Prediction of Drug Absorption Using Multivariate Statistics

William J. Egan,*,† Kenneth M. Merz, Jr.,† and John J. Baldwin‡

Center for Informatics & Drug Discovery and Pharmacopeia Laboratories, Pharmacopeia, Inc., CN 5350, Princeton, New Jersey 08543-5350

Received July 10, 2000

Literature data on compounds both well- and poorly-absorbed in humans were used to build a statistical pattern recognition model of passive intestinal absorption. Robust outlier detection was utilized to analyze the well-absorbed compounds, some of which were intermingled with the poorly-absorbed compounds in the model space. Outliers were identified as being actively transported. The descriptors chosen for inclusion in the model were PSA and AlogP98, based on consideration of the physical processes involved in membrane permeability and the interrelationships and redundancies between available descriptors. These descriptors are quite straightforward for a medicinal chemist to interpret, enhancing the utility of the model. Molecular weight, while often used in passive absorption models, was shown to be superfluous, as it is already a component of both PSA and AlogP98. Extensive validation of the model on hundreds of known orally delivered drugs, "drug-like" molecules, and Pharmacopeia, Inc. compounds, which had been assayed for Caco-2 cell permeability, demonstrated a good rate of successful predictions (74−92%, depending on the dataset and exact criterion used).

Introduction

The primary goal of the drug discovery and development process is to find a molecule possessing both good pharmacodynamic *and* good pharmacokinetic properties. Ideally, a new drug should be efficacious and selective, target-tissue(s)-specific, and orally-absorbed, cause minimal or no adverse effects due to metabolite activity or toxicity, and be distributed/excreted in such a fashion as to permit dosage once a day. Successful optimization of all these properties is an extremely challenging task. It is a goal that the pharmaceutical industry has had difficulty in reaching, as demonstrated by the high failure rates for lead compounds (>90−99%) during the development process. Only 50% of compounds fail in preclinical development, leaving many unsuitable compounds to progress into expensive clinical testing.[1] Considering that various sources estimate over 80% of compounds for which an IND has been filed fail before reaching NDA status, and considering that approximately 85% of the average total cost of an approved drug ($350−500+ million) is incurred *after* clinical testing has commenced, these many failures are very costly, indeed.[1−4]

Recently, considerable, and clearly long-overdue, interest has focused on the discovery stage assessment of pharmacokinetic properties (absorption, distribution, metabolism, excretion/ADME[5]) of compounds, as well as their pharmacological activity.[3,6−15] Good ADME/toxicity properties are just as critical as therapeutic activity. As one survey found,[16] 50.4% of development failures among 319 new chemical entities produced by seven pharmaceutical companies over a 21-year span were due to ADME/toxicity problems and occurred during clinical trials. The successful design of a maximally active compound will result in the waste of hundreds of millions of dollars per approved drug if the compound is nonselective, poorly orally-absorbed, metabolically unstable, rapidly excreted, or toxic or will not distribute into the target tissues. The pharmaceutical industry has therefore shown strong, widespread interest in the "fail fast, fail cheap" concept.

Various plans have been proposed to design molecules to have good ADME/toxicity properties. In an excellent discussion, Tarbit and Berman[13] present one feasible method: "The ability to screen combinatorial libraries of known chemical properties through high-capacity ADME screens, which model specific physiological processes of absorption, metabolism, and so on, will produce large amounts of data that will significantly aid the development of 'predictive' computational models. These models can then be used 'on line' to test compound structures and thereby aid in the design of optimized compound libraries prior to synthesis." Provided these data are of high quality, i.e., the data are sufficiently accurate and precise, are obtained under consistent and appropriate experimental conditions, and the compounds analyzed cover the entire chemical space explicitly related to each property of interest, the likelihood of constructing a reasonably useful model should be fairly high. Not only should new compounds created using combinatorial techniques be analyzed, but existing drugs and many "drug-like" compounds also should be screened extensively using these same methods and conditions, to provide consistent and comparable reference data. Care must be taken, for predictive computational models are constrained by a harsh reality − good statistical methods cannot save bad data or insufficient quantities of good data, and poor statistical methods can distort or even lead to completely erroneous conclusions despite being derived from copious amounts of the best data. Furthermore, the best form of a computational model is one which informs and aids

* To whom correspondence should be addressed. Phone: (609)-452-3794. Fax: (732)-422-0156. E-mail: wegan@pharmacop.com.
† Center for Informatics & Drug Discovery.
‡ Pharmacopeia Laboratories.

10.1021/jm000292e CCC: $19.00
Published on Web 09/19/2000

the chemical intuition of the medicinal chemist, and to accomplish those aims the model must be as understandable and easy to use as possible. A medicinal chemist will have difficulty learning from or making the best use of an incomprehensible "black box" and will be less likely to trust such a model implementation.

Although the pharmaceutical industry is at an early stage in developing ADME/toxicity models, a considerable amount of work has already been performed. One area of concentrated and visible effort has been in the computational prediction of oral absorption. Oral bioavailability is one of the most desirable attributes of a new drug, and the first step in attaining oral bioavailability is to achieve good oral absorption. Despite the present lack of large amounts of high-quality data from HTS ADME screening, significant progress has been made in the computational modeling of oral absorption. This paper describes the development and validation of a passive intestinal absorption (PIA) model which is novel in several respects when compared to previously published work and was designed to be easy to use and interpret.

Existing Computational Models for PIA

The problem of predicting PIA has been approached in a variety of ways recently.[17] Lipinski et al.[18] developed a popular, simple, and descriptive univariate model, called the "Rule of 5," by analyzing a set of 2287 compounds having United States Adopted Name (USAN) or International Nonproprietary Name (INN) designations. They chose to analyze compounds with USAN/INN designations because such designations are typically applied for prior to entry into phase II clinical trials, meaning that these compounds had satisfactorily completed phase I clinical trials, which evaluate safe dosage level and include assessments of ADME/toxicity properties. Upper bounds for property distributions were determined for properties related to lipophilicity, size, and hydrogen bonding (H-bonding), based on the 90th percentile (approximately) of the property distributions. In this model, compounds are considered less likely to be permeable when ClogP > 5, or molecular weight (MW) > 500, or the number of H-bond donors > 5, or the number of H-bond acceptors > 10, and to have particularly poor permeability if two of these bounds are exceeded. Ghose et al.[19] performed a similar descriptive analysis of 6454 compounds they considered drug-like by therapeutic class in the Comprehensive Medicinal Chemistry database and identified 80th percentile ranges for AlogP (−0.4−5.6), MW (160−480), molar refractivity (40−130), and number of atoms (20−70). These models have the benefits of being easy to use and interpret, but they do not take into account the interactions between descriptors, and while they are based upon large datasets, those datasets are only approximately related to absorption: i.e., a compound which has a USAN/INN designation is likely to be reasonably well-absorbed, but we do not know exactly how well-absorbed.

Several researchers have examined the effect of H-bonding on permeability using dynamic polar surface area (PSA_d).[20] PSA_d was computed as the van der Waals surface area of all nitrogen and oxygen atoms, plus their attached hydrogen atoms, Boltzmann averaged over each of the low-energy conformers of a molecule. PSA_d was shown to have a strong, inverse sigmoidal relationship ($r^2 = 0.94$) with percent human intestinal absorption (FA) for a set of 20 molecules whose PSA_d covered the range 53.1−242.1 Å^2.[20] The sigmoidal relationship predicts FA < 10% for PSA_d > 139 Å.^{2} Lipophilicity (ClogP ranging from −8.09 to 3.29 for the 20 molecules) had a much poorer relationship to FA ($r^2 = 0.34$), and no relationship was found with nonpolar surface area. Later work concluded that differences in the "simulated" environment (vacuum, chloroform, and water) had little effect on PSA_d.[21] Clark[22] reviewed the historical use of polar surface area (PSA) in the modeling of solvation and partition processes and demonstrated that using a single, low-energy conformer to compute PSA performed equally as well as the PSA_d method and has the advantage of being far faster to compute. Kelder et al.[23] computed the PSA values for a set of 1590 orally administered non-CNS drugs which reached at least phase II clinical trials, and the published histogram shows only a small fraction of those compounds have PSA > 150 Å.^{2} The combination of dynamic nonpolar surface area ($NPSA_d$) and PSA_d has also been used to fit a sigmoidal model to the Caco-2 cell permeability of 12 oligopeptide derivatives ($r^2 = 0.96$) and to predict the permeability of 7 more oligopeptide derivatives.[24] While a relationship of PSA to permeability has been demonstrated, the models usually do not take into account the effects of other descriptors. Also, the datasets used to build the PSA models are so small that, although a wide range of PSA was covered, the entire chemical space related to PIA is likely not covered.

