Theoretical Calculation and Prediction of Caco-2 Cell Permeability Using MolSurf Parametrization and PLS Statistics

Ulf Norinder,[1,3] Thomas Österberg,[1] and Per Artursson[2]

Received June 24, 1997; accepted August 27, 1997

Purpose. To statistically model the permeability across Caco-2 cell monolayers using theoretically computed molecular descriptors and multivariate statistics.

Methods. Seventeen structurally diverse compounds were investigated. The program MolSurf was used to compute theoretical molecular descriptors related to physico-chemical properties such as lipophilicity, polarity, polarizability and hydrogen bonding. The multivariate Partial Least Squares Projections to Latent Structures (PLS) method was used to delineate the relationship between the permeability across Caco-2 cell monolayers and the theoretically computed molecular descriptors.

Results. Excellent statistical models were derived. Properties associated with hydrogen bonding had the largest impact on diffusion through the monolayers and should be kept at a minimum to promote high permeability. High lipophilicity and the presence of surface electrons, *i.e.* valence electrons, which are not tightly bonded to the molecule, were also found to have a favorable influence to achieve high permeability.

Conclusions. The results indicate that theoretically computed molecular MolSurf descriptors in conjunction with multivariate statistics of PLS type can be used to successfully model permeability across Caco-2 cell monolayers and, thus, differentiate drugs with poor permeability from those with acceptable permeability at an early stage of the preclinical drug discovery process.

KEY WORDS: MolSurf; PLS; Caco-2 cells; quantum mechanics.

INTRODUCTION

Drug discovery programs are generally dedicated to the development of orally active drugs, since this is the preferred route of administration and often an absolute requirement from a marketing point of view. The majority of conventional low molecular weight drugs are absorbed by passive diffusion from the gut. The extent of absorption is mainly dependent on dose, solubility/dissolution rate and membrane permeability.

Solubility/dissolution rate and permeability are of equal importance and an unfavorable value of one parameter may be compensated by a favorable value of the other. The present paper focus on a new computational method (MolSurf) which can be used for prediction of passive membrane permeability. Physicochemical properties such as lipophilicity (1) or hydrogen bonding capacity (2) correlates with passive membrane permeability for structurally homogenous data sets. However, correlations between single physicochemical parameters often break down when structural diversity is introduced. On the contrary, transport studies of compounds across monolayers of human intestinal epithelial cells (*i.e.* Caco-2 cells) show good correlation even for heterogeneous sets of compounds (3). The oral absorption potential can also be predicted by in situ perfusion in the rat with good results (4). However, these methods are very costly and time consuming and require the synthesis of at least mg quantities of the test compounds. Thus it would be of great economic and scientific value if passive membrane permeability could be predicted a priori with high precision by a computational method. Work in this direction has been published recently. van de Waterbeemd *et. al.* reported that calculated polar van der Waals surface area was correlated with permeability across the blood brain barrier (5) and that calculated molecular descriptors (mainly hydrogen bonding and molecular size) could be used to estimate passive membrane permeability across Caco-2 cells (6). Palm and co-workers reported the use of dynamic polar van der Waals surface area to predict passive permeability over Caco-2 cells and rat ileum (7). In this paper we describe a new tool for the prediction of Caco-2 cell permeability. The relationship between permeability and the molecular properties have been investigated using quantitative structure-property analysis based on MolSurf (8) parametrization with the Partial Least Squares Projections to Latent Structures (PLS) method (9) as statistical engine.

METHOD OF CALCULATION

Data Set

The following 17 compounds, previously studied by van de Waterbeemd *et. al.* (6), were used: Acetylsalicylate (Ac), Alprenolol (Al), Atenolol (At), Corticosterone (Co), Dexamethasone (De), Felodipine (Fe), Hydrocortisone (Hy), Mannitol (Ma), Metoprolol (Me), Olsalazine (Ol), Practolol (Pa), Propranolol (Pr), Salicylic acid (Sa), Sulfasalazine (Su), Terbutaline (Tb), Testosterone (Te) and Warfarin (Wa).

