Smith, D.A., van der Waterbeemd, H., Curr. Opin. Chem. Biol., 3 (1999) 372.
4. Hilgers, A.R., Conradi, R.A., Burton, P.S., Pharm. Res., 7 (1990) 902.
5. Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney P.J., Adv. Drug Del. Rev., 23 (1997) 3.
6. Palm, K., Luthman, K., Ungell, A. Strandlund, G., Arturssan, P., J. Pharm. Sci., 85 (1996) 32.
7. Palm, K., Luthman, K., Ungell, A. Strandlund, G., Beigi, F., Lundahl, P., Arturssan, P., J. Med. Chem., 41 (1998) 5382.
8. Clark, D.E., J. Pharm. Sci., 88 (1999) 807.
9. Osterberg, T., Norinder, U., J. Chem. Inf. Comput. Sci., 40 (2000) 1408.
10. Ertl, P., Rohde, B., Selzer, P., J. Med. Chem., 43 (2000) 3714.
11. Clark, D.E., J. Pharm. Sci., 88, (1999) 815.
12. Oprea, T.I., Gottries, J., J. Mol. Graphics Modeling., 17 (1999) 261.
13. Yoshida, F., Topliss, J.G., J. Med. Chem., 43 (2000) 2575.
14. Lombardo, F., Blake, J.F., Curaolo, W.J., J. Med. Chem., 39 (1996) 4750.
15. Cramer, R.D., DePriest, S.A., Patterson, D.E., Hecht, P. The Developing Practice of Comparative Molecular Field

Analysis in 3D-QSAR in Drug Design: Theory, Methods and Applications. ESCOM, Leiden, 1993.
16. Cramer, R.D., Patterson, D.E., Bunce, J.D., J. Amer. Chem. Assoc., 110 (1998) 5959.
17. Martin, Y.C., Kim, K.H., Liu, C.T., Quant. Struc-Act. Relat., 1 (1996) 1.
18. Klebe, G., Persp. Drug Disc. Design., (1998) 87.
19. Ekins, S., Durst, G.L., Stratford, R.E., Thorner, D.A., Lewis, R., Loncharich, R.J., Wikel, J.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1578.
20. Ekins, S., Bravi, G., Ring, B.J., Gillespie, T.A., Gillespie, J.S., VandenBranden, M., Wrighton, S.A., Wikel, J.H., J. Pharmacol. Exp. Thera., 288 (1999) 21.
21. Segarra, V., López, M., Ryder, H., Palacios, J.M., Quant. Struct.-Act. Relat., 18 (1999) 474–481.
22. Cruciani, G., Carrupt, P.A., Testa, B., J. Mol. Structure: Theochem., 503 (2000) 17.
23. VolSurf is distributed by Tripos, Inc., 1699 S. Hanley Road, St. Louis Missouri, U.S.A.
24. Wessel, M.D., Jurs, P.C., Tolan, J.W., Muskal, S.M., J. Chem. Inf. Comput. Sci., 38 (1998), 726.
25. Oprea, T.I., Gottries, J., J. Mol. Graphics Modeling, 17 (1999) 261.
26. Yoshida, F., Topliss, J.G., J. Med. Chem., 43 (2000) 2575.
27. Sietsema, W.J., Intern. J. Clin. Pharm. Therapy Tox., 27 (1989) 179.
28. Abraham, M.H., Chadha, H.S., Mitchell, R.C., J. Pharm. Science., 83 (1994) 1257.
29. Lombardo, F., Blake, J.F., Curaolo, W.J., J. Med. Chem., 39 (1996) 4750.
30. Jørgensen, F.S., Jensen, L.H., Capion, D., Christensen, I.T. Prediction of Blood- Brain Barrier Penetration. In Rational Approaches to Drug Desig, Höltje, H.-D., and Sippl, W., (Eds.) Prous Science: Barcelona, 2001, pp. 281–285.
31. SYBYL® 6.8.1 Tripos Inc., 1699 S. Hanley Road, St. Louis, Missouri, 63144, U.S.A.
32. The Merck Index. 12^{th} Edition on CD-ROM, version 12:3 2000. Chapman & Hall / CRCnetBASE Electronic Publishing Division.
33. CONCORD was developed by R.S. Pearlman, A. Rusinko, J.M. Skell and R. Balducci at the University of Texas, Austin TX and is available exclusively from Tripos, Inc., 1699 S. Hanley Road, St. Louis Missouri, U.S.A.
34. Clark, M., Cramer, R.D. III., Van Opdenbosch, N., J. Comp. Chem., 10 (1989) 982.
35. Gasteiger, J., Marsili., Tetrahedron, 36, (1980) 3219.
36. Clark, R.D., Leonard, J.M., Strizhev, A. Pharmacophore Models and Comparative Molecular Field Analysis (CoMFA). In Pharmacophore Perception, Development, and Use in Drug Design, Güner, O.F., (Ed), International University Line: La Jolla, 2000, pp. 151–169.
37. IFO refers to the fact that the principal axes correspond to the series of mutually perpendicular axes that minimize the respective moments of inertia for the molecular graph.
38. A small negative dipole moment ($|\mu| < 0.01$) along the x or y axis is allowed if that orientation is necessary to get a large dipole moment along the z axis. This generally only comes into play for highly symmetrical molecules that are not likely to be good drug candidates anyway.
39. Klebe, G., Abraham, U., Mietzner, T., J. Med. Chem., 37 (1994) 4130.
40. Hunt, P.A., J. Comp-Aided Mol. Design., 13 (1999) 453.
41. Confort was developed by R.S. Pearlman and R. Balducci at the University of Texas, Austin TX and is distributed by Tripos, Inc., 1699 S. Hanley Road, St. Louis Missouri, U.S.A.
42. Boström, J.J., Comp-Aided Mol. Design., 15 (2001) 1137.
43. Lin, J.H., Lu, A.Y.H., Pharmacol. Rev., 49 (1997) 403.