More complex multivariate models, incorporating both linear and nonlinear relationships, have been used to model passive intestinal absorption. Camenisch et al.[25] used a sigmoidal relationship to model the effects of MW and lipophilicity (log *D* at pH 7.4) on the Caco-2 cell permeability for 36 compounds. Their results suggest as log *D* decreases from 3.66 to −4.5, there is a sigmoidal decrease in Caco-2 permeability. The shape of the sigmoid was dependent on MW. Further theoretical work by the same group concluded that pH effects may be ignored because donor and acceptor compartments often have the same pH in Caco-2 cell permeability studies. They considered the pH finding to support their earlier conclusion of a MW dependence of the sigmoidal lipophilicity−permeability relationship.[26]

Van de Waterbeemd et al.[27] constructed a series of linear models containing H-bonding and MW terms, and the combination of MW and Cad (a sum of free energy H-bond donor and H-bond acceptor factors) best fit the Caco-2 cell permeability for 17 compounds ($r^2 = 0.883$). Norinder et al.[28] used partial least-squares (PLS) regression to model the same dataset using MolSurf descriptors (which include log *P*, polarizability, numbers and strengths of H-bond acceptor nitrogen and oxygen atoms, number of H-bond donor atoms) with good results ($r^2 = 0.935$) for the training set. PLS regression using MolSurf descriptors was also applied[29] to the dataset modeled by Palm et al.[20] producing a linear model ($r^2 = 0.916$) from similar descriptor selections. A neural network approach using a genetic algorithm for descriptor selection was used to quantitatively predict human intestinal absorption.[30] As the authors pointed

out,[30] the complex interrelationships in the neural network make it difficult to gauge the contribution of an individual descriptor. While this paper used a larger dataset than most other papers (86 compounds), the inclusion of actively transported compounds and a skewed bias toward well-absorbed compounds have been criticized by Clark.[22]

Three even more complex models and their analyses are of particular interest. (1) Human jejunal permeability was well-predicted ($r^2 = 0.98$, $q^2 = 0.96$) for a small training set of 13 compounds. These compounds were selected for diversity when compared against 138 common drugs using a PLS model which selected only ClogP, H-bond donor count, and single-conformer PSA from 18 descriptors.[31] Interestingly, log *D* values at pH 5.5, 6.5, and 7.4 were not selected for inclusion in the best model. A bi-plot of the first two principal components (explaining 67% of the variance of 14 descriptors) for all 158 compounds examined (mostly common drugs) could not separate compounds known to be actively transported from passively transported compounds, and the linear model predicted much lower jejunal permeability than was measured for known actively transported compounds. (2) Models that used combinations of hydrophilic and hydrophobic factors best-explained membrane partitioning as measured by chromatography using phospolipid stationary phases on a set of 20 D-optimally designed tetrapeptides.[32] One model suggests that negative charge contributes to poor partitioning, while positive charge had a nonsignificant effect. Molecular volume and molecular surface area both had strong positive contributions to partitioning, but this size effect was considered to be due to the correlation between MW and lipophilicity. (3) A focused combinatorial optimization library of 449 compounds was designed to meet absorption related constraints (the rule of 5 and PSA < 140 Å^2), and those constraints greatly improved absorption as measured by Caco-2 cell permeability, while the percentage of compounds more active than a target molecule was simultaneously successfully optimized.[33] The first two models provide contradictory results for the importance of charge effects, and the third model, which does not consider charge effects, demonstrates that the simultaneous optimization of activity and absorption using univariate property constraints provides good results.

Rationale for the Current Work

While there has been considerable research on the computational modeling of PIA, there are several areas in which we felt improvements might be made. Specifically, we thought that it would be possible to build a better computational model for PIA by (1) focusing more on the multivariate nature of the problem and how we choose to describe the factors influencing PIA and (2) using much larger datasets which are (a) specifically related to PIA (as much as possible) and (b) contain more accurate and precise information.

To build a comprehensive computational model for PIA, we must choose which properties to use to describe a molecule, based on our understanding of the physical processes governing absorption. The properties of lipophilicity, hydrophilicity, size, and degree of ionization are generally regarded as the most crucial factors affecting the passive intestinal absorption of a molecule,[34−38] and existing computational models incorporate one or more of these factors in some fashion.

According to the fluid mosaic model,[39] the structure of a cell membrane is considered to be an interrupted phospholipid bilayer capable of both hydrophilic and hydrophobic interactions. Transcellular passage through the membrane lipid/aqueous environment is viewed as the predominant pathway for passive absorption of lipophilic compounds, while low-molecular-weight (<200), hydrophilic compounds make use of the water-filled channels of the tight junctions between membrane cells (paracellular transport).[36,37,40] For this reason, lipophilicity has been considered a key property for activity in drug design for many years[41−43] and is a common property used to estimate the membrane permeability of a molecule. Lipophilicity is often measured as the log of the partition coefficient between *n*-octanol and water (log *P*). A variety of methods[44] can estimate log *P* computationally with good results. The relationship between log *P* and permeability is nonlinear, with drops in permeability at both low and high log *P*. These nonlinearities are theorized to be due to (1) the inability of weakly lipophilic compounds to penetrate the lipid portion of the membrane and (2) the excessive partitioning of strongly lipophilic compounds into the lipid portion of the membrane and their subsequent inability to pass through the aqueous portion of the membrane.[38,43,45−47]

Conradi et al.[36] consider lipophilicity by itself to be inadequate for the estimation of a solute's ability to penetrate a membrane barrier. Instead, they argue that both hydrophobic effects and H-bonding forces must be considered, rather than just lipophilicity. The H-bonding ability (hydrophilicity) of a molecule has long been known to be an important property for membrane permeation,[48,49] and more recent models using PSA to estimate H-bonding ability have demonstrated a nonlinear relationship between PSA and permeability, with permeability declining sigmoidally as PSA increases.[20] While log *P* is generally used to estimate a compound's lipophilicity, the fact that log *P* is a ratio raises a concern about the use of log *P* to estimate hydrophilicity and hydrophobicity, in our view. This is because the use of a ratio by itself causes a loss of information. For example, a log *P* of 2.0 merely specifies that the concentration of a compound in *n*-octanol is 100-fold that in water, but it cannot tell you the actual concentration in either solvent. Thus, a second piece of information, such as provided by some measure of H-bonding, e.g., PSA, is necessary to provide the frame of reference.

Camenisch et al.[37] recently reviewed the effect that the degree of ionization has on membrane permeability. According to the pH-partition theory, only the unionized form of a compound may cross a cell membrane. However, Palm et al.[50] demonstrated that for compounds whose fraction un-ionized was less than 10%, a state which would be common for a large number of drugs over the pH range encountered during intestinal absorption, the ionized form contributes significantly to permeability across Caco-2 cell membranes. As discussed by these authors,[37,50] a number of examples exist where membrane permeability is greater than would be expected from pH-partition theory.

To estimate the general effect of charge on a molecule's absorption, we performed a qualitative analysis of the well-absorbed (WAbs) dataset (199 compounds with absorption ≥ 90%, fully described below). pK_a values were collected from a medicinal chemistry text[51] or computed using commercial software (ACD/pK_a DB v4.0, Advanced Chemistry Development, Inc., Toronto, Canada) for all compounds in the WAbs dataset. The fraction ionized was computed for each compound at pH values of 5, 7.4, and 8, and compounds were classified according to whether they were less than 10% un-ionized for each pH value, similar to the method used by Palm et al.[50] At pH values of 5, 7.4, and 8, the percentages of compounds less than 10% un-ionized were 60.8%, 61.8%, and 55.8%, respectively. This qualitative analysis of a reasonably large dataset lends support to the conclusion of Palm et al.[50] that the contribution of the ionized form to absorption is significant and suggests that further work is required to better understand the effects of molecular charge on absorption. Due to the field's lack of understanding of the full effects of charge on absorption, we decided to not include charge as an explicit factor in our modeling effort and thus considered only lipophilicity, hydrophilicity, and size.