Caco-2 Cell Permeability Data

The experimental permeability values for the data set compounds were taken from Artursson and Karlsson (3) and are given in Table 1.

Calculated MolSurf Parameters

A summary of the computational protocol described in this section is depicted in Figure 1.

Conformational Analysis

The structures of the investigated compounds were built in Macromodel (10) and were modeled in their neutral forms. The three-dimensional structures were determined by Monte-Carlo—based conformational analysis performed with the MacroModel program package using the Merck Molecular Force

[1] Astra Pain Control AB, S-151 85 Södertälje, Sweden.
[2] Department of Pharmaceutics, Uppsala Biomedical Centre, Uppsala University, Box 580, S-751 23 Uppsala, Sweden.
[3] To whom correspondence should be addressed. (e-mail: ulf.norinder@pain.se.astra.com)

ABBREVIATIONS: PCA, principal component analysis; PLS, Partial Least Squares Projections to Latent Structures; PRESS, Predictive Residual Error Sum of Squares; HBA, Hydrogen Bond Acceptor; HBD, Hydrogen Bond Donor.

0724-8741/97/1200-1786$12.50/0

Table 1. Experimental, Calculated and Predicted Permeability Values Over Caco-2 Cells

		PLS model of log (Permeability)[a]					
		1		2		3	4
Compound	exp.[b]	calc.[c]	pred.[d]	calc.	pred.	calc.	calc.
At	−6.700	−6.247		−6.337		−6.174	−6.323
De	−4.903	−5.062		−5.044		−5.082	−4.988
Ma	−6.745	−6.837		−7.127		−6.693	−6.863
Ol	−6.959	−6.875		−6.747		−6.804	−6.719
Pr	−4.378	−4.886		−4.688		−4.816	−4.785
Sa	−4.924	−4.996		−5.078		−5.258	−5.268
Su	−6.886	−7.065		−7.042		−7.229	−7.251
Tb	−6.420	−6.322		−6.071		−6.384	−6.081
Wa	−4.417	−4.041		−4.197		−4.125	−4.316
Ac	−5.620		−4.623		−4.851	−4.997	−5.112
Al	−4.393		−4.864		−4.653	−4.769	−4.731
Co	−4.263		−4.449		−4.521	−4.459	−4.549
Fe	−4.644		−4.420		−4.295	−4.411	−4.407
Hy	−4.668		−5.090		−5.095	−5.051	−5.023
Me	−4.569		−4.779		−4.490	−4.692	−4.612
Pa	−6.046		−5.845		−5.823	−5.839	−5.902
Te	−4.286		−4.012		−3.785	−4.038	−3.892

[a] PLS models; 1 = Based on training set compounds with all variables, 2 = Based on training set compounds with the reduced set of variables, 3 = Based on all compounds with all variables, 4 = Based on all compounds with the reduced set of variables.
[b] Experimental log (Permeability) values taken from ref. 3.
[c] Calculated/fitted log (Permeability) values for the training set.
[d] Predicted log (Permeability) values for the test set.

Field (MMFF). One hundred starting conformations were generated for each structure. The energy minimizations were performed in vacuum. Unique minimized conformations within 5 kJ/mol of the lowest energy conformation were saved for further studies.

Semi-empirical Calculations

The conformation with the lowest found energy from the previous conformational analysis was subjected to a geometry optimization (energy minimization) using the semi-empirical quantum chemistry based AM1 method available in the Spartan program (11).

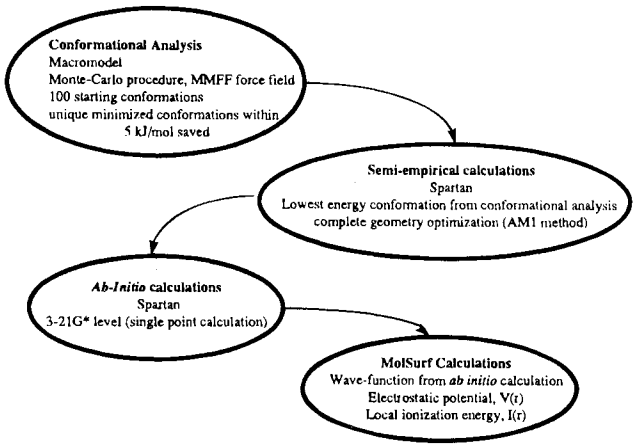

Fig. 1. The computational protocol used to compute the theoretical MolSurf molecular descriptors.