Dataset Construction

The quality and quantity of data are of paramount concern. Most published models for passive intestinal absorption have been constructed from small-sized datasets which in many cases do not cover the entire chemical space associated with the property of absorption, as measured by factors deemed relevant, e.g., lipophilicity (log *P*). Consequently, we considered assembling a large set of data on compound absorption or permeability from the literature. Two types of data are readily available: the reported percent intestinal absorption, generally in humans, and the permeability of compounds as measured by the in vitro Caco-2 cell permeability assay.[52,53] Modeling human absorption data is obviously the best approach, because it is the actual property we are interested in predicting in silico, but use of human absorption data has the drawback that only a small amount of new data can be added for validation and model improvement purposes as time passes, due to the difficulties and costs of obtaining human intestinal absorption data. On the other hand, Caco-2 cell permeability assays have the advantage of greatly increased throughput (comparatively) and lower cost and have also been shown to have reasonable correlation with human absorption.[55] Unfortunately, considerable inter- and intralaboratory variability exists in Caco-2 cell permeability measurements. Artursson et al.[52] discussed the sources of variability in Caco-2 cell permeability assays and compared four calibration curves between percent human absorption and Caco-2 cell permeability, finding high interlaboratory variability; the curves were shifted relative to one another by approximately 0.25−1.75 log apparent permeability units. To assess intralaboratory variability, we randomly surveyed published Caco-2 cell permeability studies and found five studies which reported information on mean and standard deviation values for replicate measurements.[54,56−59] The average percent relative standard deviations (100 × standard deviation ÷ mean, %RSD) for the five studies were 5.6%, <10%, 10.3%, 12.7%, and 28.3%. We concluded that the inter- and intralaboratory variabilities were too high to combine published Caco-2 cell permeabilities from different sources to form one large dataset.

Use of in vivo human passive absorption data carries the same risk of high variability. Decades worth of data collected using different experimental methods and under different conditions are very likely not comparable, except at the extremes. The variability observed in Caco-2 cell permeability assays suggests that our ability to precisely measure passive intestinal absorption in vivo is equally limited. Well- and poorly-absorbed compounds should be easily separated by a statistical pattern recognition model, despite the likely high variability in measurement. To quantitatively predict differences in the percent absorption for compounds whose measured percent absorptions are similar (within 10−20%, roughly) would be far more difficult, because of numerous reports in the literature that lipophilicity (log *P*) and H-bonding ability (PSA) are nonlinearly related to permeability and hence percent absorption. This suggests that we should not build a quantitative model for passive absorption unless we have highly precise data. The observed nonlinearities are steep drops in absorption/permeability, and if our measurements of absorption/permeability are poor, the measurement imprecision limits the possible fit of a nonlinear quantitative model for data in those regions of sharp change.

Consequently, we decided to model passive intestinal absorption using a statistical pattern recognition method applied to a large set of literature data of compounds with high (≥90%) and low (<30%) human percent intestinal absorption. Validation was performed using various literature derived datasets and Caco-2 cell permeability assay results for compounds developed internally at Pharmacopeia, Inc. The six datasets are described in detail below.

The Datasets. 1 and 2. Well-Absorbed (WAbs) and Poorly-Absorbed (PAbs) Compounds Datasets. A list of compounds with good and poor absorption were compiled from a variety of literature sources.[20,30,60−65] The WAbs dataset contains 199 compounds described in the literature as having absorption ≥ 90% or an oral bioavailability ≥ 90% (which implicitly requires an absorption of at least 90%). The PAbs compound dataset contains 35 compounds described as having an absorption < 30%. Discrepancies between absorption values listed in different sources were examined to determine if bioavailability had been reported instead of absorption or if other factors were causing the discrepancy, e.g., formulation effects or food−drug interactions. If the discrepancy could not be resolved, the compound in question was not included. Active transport mechanisms were not taken into account in the creation of these two datasets but were considered in the analysis below. Quaternary amines were excluded from the PAbs datset.

3. Comprehensive Medicinal Chemistry (CMC) Dataset. The CMC Database (CMC 3-D 99.1, MDL Information Systems, Inc., San Leandro, CA) contains 7577 entries and was used to select compounds deemed to be drug-like by therapeutic category (class). Following published methodologies,[19,66] we initially eliminated compounds in the following classes: radiopague and contrast agents, disinfectants, spermicides, wetting agents, flavorings, pharmaceutical aids, surgical aids, dental, surfactants, sunscreen and ultraviolet screens, preservatives, aerosols, chelating agents, insecticides, astringents, herbicides, solvents, laxatives, sweeteners, adhesives, dentistry, veterinary, buffers, and scabicides. We then examined the classes of the remaining 6273 compounds to see if further culling was warranted, because the CMC database is updated several times each year. After further review, a number of the remaining classes were considered non-drug-like and compounds in the following classes were also removed: antacids, alcohol denaturants, alkalizing agents, ammonium detoxicants, bases for collodion, blood substitutes and blood volume determinations, body imaging, calcium supplements and calcium replenishers, caustics, avian, chlorinating agents, poultry, complexing agents, detergents, diagnostic aids, emulsions and emulsifiers, indicators, MRI agents, potassium-removing resins, prosthetic aids, replenishers, rodenticides, tooth discoloration inhibitors, oleaginous vehicles, supplements, pH sensing agents, radioactive and radioprotective agents, repellents (arthropod), topical, swine, and hematinic and antianemic agents. Compounds containing X and Li atom entries, as well as several entries with structural problems, were also eliminated, leaving 5836 compounds which were deemed reasonably drug-like by therapeutic category.

4. USAN/INN Dataset. This dataset contains 8504 compounds extracted from the World Drug Index (WDI, March 1998, Derwent Information, London, U.K.) which have either USAN (United States Adopted Name, 7572 compounds) or INN (International Nonproprietary Name, 6489 compounds) designations.

5. Physician's Desk Reference (PDR) Dataset. A list of 438 drugs which are marketed in orally available forms (tablets, capsules, caplets, and liquid suspensions) was compiled from a thorough search of the PDR Electronic Library,[65] a searchable database of the Physician's Desk Reference. While the CMC and USAN/INN datasets represent approximate measures of drug-likeness, the PDR compounds are all orally delivered and represent a more exact standard for comparing absorption potential.

6. Pharmacopeia Compounds. Pharmacopeia, Inc.'s discovery and lead optimization efforts include the use of a standard Caco-2 cell permeability assay, performed as apical-to-basolateral transport experiments in the absence of P-glycoprotein inhibition.[54,67−71] The dataset is composed of 446 compounds selected from various programs at Pharmacopeia, Inc. for determination of apparent Caco-2 cell permeability (P_{app}, nm/s) in accordance with program requirements. Structural information is confidential. Upon the basis of the P_{app} values determined experimentally for standard compounds, compounds with $P_{app} < \sim 34$ nm/s are considered likely to be poorly-absorbed (<30%) and compounds with $P_{app} > \sim 100$ nm/s are considered likely to be well-absorbed (>90%).

Calculations. Structures were converted to neutral form, where necessary, to facilitate property calculation. Structures were OFF energy-minimized using Cerius^2 4.0 versions ccI and ccJ (Molecular Simulations, Inc., San Diego, CA), with the minimization terminating after 1000 iterations using the default "high-convergence" settings. Descriptors were also computed in Cerius^2 using default settings, unless otherwise specified. PSA was calculated as the van der Waals surface area of oxygen and nitrogen atoms, including any attached hydrogen atoms, with the modified Jurs-TPSA descriptor in Cerius^2 (polar atoms option set to N,O and probe radius set to 0 Å). AlogP[72] was computed using the Cerius^2 4.0 AlogP98 descriptor. Statistical models were created using MATLAB 5.3 (The Mathworks, Inc., Natick, MA) and Cerius^2 4.0. pK_a values were computed using ACD/pK_a DB v4.0 (Advanced Chemistry Development, Inc., Toronto, Canada).

Descriptors. Upon the basis of our examination of the relevant literature, the most appropriate factors to consider in a passive absorption model are lipophilicity, hydrophilicity, and size. We chose AlogP98, PSA, and MW as variables to measure these factors. All three factors are interrelated, and the correlations of the chosen variables with each other must be taken into account when building a model. Therefore, we computed all three descriptors for the CMC dataset, autoscaled each descriptor to remove the effect of unequal variances,[73] and analyzed the autoscaled CMC dataset with principal component analysis (PCA)[74,75] to assess the numbers of independent contributions to the variance of the dataset. Despite the potential pitfalls of assigning physical meaning to the abstract, orthonormal linear combinations of the original variables which are called principal components (PCs),[76] it is revealing that the first two PCs account for 96.64% of the variance in the autoscaled CMC dataset.