Ab-Initio Calculation

A quantum mechanical *ab initio* calculation using a 3-21G* basis set without further geometry optimization (single point calculation) was subsequently performed on all AM1 optimized (minimized) conformations using the Spartan program (11). The *ab initio* calculations were done to obtain a wave function, *i.e.* a quantum chemical description, of good quality for each investigated compound (12).

MolSurf Calculations

The wave-function from each *ab initio* calculation was used by MolSurf (8) to compute various properties related to the molecular valence region. The chemical behavior and, hence, the calculated properties depend on the distribution of electrons and energy in the valence region. This region is represented by a surface of constant electron density (0.001 electrons/bohr^3) encompassing the molecule. The electrostatic potential, V(r), and the local ionization energy, I(r), are calculated at points evenly distributed (0.28 bohr apart) on this surface. The former property, V(r), is related to the potential registered by a probe of positive unit charge positioned at each of the points on the surface. Similarly, the latter property, I(r), is the energy required to remove an electron from the molecule at each of the points on the surface. The computed descriptors describe properties such as base strength, hydrophobicity, hydrogen

Table 2. Computed MolSurf Descriptors

#	Descriptor	Designation
1	Surface	
2	Octanol/Water Partition Coefficient	logP
3	Polarizability	
4	Polarity	
5	Lewis base	LB
6	Lewis acid	LA
7	Hydrogen bond acceptor strength for oxygen atoms	HBAo
8	Number of hydrogen bond acceptor oxygen atoms	#HBAo
9	Hydrogen bond donor strength	HBD
10	Number of hydrogen bond donor atoms	#HBD
11	Sum of HBAn, HBAo and HBD	ΣHB
12	Hydrogen bond acceptor strength for nitrogen atoms	HBAn
13	Number of hydrogen bond acceptor nitrogen atoms	#HBAn

bonding, polarity as well as polarizability (see Table 2 for a list of calculated descriptors).

MolSurf parameters were calculated for the entire compound as well as for individual atoms contributing to hydrogen bonding. The number of possible hydrogen bond acceptors and donors, respectively, were also used as descriptors. The former were partitioned into two categories of oxygen and nitrogen type. This division of hydrogen bond acceptor types also applies to the corresponding variables using the actual computed hydrogen bond acceptor strengths. Additionally, the sum of the hydrogen bond acceptor and donor strengths was used as a descriptor.

Statistical Analysis

Training Set Selection

A training set consisting of 9 molecules was selected using the maximin approach of Marengo and Todeschini (13). The method works through an exchange algorithm where, in each cycle, a substitution is selected to provide the maximum increase of the minimum distance between the currently selected compounds. The procedure provides a final uniform distribution of the selected compounds from all available structures in parameter (chemical property) space. One hundred random starting points were used. The solution of 9 selected compounds with the largest found minimum distance between two compounds was used as the training set. The remaining 8 structures were used as an external test set to assess the predictivity of the derived models. The parameter space in this case was the scores of the first 5 principal components from a principal component analysis (PCA, see section on principal component analysis) on the 13 MolSurf parameters (see Table 2 for a list) computed for each compound of the entire data set.

Principal Component Analysis

A principal component analysis (PCA) (14–16) was performed on the descriptor matrix for the data set. In short, a PCA reexpresses the descriptor matrix X (*e.g.* a set of data collected for a number of compounds) as a mean vector plus the product of a few column score matrix T times a few row matrix P′. The PC-scores t contained in T provide the best linear summary of X with respect to compound description. These variables, often referred to as principal properties, are well suited as 'condensed' descriptors since they are few and orthogonal (independent) and are possible to interpret in terms of what particular aspects of the structures they reflect. In PCA, and other statistical methods as well, it is important to give the variables in the matrix an equal chance (weight) to influence the analysis regardless of their respective scales. This was ensured by an autoscaling procedure where the variance of each column was scaled to unit variance.

PLS Analysis

The relationship between the experimentally determined Caco-2 cell permeability values (log scale) and the computed MolSurf properties for the data set compounds was determined using the PLS (Partial Least Squares Projections to Latent Structures) method (9). The PLS method used in this work calculates one component at a time and stops when the added information becomes insignificant, as determined by a cross-validatory procedure (see below for details). In this way PLS summarizes the original variables stored in the descriptor matrix X as a few orthogonal new variables called scores (t) which are collected in the (few) column score matrix T. The scores are linear combinations of the original variables. The PLS method solves the problems of forming the model Y = f(T) and finding the coefficients (loadings) of the original variables that form each t at the same time. The loadings are collected in the row matrix P′. The PLS model, as expressed through the score matrix T, can subsequently be transformed into 'regression' coefficients of the original variables for comparison purposes so that the influence of each variable can be analyzed in a straight forward manner as is the case for ordinary multiple regression methods. The number of significant components is determined using a leave-one-out cross-validatory procedure (LOO-cv) (17). In such a procedure each compound is removed from the data set once and the remaining compounds are used to develop each model. The left out compound is then predicted from the developed model. The sum of the squared difference between predicted activity and experimental activity for the left out compounds (predictive residual sum of squares; PRESS) is computed. If the PRESS value is smaller for the latest calculated PLS component compared to the previous component then the former component is judged to be significant and kept in the model. Thus, the model with the smallest computed PRESS value is used. In order to avoid overfitting of the data a maximum of five PLS components was set as a limit.

A simple variable selection was also performed using a leave-one-out approach where each variable was left out of the model and its importance for predictivity, as judged by a leave-one-out cross-validation procedure (ref. 16, see above for an explanation), of the training set was assessed. If the predictivity of the model increased then the variable in question was permanently removed from model and otherwise the variable was kept permanently in the model.

Four analyses were performed: Two on the 9 training set compounds (see section on Training set selection for details) and two other on the entire data set of 17 structures.

Table 3. PLS Statistics of the Derived Caco-2 Permeability Models[a]

Model	R^2	cv-R^2	Npc	N^{tr}	sdev	F	$RMSE^{tr}$	p	$RMSEcv^{tr}$	N^{te}	$RMSEp^{te}$
1	0.932	0.738	2	9	0.340	40.88	0.277	<0.001	0.543	8	0.453
2	0.935	0.849	2	9	0.331	43.23	0.270	<0.001	0.412	8	0.409
3	0.901	0.791	2	17	0.352	63.43	0.319	<0.001	0.462		
4	0.909	0.852	2	17	0.336	70.27	0.305	<0.001	0.390		

[a] Model: see Table 1 for an explanation; R^2: ordinary correlation coefficient; cv-R^2: cross-validated (LOO) correlation coefficient; Npc: Number of PLS components; N^{tr}: Number of compounds in the training set; sdev: standard deviation; F: ordinary F-value; $RMSE^{tr}$: Ordinary root mean squared error for the dependent variable of the training set; p: level of significance; $RMSEcv^{tr}$: Root mean squared error for the dependent variable from the cross-validation procedure of the training set; N^{te}: Number of compounds in the test set; $RMSEp^{te}$: Root mean squared error for the dependent variable of the test set.

RESULTS

Principal Component Analysis

The PCA (principal component analysis) resulted in 5 principal components which explained 86.4% of the variance in the original matrix. The first to fifth component explained 24.8, 28.2, 15.3, 9.6 and 8.5%, respectively.

Training Set Selection

The selection of 9 training set compounds out of the 17 available molecules gave the following result. Compounds Atenolol (At), Dexamethasone (De), Mannitol (Ma), Olsalazine (Ol), Propranolol (Pr), Salicylic acid (Sa), Sulfasalazine (Su), Terbutaline (Tb) and Warfarin (Wa) were selected as training set.