Figure 1 is a set of bi-plots of the CMC dataset plotting (A) MW vs AlogP98, (B) MW vs PSA, and (C) AlogP98 vs PSA. MW is shown to have a hyperbolic bounded relationship with AlogP98, with the lower limit of MW increasing at both negative and positive extremes of AlogP98. For the majority of the compounds, MW increases with increasing AlogP98. Closer inspection shows that for AlogP98 > 0, MW trends upward with increasing AlogP98, and this trend reverses abruptly at AlogP98 = 0, where MW begins to trend upward with decreasing AlogP98. MW is shown to generally increase with PSA and has a lower bound which increases as PSA increases. AlogP98 generally decreases as PSA increases.

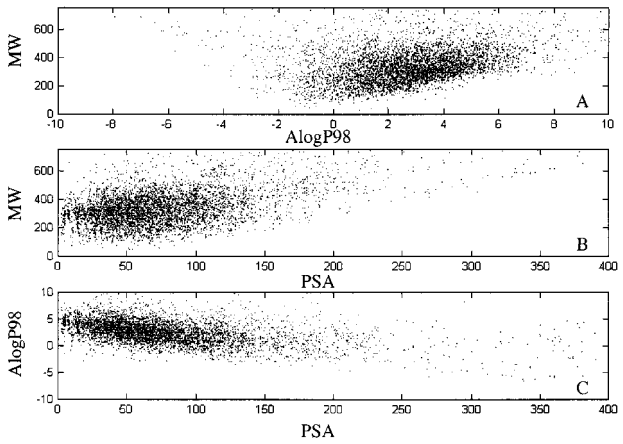

Figure 1. Three bi-plots of the CMC dataset for (A) MW vs AlogP98, (B) MW vs PSA, and (C) AlogP98 vs PSA, demonstrating the interrelationships between MW and the other two descriptors.

The large amount of the variance explained by the first two PCs (96.64%) in the CMC dataset and the obvious interrelationships shown in the bi-plots in Figure 1 indicate that one of the descriptors is likely redundant, i.e., containing information identical or very similar to that contained by the other two descriptors. It is well-known that the size of a molecule is related to log *P*.[77] For PSA to increase, the number of nitrogen and oxygen atoms (and any attached hydrogen atoms) must by definition increase as well, thereby increasing the MW of the molecule. Furthermore, nitrogen and/or oxygen atoms have a negative contribution to log *P*,[72] depending on topology, indicative of their H-bonding ability and preference for an aqueous environment, which is the factor PSA is used to estimate. Therefore, we concluded that MW was the redundant descriptor.

This conclusion contradicts the results of Camenisch et al.[25,26] who found a MW dependence of the sigmoidal lipophilicity−permeability relationship in their data. However, the demonstrated relationships of MW to both PSA and log *P*, the sigmoidal relationship between PSA and permeability,[20] and the known physical importance of H-bonding and lipophilicity to membrane permeation[38,43,45,46,48,49] all support the conclusion that log *P* (AlogP98) and PSA are the most relevant descriptors. This then suggests that MW is simply providing some information regarding H-bonding ability in the log *D*-based model proposed by Camenisch et al.[25,26] The use of PSA to provide a reference point for log *P* simply equates to using a more exact measure for the reference point.

Modeling

As discussed, we consider a pattern recognition model to be the most appropriate type of model for this problem, given the available quantity and quality of data. The simplest type of pattern recognition model answers the following question: Is a new molecule similar to some class of molecules of interest? Use of standard multivariate methods also permits us to take into account the effect that one variable may influence the permissible boundaries of another variable; see Rencher[78] for an excellent discussion.

The initial pattern recognition model is shown in Figure 2. The WAbs and PAbs compounds were plotted against AlogP98 vs PSA. A 95% confidence ellipse for the WAbs dataset was also computed and plotted. The 95% confidence ellipse represents the region of chemical space where we can expect to find well-absorbed compounds (≥90%) 95 out of 100 times, if certain statistical assumptions have not been violated. The confidence

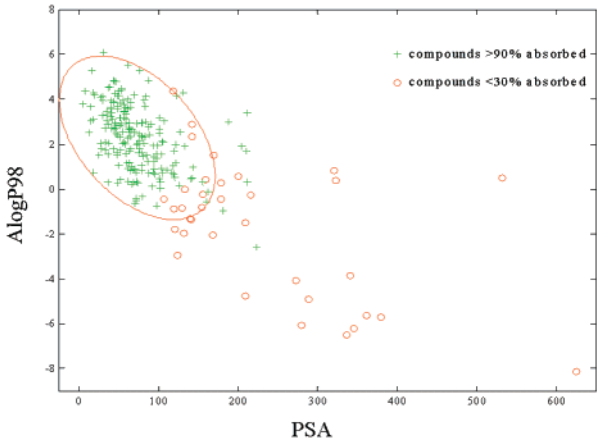

Figure 2. WAbs and PAbs compounds plotted on PSA−AlogP98 axes with a standard 95% confidence ellipse.

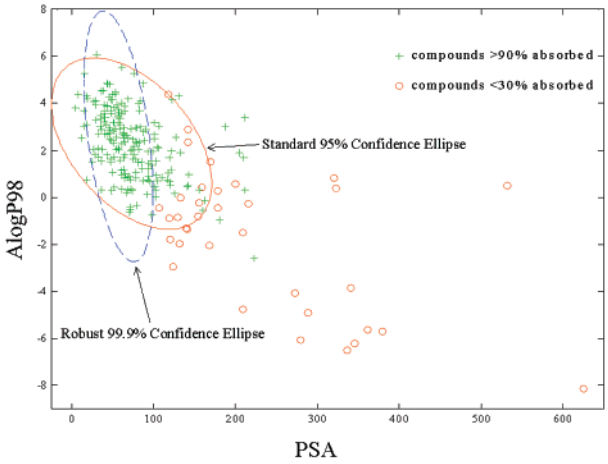

Figure 3. WAbs and PAbs compounds plotted on PSA−AlogP98 axes with the standard 95% confidence ellipse and the robust 99.9% confidence ellipse derived from SHV.

ellipse was computed using Hotelling's T^2.[78,79] The confidence ellipse takes into account the interactions (correlations) between the descriptors and would be a circle if the descriptors AlogP98 and PSA were totally uncorrelated for these compounds. As Figure 2 shows, there is overlap between a number of the well-absorbed compounds and the poorly-absorbed compounds. To investigate the cause(s) of this overlap, we employed a robust outlier detection method.

Barnett and Lewis[80] define an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data." Standard multivariate techniques for detecting outliers in multivariate data are unreliable, and robust techniques have been developed to address this problem. Egan and Morgan[81] comprehensively reviewed the extant methods for detecting multivariate outliers and developed the robust smallest-half volume (SHV) method for multivariate outlier detection. The SHV is conceptually simple, faster to compute than other robust techniques, and robust for data having up to 25−45% outliers. The result of the application of the SHV outlier detection method to the WAbs dataset is shown in Figure 3, where a 99.9% confidence ellipse, based on the 50% of compounds in the WAbs dataset selected by SHV as being the most similar to each other, was overlaid on the plot in Figure 3. WAbs compounds outside the

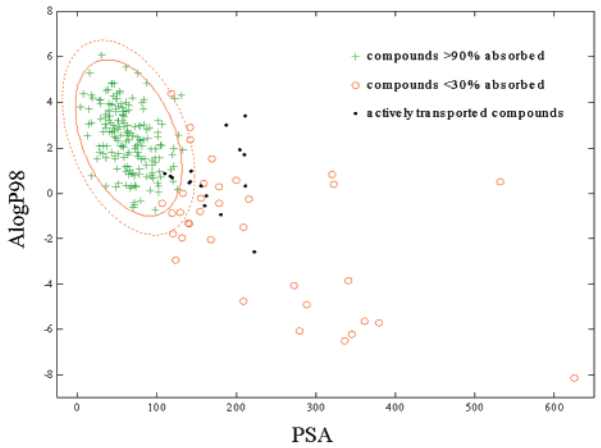

Figure 4. Plot of WAbs and PAbs datasets on PSA−AlogP98 axes with well-absorbed compounds which were identified as being actively transported represented as a separate class. The 95% and 99% (dotted) confidence ellipses, based on the remaining compounds in the WAbs dataset, are also plotted.

99.9% robust confidence ellipse, ordered by their approximate Mahalanobis distances, per Egan and Morgan,[81] were selected for closer examination.