PLS Analysis

The PLS analysis of the training set compounds using all the computed MolSurf descriptors (model 1) resulted in 2 significant PLS components according to cross-validation with $R^2 = 0.932$, cross-validated $R^2 = 0.738$, s = 0.340, F = 40.87, RMSE = 0.277 and $p < 0.001$. The corresponding analysis using the descriptors remaining after variable selection, *i.e.* excluding variables 5,6,9 and 11 (model 2) resulted in 2 significant PLS components according to cross-validation with $R^2 = 0.935$, cross-validated $R^2 = 0.849$, s = 0.331, F = 43.23, RMSE = 0.270 and $p < 0.001$. The PLS analysis on the entire data set also resulted in 2 significant PLS components according to cross-validation with $R^2 = 0.901$, cross-validated $R^2 = 0.791$, s = 0.352, F = 63.434, RMSE = 0.319 and $p < 0.001$ using all variables (model 3) while the corresponding values for the analysis based on the reduced set of parameters, *i.e.* excluding descriptors # 5, 6, 9, and 11, (model 4) were 0.909, 0.852, 0.336, 70.27, 0.305 and $p < 0.001$, respectively.

The results of all four PLS analyses are summarized in Table 3. The values for the PLS coefficients of the reduced parameter set models 2 and 4 are given in Table 4. Plots of experimental vs. calculated/predicted permeabilities for models 2 and 4 are shown in Figures 2 and 3, respectively.

DISCUSSION

The result of the training set selection is satisfactory not only from a statistical point of view since PLS models with good statistics and predictivity were developed (see below for further discussions) but also from a structural point of view.

Table 4. Scaled PLS Regression Coefficients of the Models Based on the Reduced Set of Variables[a]

MolSurf descriptor	Model 2	Model 4
Surface	0.0782	0.0922
logP	0.1289	0.1234
Polarizability	0.1040	0.1138
Polarity	−0.0864	−0.1561
HBAo	−0.0558	−0.0279
#HBAo	−0.2101	−0.1795
#HBD	−0.4662	−0.3582
HBAn	−0.3474	−0.3435
#HBAn	−0.3422	−0.3525

[a] Model: see Table 1 for an explanation.

The 9 compounds selected as the training set covers the structural classes of the data set rather well. Thus, there are representatives of both β-adrenergic compounds, steroids, organic acids, azo compounds as well as carbohydrates in the training set.

At the start of this investigation we did not separate the hydrogen bonding acceptors (HBAs) into oxygen and nitrogen types. However, during the first PLS analyses of the training set we noticed that compounds containing nitrogen hydrogen bonding acceptors were poorly predicted. This gave us the idea to separate the HBAs into nitrogen and oxygen types. This division resulted in a much better model with good internal as

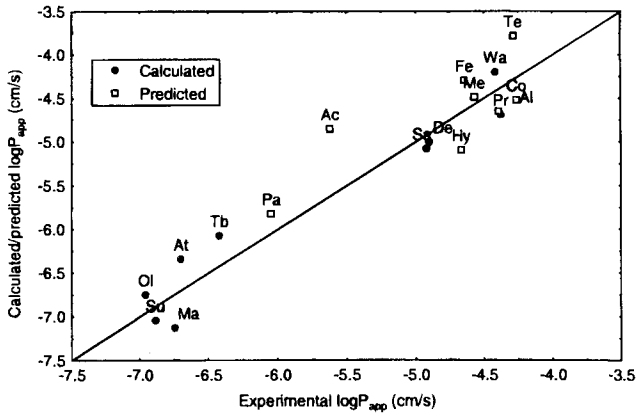

Fig. 2. Relationship between experimental and calculated/predicted permeability (PLS model 2) over Caco-2 cell monolayers for the model drugs.

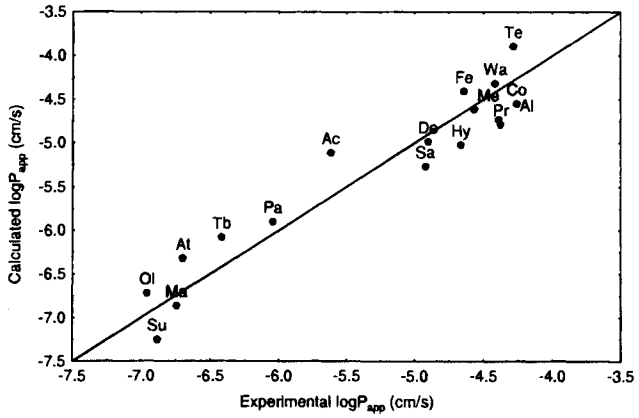

Fig. 3. Relationship between experimental and calculated permeability (PLS model 4) over Caco-2 cell monolayers for the model drugs.

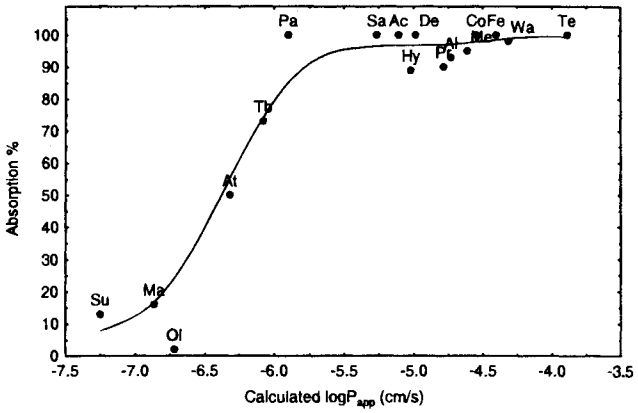

Fig. 4. Sigmoidal relationship between calculated permeability over Caco-2 cell monolayers for the 17 structurally heterogenous model drugs and absorption in humans after oral administration.

well as external predictivity as assessed by the cross-validatory procedure and the test set compounds, respectively. At this point we performed the variable selection. By removing the descriptors that proved detrimental to predictivity *i.e.* variables 5, 6, 9 and 11, for the training set we were able to develop a statistically better model with improved predictivity for both the training set and the test set (see Table 3). Both training set models (1 & 2) are well balanced since the predicted RMSE values of these models ($RMSEcv^{tr}$) are comparable to the corresponding values for the test set.

The same type of statistical improvement with respect to internal validation and other statistics were found for the PLS models based on the entire data set when removing variables 5, 6, 9, and 11 (the same variables as for model 2). Furthermore, since both final models (2 & 4) using the reduced set of variables show the same overall statistics the change in coefficients between the models is small and the correlation coefficient between the two sets of coefficients is high (0.982).

What are the physico-chemical interpretations from the results of the PLS analyses? The most important factors influencing the model are associated with hydrogen bonding. Thus, variables such as the number of possible hydrogen donor atoms as well as the number of hydrogen bond acceptor nitrogens have the greatest impact along with the actual strength of the hydrogen bond in the latter case. All these properties should be kept to a minimum to facilitate high permeability. This fact was also observed by van de Waterbeemd *et.al.* in their work (6). Additional factors that are important for high permeability are the absence of hydrogen bond acceptors related to oxygen atoms as well as an over-all non-polar character of the structure. An interesting observation is that the derived models attribute greater (or equal in the case of HBAn) weight to the hydrogen bond variables related to the number of possible sites that may engage in such interactions than the actual strength of the same interactions. One tentative explanation may be that it is more detrimental to the transport of the molecule across the membrane to have many, but weaker, hydrogen bond interactions, possibly arranged in some sort of network configuration, than a few but stronger hydrogen bond contacts. As may be expected for processes of this kind, high lipophilicity (logP) of the compound is also favorable for efficient permeability. Further more, the presence of polarizable electrons, *e.g.* conjugated and aromatic substructures as well as the larger halogens, are also beneficial for the compound to penetrate the membrane effectively. One interpretation of this property may be that compounds having many polarizable electrons can act as molecular chameleons. Thus, the polarizable electrons are first used to promote induced interactions of an electrostatic nature, *e.g.* induced dipole-dipole and induced dipole-induced dipole interactions, in the more hydrophilic phase as the compound is transported towards the membrane surface. Once the compound reaches the membrane and starts to penetrate across the membrane the same polarizable electrons are now used to minimize the electrostatic interactions in order to become as unpolar and electrostatically 'inert' as possible in the more lipophilic phase. Part of the information content of the polarizable variable may also be related to charge-transfer interactions. An indication towards this interpretation stems from the fact that the addition of a charge-transfer term to the models results in a positive PLS coefficient, *i.e.* such interactions seem to promote high permeability, although the addition of the term itself does not alter the overall statistics of the derived PLS models (18).