As stated earlier, we did not consider active transport mechanisms when assembling the WAbs dataset. Either active transport mechanisms[82] caused compounds to be well-absorbed when they lay commingled with poorly-absorbed compounds in the region of AlogP98−PSA space or some other factor (e.g., charge) needs to be included in the model. Considerable evidence was found in the literature to support the hypothesis that active transport mechanisms were the cause. Ten of the outliers are antibacterial agents which are known to be actively transported: amoxicillin, cefaclor, cefadroxil, cefamandole, cefazolin, cefprozil, cephalexin, cephradine, doxycycline, and minocycline.[82−87] Evidence also exists for the involvement of carrier-mediated transport in the absorption of methotrexate and L-leucovorin,[88−91] as well as L-dopa.[92] Three glycosides (digitoxin, digoxin, and gitoxin) were identified as outliers in the dataset and are also actively transported.[93,94] Finally, rifampin was identified as an extreme outlier, which is interesting because one study has shown rifampin to pass efficiently in both directions through Caco-2 cells in a concentration-dependent, nonsaturable fashion, and the researchers concluded that this is suggestive of passive diffusion down a concentration gradient.[95] However, rifampin is a large, polar antimycotic (MW = 822, PSA = 211.9), and similar compounds are poorly-absorbed, suggesting that rifampin may be actively transported via a nonsaturable mechanism. Tsuji and Tamai's[82] review of active-transport mechanisms and substrates was used to check the entire WAbs dataset to determine if any other compounds were actively transported. Also, the compilation of *p*-glycoprotein (pGp) efflux substrates assembled by Seelig[96] was used to cross-check the PAbs dataset for pGp efflux substrates; only 1 of the 35 compounds (doxorubicin, a pGp inducer) included in the PAbs dataset was listed. Figure 4 plots the well-absorbed compounds identified as being actively transported with a different marker style to delineate them from all other compounds. The 95% and 99% confidence ellipses were computed for the WAbs dataset, excluding actively transported compounds. Note: The 99% confi-

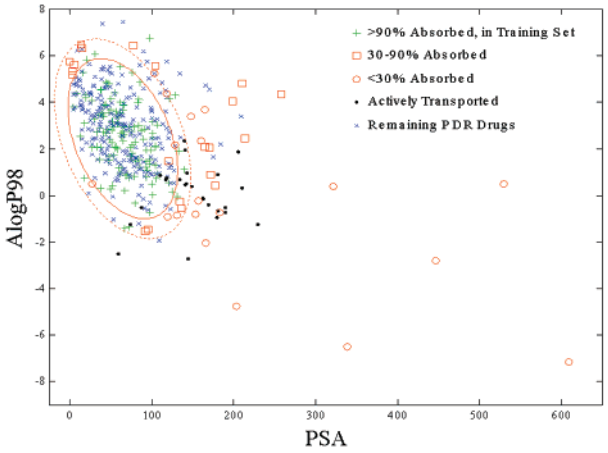

Figure 5. Plot of the PDR dataset on PSA−AlogP98 axes with the 95% and 99% (dotted) confidence ellipses from the model.

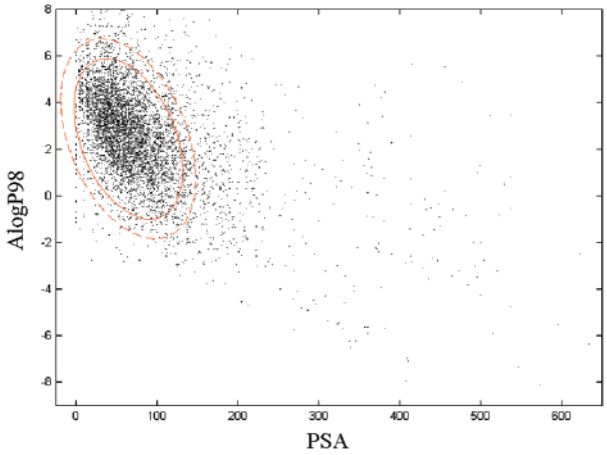

Figure 6. Plot of the CMC dataset on PSA−ALogP98 axes with the model 95% and 99% (dashed) confidence ellipses also shown.

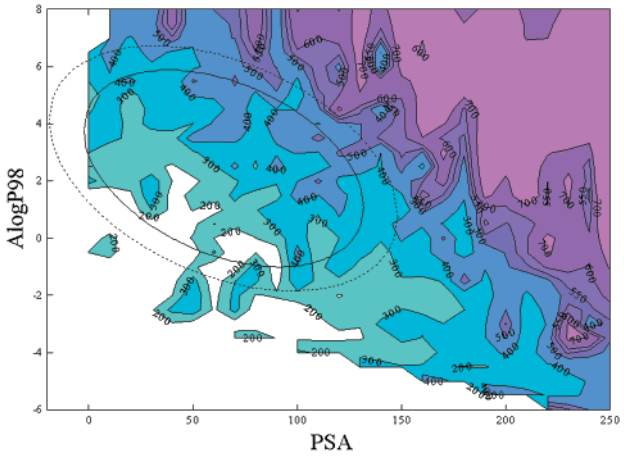

Figure 7. Contour plot of MW on PSA−AlogP98 axes for the CMC dataset. Model 95% and 99% (dotted) confidence ellipses are shown.

dence ellipse is bigger than the 95% confidence ellipse because to increase the probability that the ellipse contains more compounds, the space enclosed by the ellipse must increase.

Inspection of Figure 4 reveals several interesting features. The upper limit of PSA as defined by the 95% confidence ellipse is 131.6 Å^2, and the upper limit of PSA as defined by the 99% confidence ellipse is 148.1 Å^2; both of these PSA limits are similar to those reported in the literature for univariate PSA cutoffs.[23,97] The 95% confidence ellipse also defines an upper limit on AlogP98 (5.88) which decreases as PSA decreases, demonstrating the interaction between the two descriptors. Poorly-absorbed compounds hug the 95% confidence ellipse, and there appears to be a sharp change in absorption between the 95% and 99% confidence ellipses. This is consistent with literature reports of significant nonlinearities in the univariate relationships between lipophilicity and H-bonding ability, as discussed above.

The model was tested on three literature datasets. For the PDR dataset, 77.4% of the 438 orally delivered compounds were inside the 95% confidence ellipse and 87.4% were inside the 99% confidence ellipse. Compounds were examined and classified according to literature sources on absorption as to whether they are actively transported, moderately absorbed (30−90% absorbed), or poorly-absorbed (<30% absorbed), or if no precise determination has been made.[20,30,60−65,82] All compounds in the PDR dataset which were included in the WAbs dataset were considered as a separate class. Figure 5 plots the categorized PDR dataset and shows that the majority of the compounds which are actively transported or poorly to moderately absorbed are outside the region of chemical space considered to be statistically similar to the region occupied by well-absorbed compounds (≥90% absorbed). A large majority of compounds (88.6%) which could not be classified due to lack of explicit literature values are inside the model's 95% confidence ellipse. Excluding actively transported compounds, 81.4% of the remaining compounds in the PDR dataset were inside the 95% confidence ellipse and 90.1% of the remaining compounds in the PDR dataset were inside the 99% confidence ellipse. One would expect the majority of orally delivered drugs to be predicted to be well-absorbed, and these results are an excellent confirmation of that hypothesis.

Similar proportions of both the CMC and USAN/INN datasets were predicted to be well-absorbed. For the CMC dataset, 75.0% of the compounds are inside the model's 95% confidence ellipse and 83.5% of the compounds are inside the model's 99% confidence ellipse. For the USAN/INN dataset, 74.3% of the compounds are inside the 95% confidence ellipse and 82.9% of the compounds are inside the 99% confidence ellipse. Figure 6 plots the CMC dataset on the PSA−AlogP98 axes with the 95% and 99% confidence ellipses.

The CMC dataset also provides additional confirmation that the information contained in MW is already included in the PSA and AlogP98 descriptors. Figure 7 is a contour plot of MW on the PSA−AlogP98 axes for the CMC dataset. At low PSA and high AlogP98, MW is still only in the range 400−500, and at high PSA and low AlogP98, MW is in the range 200−400; both MW ranges are in the range generally considered acceptable for small-molecule drug design.[18] The 95% and 99% confidence ellipses create acceptable hydrophilicity and lipophilicity bounds individually on PSA and AlogP98, where the MW ranges are still low. Moreover, the confidence ellipses bound the interaction between hydrophilicity and lipophilicity, at moderate PSA and moderate AlogP98, where MW increases unacceptably,

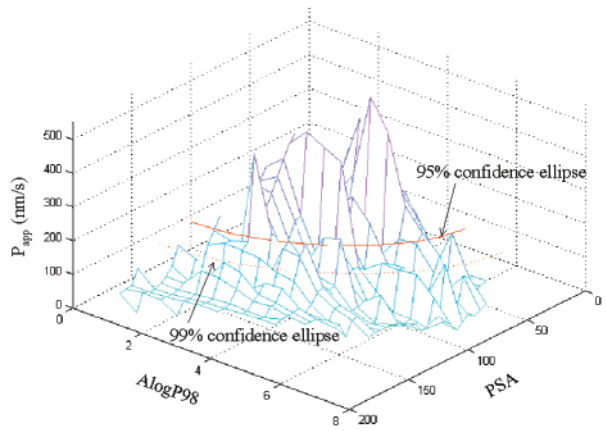

Figure 8. 3-D smoothed surface of Caco-2 P_{app} on PSA−AlogP98 axes for 446 Pharmacopeia, Inc. compounds with higher lipophilicity. 95% and 99% confidence ellipses are offset higher to make them visible.

due to the relationship between MW and hydrophilicity and lipophilicity.

The ability of the pattern recognition model to correctly classify known orally available and drug-like compounds as well-absorbed is promising. However, the poorly-absorbed compounds in the PAbs dataset have high PSA and low AlogP98 values. Therefore, we selected Pharmacopeia, Inc. compounds which covered the more lipophilic region of chemical space and for which Caco-2 permeability had been determined to further validate the utility of the PSA−AlogP98 pattern recognition model. Figure 8 plots the P_{app} surface, smoothed using triangle-based cubic interpolation, against PSA−AlogP98 with the relevant portions of the model's 95% and 99% confidence ellipses offset to make them visible. Figure 8 clearly shows further experimental evidence of the sharp, nonlinear drop in permeability as a function of both lipophilicity and hydrophilicity, as estimated by PSA and AlogP98. When the Caco-2 cell permeability data is assessed in terms of where the compounds lie in the PSA−AlogP98 model space, the usefulness of the model is apparent (Table 1): 61.8% of the highly permeable compounds (P_{app} > 100 nm/s) are inside the 95% confidence ellipse, and 91.5% of the moderately to highly permeable compounds (P_{app} > 34 nm/s) are within the 99% confidence ellipse. Only 20.6% of the poorly permeable compounds are inside the 95% confidence ellipse. The permeability is mixed for the compounds located between the 95% and 99% confidence ellipses of the model, a region of sharp change in permeability, based on all available evidence. Despite the inherent error in the calculated AlogP98 values and the moderate imprecision of the Caco-2 cell permeability assay, the model performs reasonably well.

Discussion and Conclusions

In this paper we described the development of a general computational model for human passive intestinal absorption. The descriptors chosen for inclusion in the model were AlogP98 and PSA. This choice was based on consideration of the physical processes involved in membrane permeability and the fact that PSA provides a reference point for AlogP98. It was critical that PSA serve as a reference point for AlogP98 since the latter descriptor is a ratio of lipophilicity to hydrophilicity which contains no information on the absolute measure of either factor. Larger datsets (several hundred compounds or larger) were collected, so as to cover as thoroughly as possible the chemical space related to passive intestinal absorption (defined by PSA and AlogP98). Due to the variability in published Caco-2 cell permeability assay results, we used literature data on compounds known to be well- and poorly-absorbed in humans. Because of the categorical nature of the data, we chose to use a statistical pattern recognition model.

Table 1. Percentages of 446 Pharmacopeia, Inc. Compounds in Each of Three Permeability Categories vs Three Regions of the Absorption Model

model region/P_{app}	<34 nm/s: poor	34−100 nm/s: moderate	100 nm/s: high
inside 95% confidence ellipse	20.6%	29.5%	61.8%
95−99% confidence ellipse	38.7%	41.9%	29.7%
outside 99% confidence ellipse	40.6%	28.7%	8.5%

The resultant model has several advantages. The descriptors PSA and AlogP98 are physically meaningful and easily related to structure, making them relatively straightforward for a medicinal chemist to interpret. Active transport and efflux mechanisms were accounted for in the model-building process, greatly reducing the potential for those mechanisms to bias the model. Robust outlier detection enabled actively transported compounds to be identified, and a check of relevant literature found only one of the poorly-absorbed compounds to be a known substrate for *p*-glycoprotein efflux. The dataset used to build the model was sufficiently large to provide good coverage of chemical space related to passive intestinal absorption. The interaction between hydrophilicity and lipophilicity was also taken into account and was necessary to discriminate between well- and poorly-absorbed compounds. Extensive validation of the model on known orally delivered drugs, drug-like molecules, and Pharmacopeia, Inc. compounds which had been assayed for Caco-2 cell permeability demonstrated a reasonably good rate of successful predictions (74−92%, depending on dataset and criterion).

This approach does, however, have a number of drawbacks. Ideally, we would like a quantitative model to more exactly handle the nonlinear relationships between absorption and hydrophilicity and lipophilicity, not a pattern recognition model which provides a yes/no answer. Human absorption data was used to create the model, and significant quantities of additional human data would be very difficult to obtain. However, assays such as the Caco-2 cell permeability assay can provide additional information, provided the variability issues are addressed. Several factors were not considered, including the effects of charge/dissociation and the changes in conformation due to solvent interactions, which may be significant, especially for particular series. Solubility was only considered implicitly, in that the WAbs dataset compounds had to be sufficiently soluble to dissolve for absorption measurements to be made. Calculated log *P* was included as a descriptor, and calculated log *P* values have been shown by Ghose et al.[19] to have sufficient error (rmse 0.26−1.17 for ClogP, rmse 0.39−0.75 for AlogP98, depending on number of atoms) to distort the model predictions. This is because permeability changes sharply in certain regions and the

reported errors for ClogP and AlogP98 are sufficient to push the compounds outside the confidence ellipse(s) we are using as boundaries for those regions of sharp change.

Computational models for the prediction of ADME properties have clearly generated significant interest, due to their potential to greatly reduce both the time and cost required to discover and develop a new drug. Computational ADME models, such as the passive intestinal absorption model described herein, will only be useful if they can predict with reasonable accuracy (70%+ correct) the ADME property of interest for a broad range of compounds and provide insight into the nature of the structure−property relationship. To accomplish this goal requires significant amounts of accurate, precise, and consistent data obtained over large regions of chemical space strongly related to the ADME property of interest. Although there are large quantities of experimental ADME observations in the literature, the differences in experimental methodologies, inter- and intralaboratory variability, and focus of those experiments on particular series of interest cause the actual amount of information in the literature data to be much lower than it appears to be, and far more difficult to extract.

Regardless of how pharmaceutical companies approach computational ADME modeling (in-house, via consortia,[98] or through the purchase of third-party software), the critical issue in modeling ADME properties is obtaining more and better data. Traditionally, pharmaceutical companies have concentrated on the project immediately at-hand, and systematic, comprehensive exploration of structure−ADME property relationships has simply not occurred. The use of small datasets, in our opinion, runs multiple risks: (1) missing a relationship because a region of chemical space simply was not covered; (2) discovering a relationship but missing interrelationships; and (3) not having enough data to tell if an observed relationship (or interrelationship) actually does exist or is merely an artifact of the small sample size. Fully realizing the potentially enormous benefits of computational ADME modeling will require the collection of large experimental datasets containing consistent, accurate, and precise ADME properties.

Acknowledgment. The authors thank Anfan Wu and Dr. Kirk McMillan (In Vitro Pharmacology Group, Pharmacopeia Laboratories, Pharmacopeia, Inc.) and I-Ping Cheng (CIDD, Pharmacopeia, Inc.) for their assistance with the Caco-2 cell permeability data. W.J.E. thanks Dr. Marie C. Egan (The College of New Jersey) for her insightful comments.