Thus PLS models with good predictability and overall statistical quality have been derived for the transport of small molecules across Caco-2 cell monolayers. Furthermore, these statistical PLS models are based on a physico-chemical MolSurf parametrization that makes them easy to interpret with respect to structural requirements that promote high permeability. Also, the parametrization gives rise to new possible insights with respects to the mechanisms operating during the transport of small molecules across membranes. However, as was mentioned in the introduction the main use of Caco-2 cell permeability data is to predict oral absorption in humans. Figure 4 illustrates the sigmoidal relationship between calculated permeability across Caco-2 cell monolayers and oral absorption in humans. Further work is under way to model oral absorption in humans directly and to investigate transport across other types of membranes using MolSurf parametrization and multivariate PLS statistics. We hope to report the results of these studies in the near future.

REFERENCES

1. Y. C. Martin. *J. Med. Chem.* **24**:229–237 (1981).
2. W. D. Stein. The molecular basis of diffusion across membranes. *In The Movement of Molecules Across Cell Membranes,* Academic Press, New York, 1967, pp. 65–125.

3. P. Artursson and J. Karlsson. *Biochem. Biophys. Res. Commun.* **175**:880–885 (1991).
4. L. Amidon, P. J. Sinko, and D. Fleisher. *Pharm. Res.* **5**:651–654 (1988).
5. H. van de Waterbeemd and M. Kansy. *Chimia* **46**:299–303 (1992).
6. H. van de Waterbeemd, G. Camenisch, G. Folkers, and O. A. Raevsky. *Quant. Struct.-Act. Relat.* **15**:480–490 (1996).
7. K. Palm, K. Luthman, A-L. Ungell, G. Strandlund, and P. Artursson. *J. Pharm. Sci.,* **85**:32–39 (1996).
8. MolSurf version 2.0, Qemist AB, Hertig Carls allé 29, S-691 41 Karlskoga, Sweden, e-mail: par.sjoberg@mbox309.swipnet.se.
9. S. Wold, E. Johansson, and M. Cocchi. PLS-Partial least-squares projections to latent structures. In H. Kubinyi (ed.), *3D QSAR in Drug Design,* ESCOM, Leiden, 1993, pp. 523–550.
10. Macromodel version 5.5, Dept. Chem., Columbia Univ., New York, NY 10027, USA.
11. Spartan version 4.1, Wavefunction, Inc., 18401 Von Karman Ave., #370, Irvine, CA 92715, USA.
12. The MolSurf program presently requires an *ab initio* wavefunction due to the program's current parametrization and underlying methodology. Work is currently under way that will enable the use of semi-empirical AM1 calculations throughout the quantum mechanical parts of the computational protocol.
13. E. Marengo and R. Todeschini. *Chem. Intel. Lab. Systems* **16**:37–44 (1992).
14. J. E. Jackson. *A Users Guide to Principal Components,* Wiley & Sons, New York, 1991.
15. I. T. Jolliffe. *Principal Component Analysis,* Springer Verlag, New York, 1986.
16. E. R. Malinowski. *Factor Analysis in Chemistry,* 2nd ed., Wiley & Sons, New York, 1991.
17. S. Wold. *Technometrics* **20**:379–405 (1979).
18. U. Norinder, unpublished results.