References

(1) Arlington, S. Pharma 2005-An Industrial Revolution in R&D. *Pharm. Exec.* **2000**, *January*, 74−84.
(2) Miller, H. I. Rising Costs Hold Up Drug Discovery. *Nature* **1998**, *395*, 835.
(3) Sinko, P. J. Drug Selection in Early Drug Development: Screening for Acceptable Pharmacokinetic Properties Using Combined in vitro and Computational Approaches. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 42−48.
(4) DiMasi, J. A. Success Rate for New Drugs Entering Clinical Testing in the United States. *Clin. Pharmacol. Ther.* **1995**, *58*, 1−14.
(5) Caldwell, J.; Gardner, I.; Swales, N. An Introduction to Drug Disposition: The Basic Principles of Absorption, Distribution, Metabolism, and Excretion. *Toxicol. Pathol.* **1995**, *23*, 102−114.
(6) Lin, J. H.; Lu, A. Y. H. Role of Pharmacokinetics and Metabolism in Drug Discovery and Development. *Pharmacol. Rev.* **1997**, *49*, 403−449.
(7) Smith, D. A.; van de Waterbeemd, H. Pharmacokinetics and Metabolism in Early Drug Discovery. *Curr. Opin. Chem. Biol.* **1999**, *3*, 373−378.
(8) Lipper, R. A. E Pluribus Product. *Modern Drug Discovery* **1999**, *2*, 55−60.
(9) Peet, N. P. Selecting Leads with Pharmacokinetic Data. *Modern Drug Discovery* **1999**, *2*, 21.
(10) Rodrigues, A. D. Preclinical Drug Metabolism in the Age of High-Throughput Screening: An Industrial Perspective. *Pharm. Res.* **1997**, *14*, 1504−1510.
(11) Watt, A. P.; Morrison, D.; Evans, D. C. Approaches to Higher-throughput Pharmacokinetics (HTPK) in Drug Discovery. *Drug Discovery Today* **2000**, *5*, 17−24.
(12) Gibbons, J. A.; Taylor, E. W.; Braeckman, R. A. ADME/PK Assays in Screening for Orally Active Drug Candidates. In *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; Gordon, E. M., Kerwin, J. F., Eds.; Wiley-Liss, Inc.: New York, 1998; pp 453−474.
(13) Tarbit, M. H.; Berman, J. High-throughput Approaches for Evaluating Absorption, Distribution, Metabolism, and Excretion Properties of Lead Compounds. *Curr. Opin. Chem. Biol.* **1998**, *2*, 411−416.
(14) Smith, D. A.; Jones, B. C.; Walker, D. K. Design of Drugs Involving the Concepts and Theories of Drug Metabolism and Pharmacokinetics. *Med. Res. Rev.* **1996**, *3*, 243−266.
(15) Caldwell, G. W. Compound Optimization in Early- and Late-Phase Drug Discovery: Acceptable Pharmacokinetic Properties Utilizing Combined Physicochemical, in vitro, and in vivo Screens. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 30−41.
(16) Prentis, R. A.; Lis, Y.; Walker, S. R. Pharmaceutical Innovation by the Seven UK-owned Pharmaceutical Companies (1964−1985). *Br. J. Clin. Pharmacol.* **1988**, *25*, 387−396.
(17) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.
(18) 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 Deliv. Rev.* **1997**, *23*, 3−25.
(19) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(20) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14*, 568−571.
(21) Palm, K.; Luthman, K.; Ungell, A.; 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*, 5382−5392.
(22) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.
(23) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J. Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(24) Stenberg, P.; Luthman, K.; Artursson, P. Prediction of Membrane Permeability to Peptides from Calculated Dynamic Molecular Surface Properties. *Pharm. Res.* **1999**, *16*, 205−212.
(25) Camenisch, G.; Alsenz, J.; van de Waterbeemd, H.; Folkers, G. Estimation of Permeability by Passive Diffusion through Caco-2 Cell Monolayers Using Drugs' Lipophilicity and Molecular Weight. *Eur. J. Pharm. Sci.* **1998**, *6*, 313−319.
(26) Camenisch, G.; Folkers, G.; van de Waterbeemd, H. Shape of Membrane Permeability−Lipophilicity Curves: Extension of Theoretical Models with an Aqueous Pore Pathway. *Eur. J. Pharm. Sci.* **1998**, *6*, 321−329.
(27) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480−490.
(28) Norinder, U.; Osterberg, T.; Artursson, P. Theoretical Calculation and Prediction of Caco-2 Cell Permeability Using MolSurf Parametrization and PLS Statistics. *Pharm. Res.* **1997**, *14*, 1786−1791.

(29) Norinder, U.; Osterberg, T.; Artursson, P. Theoretical Calculation and Prediction of Intestinal Absorption of Drugs in Humans Using MolSurf Parametrization and PLS Statistics. *Eur. J. Pharm. Sci.* **1999**, *8*, 49−56.
(30) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(31) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernas, H.; Karlen, A. Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters. A Multivariate Data Analysis Approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(32) Alifrangis, L. H. C., I. T.; Berglund, A.; Sandberg, M.; Hovgaard, L.; Frokjaer, S. Structure−Property Model for Membrane Partitioning of Oligopeptides. *J. Med. Chem.* **2000**, *43*, 103−113.
(33) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.
(34) Chan, O. H.; Stewart, B. H. Physicochemical and Drug-delivery Considerations for Oral Drug Bioavailability. *Drug Discovery Today* **1996**, *1*, 461−473.
(35) Navia, M. A.; Chaturvedi, P. R. Design Principles for Orally Bioavailable Drugs. *Drug Discovery Today* **1996**, *1*, 179−189.
(36) Conradi, R. A.; Burton, P. S.; Borchardt, R. T. Physicochemical and Biological Factors that Influence a Drug's Cellular Permeability by Passive Diffusion. *Methods Princ. Med. Chem.* **1996**, *4*, 233−252.
(37) Camenisch, G.; Folkers, G.; van de Waterbeemd, H. Review of Theoretical Passive Drug Absorption Models: Historical Background, Recent Developments and Limitations. *Pharm. Acta Helv.* **1996**, *71*, 309−327.
(38) Kararli, T. T. Gastrointestinal Absorption of Drugs. *Crit. Rev. Ther. Drugs Carrier Syst.* **1989**, *6*, 39−86.
(39) Singer, S. J.; Nicolson, G. L. The Fluid Mosaic Model of the Structure of Cell Membranes. *Science* **1972**, *175*, 720−731.
(40) Lennernäs, H. Human Jejunal Effective Permeability and Its Correlation with Preclinical Drug Absorption Models. *J. Pharm. Pharmacol.* **1997**, *49*, 627−638.
(41) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and their Uses. *Chem. Rev.* **1971**, *71*, 525−616.
(42) Hansch, C.; Dunn, W. J. Linear Relationships between Lipophilic Character and Biological Activity of Drugs. *J. Pharm. Sci.* **1972**, *61*, 1−19.
(43) Hansch, C.; Clayton, J. M. Lipophilic Character and Biological Activity of Drugs. II. Parabolic Case. *J. Pharm. Sci.* **1973**, *62*, 1−21.
(44) Carrupt, P.; Testa, B.; Gaillard, P. Computational Approaches to Lipophilicity: Methods and Applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol. 11, pp 241−315.
(45) Martin, Y. C. A Practitioner's Perspective of the Role of Quantitative Structure−Activity Analysis in Medicinal Chemistry. *J. Med. Chem.* **1981**, *24*, 229−237.
(46) Wils, P.; Warnery, A.; Phung-ba, V.; Legrain, S.; Scherman, D. High Lipophilicity Decreases Drug Transport Across Intestinal Epithelial Cells. *J. Pharmacol. Exp. Ther.* **1994**, *269*, 654−658.
(47) Gobas, F. A.; Lahittette, J. M.; Garofolo, G.; Shiu, W.; Y.; MacKay, D. A Novel Method for Measuring Membrane-Water Partition Coefficients of Hydrophobic Organic Chemicals: Comparison with 1-Octanol−Water Partitioning. *J. Pharm. Sci.* **1988**, *77*, 265−272.
(48) Wright, E. M.; Diamond, J. M. Patterns of Nonelectrolyte Permeability. *Proc. R. Soc. B* **1969**, *172*, 227−271.
(49) Diamond, J. M.; Wright, E. M. Molecular Forces Governing Nonelectrolyte Permeation through Cell Membranes. *Proc. R. Soc. B* **1969**, *172*, 273−316.
(50) Palm, K.; Luthman, K.; Ros, J.; Gråsjo, J.; Artursson, P. Effect of Molecular Charge on Intestinal Epithelial Drug Transport: pH-Dependent Transport of Cationic Drugs. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 435−443.
(51) Delgado, J. N., Remers, W. A., Eds. *Wilson and Gisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry*, 10th ed.; Lippincott Raven Publishers: New York, 1998; Appendix A.
(52) Artursson, P.; Palm, K.; Luthman, K. Caco-2 Monolayers in Experimental and Theoretical Predictions of Drug Transport. *Adv. Drug Deliv. Rev.* **1996**, *22*, 67−84.
(53) Gan, L. L.; Thakker, D. R. Applications of the Caco-2 Model in the Design and Development of Orally Active Drugs: Elucidation of Biochemical and Physical Barriers Posed by the Intestinal Epithelium. *Adv. Drug Deliv. Rev.* **1997**, *23*, 77−98.
(54) Artursson, P.; Karlsson, J. Correlation Between Oral Drug absorption in Humans and Apparent Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cells. *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880−885.
(55) Yee, S. In Vitro Permeability Across Caco-2 Cells (Colonic) Can Predict In Vivo (Small Intestine) Absorption in Man − Fact or Myth. *Pharm. Res.* **1997**, *14*, 763−766.
(56) Pade, V.; Stavchansky, S. Link Between Drug Absorption Solubility and Permeability Measurements in Caco-2 Cells. *J. Pharm. Sci.* **1998**, *87*, 1604−1607.
(57) Chong, S.; Dando, S. A.; Soucek, K. M.; Morrison, R. A. In Vitro Permeability Through Caco-2 Cells is not Quantitatively Predicitve of In Vivo Absorption for Peptide-like Drugs Absorbed via the Dipeptide Transporter System. *Pharm. Res.* **1996**, *13*, 120−123.
(58) Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating Partitioning and Caco-2 Cell Permeability of Structurally Diverse Small Molecular Weight Compounds. *Pharm. Res.* **1998**, *15*, 1490−1494.
(59) Irvine, J. D.; Takahashi, L.; Lockhart, K.; Cheong, J.; Tolan, J. W.; Selick, H. E.; Grove, R. E. MDCK (Madin-Darby Canine Kidney) Cells: A Tool for Membrane Permeability Screening. *J. Pharm. Sci.* **1999**, *88*, 28−33.
(60) Chiou, W. L.; Barve, A. Linear Correlation of the Fraction of Oral Dose Absorbed of 64 Drugs between Humans and Rats. *Pharm. Res.* **1998**, *15*, 1792−1795.
(61) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J. Med. Chem.* **1998**, *41*, 1007−1010.
(62) Sietsema, W. K. The Absolute Oral Bioavailability of Selected Drugs. *Int. J. Clin. Pharmacol., Ther. Toxicol.* **1989**, *27*, 179−211.
(63) Gilman, A. G., Hardman, J. G., Limbird, L. E., Molinoff, P. B., Ruddon, R. W., Eds. *The Pharmacological Basis of Therapeutics*, 9th ed.; McGraw-Hill: New York, 1996.
(64) Anderson, P. O., Knoben, J. E., Troutman, W. G., Eds. *Handbook of Clinical Drug Data*, 9th ed.; Appleton & Lange: Stamford, 1999.
(65) *The PDR Electronic Library*, release 1999.2; Medical Economics Co., Inc.: Montvale, NJ, 1999.
(66) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(67) Meunier, V.; Bourrie, M.; Berger, Y.; Fabre, G. The Human Intestinal Epithelial Cell Line Caco-2; Pharmacological and Pharmacokinetic Applications. *Cell Biol. Toxicol.* **1995**, *11*, 187−194.
(68) Briske-Anderson, M. J.; Finley, J. W.; Newman, S. M. The Influence of Culture Time and Passage Number on the Morphological and Physiological Development of Caco-2 Cells. *Proc. Soc. Exp. Biol. Med.* **1997**, *214*, 248−257.
(69) Brayden, D. J. Human Intestinal Epithelial Cell Monolayers as Prescreens for Oral Drug Delivery. *Pharm. News* **1997**, *4*, 11−15.
(70) Yamashita, S.; Tanaka, Y.; Endoh, Y.; Taki, S.; Toshiyasu, N.; Tanekazu, S.; Sezaki, H. Analysis of Drug Permeation Across Caco-2 Monolayer: Implication for Predicting in vivo Drug Absorption. *Pharm. Res.* **1997**, *14*, 486-491.
(71) Taylor, E. W.; Gibbons, J. A.; Braeckman, R. A. Intestinal Absorption Screening of Mixtures from Combinatorial Libraries in the Caco-2 Model. *Pharm. Res.* **1997**, *14*, 572−577.
(72) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(73) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; John Wiley & Sons: New York, 1986.
(74) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(75) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.
(76) Gould, S. J. *The Mismeasure of Man*; W. W. Norton & Co.: New York, 1996.
(77) Buchwald, P.; Bodor, N. Octanol−Water Partition: Searching for Predictive Models. *Curr. Med. Chem.* **1998**, *5*, 353−380.
(78) Rencher, A. C. *Methods of Multivariate Analysis*; John Wiley & Sons: New York, 1995.
(79) Jackson, J. E. Principal Components and Factor Analysis: Part I − Principal Components. *J. Quality Technol.* **1980**, *12*, 200−213.
(80) Barnett, V.; Lewis, T. *Outliers in Statistical Data*, 3rd ed.; John Wiley & Sons: New York, 1994; p 7.
(81) Egan, W. J.; Morgan, S. L. Outlier Detection in Multivariate Analytical Chemical Data. *Anal. Chem.* **1998**, *70*, 2372−2379.
(82) Tsuji, A.; Tamai, I. Carrier-Mediated Intestinal Transport of Drugs. *Pharm. Res.* **1996**, *13*, 963−977.
(83) Tsuji, A. Intestinal Uptake of β-lactam Antibiotics and its Relationship to Peptide Transport. *Adv. Biosci.* **1987**, *65*, 125−131.

(84) Bretschneider, B.; Brandsch, M.; Neubert, R. Intestinal Transport of β-lactam Antibiotics: Analysis of the Affinity at the H+/Peptide Symporter (PEPT1), the Uptake into Caco-2 Cell Monolayers and the Transepithelial Flux. *Pharm. Res.* **1999**, *16*, 55−61.
(85) Meshali, M. M.; Attia, I. E. Transport Mechanism of Some Naturally Occurring Tetracyclines Across Everted Rat Gut. *Can. J. Pharm. Sci.* **1978**, *13*, 42−45.
(86) Barcina, Y.; Ilundain, A.; Larralde, J. Effect of Amoxicillin, Cephalexin, and Tetracycline-hydrochloride on Intestinal L-leucine Transport in the Rat *In Vivo*. *Drug-Nutr. Interact.* **1988**, *5*, 283−288.
(87) Snyder, N. J.; Tabas, L. B.; Berry, D. M.; Duckworth, D. C.; Spry, D. O.; Dantzig, A. H. Structure−activity Relationship of Carbacephalosporins and Cephalosporins: Antibacterial Activity and Interaction with the Intestinal Proton-Dependent Dipeptide Transport Carrier of Caco-2 Cells. *Antimicrob. Agents Chemother.* **1997**, *41*, 1649−1657.
(88) Dedrick, R. L.; Zaharko, D. S.; Lutz, R. J. Transport and Binding of Methotrexate In Vivo. *J. Pharm. Sci.* **1973**, *62*, 882−890.
(89) Chungi, V. S.; Bourne, D. W. A.; Dittert, L. W. Competitive Inhibition between Folic Acid and Methotrexate for Transport Carrier in the Rat Small Intestine. *J. Pharm. Sci.* **1979**, *68*, 1552−1553.
(90) Cercos-Fortea, T.; Casabo, V. G.; Nacher, A.; Cejudo-Ferragud, E.; Polache, A.; Merino, M. Evidence of Competitive Inhibition of Methotrexate Absorption by Leucovorin Calcium in Rat Small Intestine. *Int. J. Pharm.* **1997**, *155*, 109−119.
(91) Honscha, W.; Petzinger, E. Cloning and Characterization of the Hepatocellular Methotrexate Transport System. *Nova Acta Leopold.* **1998**, *78*, 135−140.
(92) Shindo, H.; Komai, T.; Kawai, K. Studies on the Metabolism of D- and L- Isomers of 3,4-dihydroxyphenylalanine (DOPA) V. Mechanism of Intestinal Absorption of D- and L-DOPA-14C in Rats. *Chem. Pharm. Bull. (Tokyo)* **1973**, *21*, 2031−2038.
(93) Greenberger, N. J.; Caldwell, J. H. Intestinal Absorption of Tritium-Labeled-Digitalis Glycosides in Experimental Animals and Man. *Basic Clin. Pharmacol. Digitalis, Proc. Symp.* **1972**, Meeting Date 1970, 15−47.
(94) Braun, W.; Damm, K. H. Drug Interactions in Intestinal Absorption of 3H-Digitoxin in Rats. *Experientia* **1976**, *32*, 613−614.
(95) Ranaldi, G.; Islam, K.; Sambuy, Y. Epithelial Cells in Culture as a Model for the Intestinal Transport of Antimicrobial Agents. *Antimicrob. Agents Chemother.* **1992**, *36*, 1374−1381.
(96) Seelig, A. A General Pattern for Substrate Recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.
(97) Palm, K.; Luthman, K.; Ungell, A.; Strandlund, G.; Artursson, P. Correlation of Drug Absorption with Molecular Surface Properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.
(98) Johnson, D. The Discovery-Development Interface has become the New Interfacial Phenomenon. *Drug Discovery Today* **1999**, *4*, 535−536.

JM000292E