Part Four
Computational Approaches to Drug Absorption and Bioavailability

14
Calculated Molecular Properties and Multivariate Statistical Analysis

Ulf Norinder

Abbreviations

2D	Two dimensional
3D	Three dimensional
AD	Applicability domain
ADME	Absorption, distribution metabolism, and excretion
ADMET	Absorption, distribution metabolism, excretion, and toxicity
ANN	Artificial neural network
ARD	Automatic relevance determination
BCI	Bernard chemical information
BCUT	Burden, CAS, University of Texas descriptors
BNN	Bayesian neural network
C4.5	Decision trees using information entropy
CART	Classification and regression tree
Clog*P*	Calculated partition coefficient between octanol and water
CoMFA	Comparative molecular field analysis
CV	Cross-validation
DECORATE	Diverse ensemble creation by oppositional relabeling of artificial training examples
F_a	Fraction absorbed
FFD	Fractional factorial design
FIRM	Formal inference-based recursive modeling
FLAPs	Fingerprints for ligands and proteins
FNHS	Fractional negative hydrophobic surface area
FPHS	Fractional positive hydrophobic surface area
GAs	Genetic algorithms
GP	Genetic programming
G-REX	Genetic rule extraction
hERG	Human ether-a-go-go related gene
HMLP	Heuristic molecular lipophilicity potential

LDA	Linear discriminant analysis
LOO-CV	Leave-one-out cross-validation
LMO-CV	Leave-multiple-out cross-validation
MACC	Maximum auto- and cross-correlation
MIF	Molecular interaction field
MLP	Molecular lipophilicity potential
MLR	Multiple linear regression
MOE	Molecular operating environment
NN	Neural network
OD	Onion design
P-gp	P-Glycoprotein
PCA	Principal component analysis
RDS	Rule discovery system
PLS	Partial least square projection to latent structures
QSPR	Quantitative structure–property relationship
RNH	Relative hydrophilicity
RPH	Relative hydrophobicity
PNHS	Partial negative hydrophobic surface area
PPHS	Partial positive hydrophobic surface area
PSA	Polar surface area
QSAR	Quantitative structure–activity relationship
RP	Recursive partitioning
SFD	Space-filling design
SVM	Support vector machine
TPSA	Topological polar surface area
WDI	World Drug Index
WHIM	Weighted holistic invariant molecular descriptors
WNHS	Weighted negative hydrophobic surface area
WPHS	Weighted positive hydrophobic surface area

Symbols

A%	Percentage absorbed
A	Abraham hydrogen-bond acidity parameter
B	Abraham hydrogen-bond basicity parameter
C_d	Hydrogen-bond donor factor
C_a	Hydrogen-bond acceptor factor
δ, δ^v	Kier–Hall molecular connectivity chi parameter
E	Abraham excess molar refraction parameter
log P	$\log_{10}$ of partition coefficient between octanol and water
S	Dipolarity/polarizability solute–solvent interactions
V	McGowan characteristic volume
q^2	Cross-validated coefficient of determination
r^2	Coefficient of determination (correlation coefficient)

14.1 Introduction

To derive statistically good and predictive models, there are some aspects to consider. The investigated data set should have a reasonable spread with respect to continuous target values, for example, biological activities or some ADME-related property, of approximately three orders of magnitude or more, and the target values should also be reasonably well distributed. For deriving good and predictive classification models, the investigated classes should either be well balanced from the start, that is, the number of objects in each class are approximately the same, or should be balanced during the analysis by some appropriate weighting scheme. However, there are additional requirements that will need attention to consider a derived model robust with good forecasting ability. The investigated objects, for example, chemical structures, need to be well described for the statistical analysis to find adequate information among the collected independent variables (descriptors) to correlate to the corresponding dependent variable (target value). Today, there exist a large number of different descriptors as well as programs, for example, Dragon [1], Molconn-Z [2], MOE [3], Sybyl [4], and others, with which to calculate these variables. It is easy to rapidly calculate several thousands of descriptors for relatively large data set in the order of 10 k and upward. What kinds of descriptors are useful for modeling ADME properties? How correlated are the variables? Should 2D and/or 3D descriptors be utilized? Does the large magnitude of descriptors used has implications for the choice of statistical method or methods to be employed for analysis? What is the applicability domain (AD) of the derived model? How can this domain be quantified and used to advise users about the limitations of the model in question?

This chapter will try to answer some of these questions and investigate various approaches to derive statistically sound, robust, and predictive *in silico* models.

14.2 Calculated Molecular Descriptors

14.2.1 2D-Based Molecular Descriptors

Among the advantages with 2D-based descriptors are their rapid speed of computation for large sets of compounds and that they do not require 3D structures. Thus, these descriptors avoid the problem and compute times associated with 3D structure generation and conformational analysis, even though there are programs available that generate reliable 3D structures, for example, CORINA [5].

The 2D-based descriptors are sometimes divided into different types of descriptors such as constitutional, fragment, and functional group-based as well as topological descriptors.

14.2.1.1 Constitutional Descriptors

The constitutional descriptors are typically descriptors such as molecular weight, the number of various *x*-membered rings, the number of different types of atoms – for example, atoms of carbon, oxygen, nitrogen, and different halogens – and bonds, for example, single, double, triple, and aromatic. These kinds of descriptors have been used as one part of the structure description for modeling different ADMET end points [6–11]. Particularly, descriptors related to nitrogen and oxygen atoms have been found to be useful in deriving good ADME models since these descriptors capture the importance of hydrogen bonding for absorption and solvation processes. Another constitutional descriptor that has been frequently used is the number of rotatable bonds. This parameter is an attempt to easily obtain a crude estimation of entropy. In addition, incorporation of descriptors such as counts of various *x*-membered rings may provide important information regarding the influence of molecular self-association, for example, π–π interactions, in problems related to solubility [12].

A significant advantage of using constitutional descriptors is the ease of interpretation. It is straightforward for a researcher to understand the impact of these descriptors on derived statistical structure–property models.

14.2.1.2 Fragment- and Functional Group-Based Descriptors

The fragment and functional group based descriptors also represent a large and diverse number of available descriptors, and the division between these two sets of descriptors is rather fuzzy. These types of descriptors are also frequently called fingerprints, bits, or keys; for example, Scitegic fingerprints [13], MDL keys [14], and so on. This group of descriptors can vary significantly in the number of generated descriptors, the size of fragments identified, and the technique employed to store the descriptors. One may distinguish between two major approaches: a predefined set of patterns to be identified, also called a dictionary-based set, or a set of patterns that will vary depending upon the set of chemical structures that is under investigation. The former types of descriptors are generated using BCI fingerprints [15], Leadscope fingerprints [16], or MDL keys [17]. Many times user-defined fingerprints are of the same kind, for example, the Bursi alerts for toxicological screening [18].

The Leadscope fingerprints (the set contains some 27 k descriptors) vary markedly in size. These fingerprints cover structural patterns from small functional groups to rather large substructural moieties, whereas the Dragon functional group fingerprints (154 descriptors) [1] are restricted to identifying mostly small function groups, for example, ketones, amides carboxylic acids, esters, and alcohols. The descriptor sets generated using Daylight [19], Unity [4], or Scitegic fingerprints [13] represent the latter kinds of fingerprints, that is, the *in situ*-generated patterns. These fingerprints are usually "hashed" onto a fingerprint vector of predefined length, many times of lengths 512, 1024, or 2048, using a pseudorandom number generator. Owing to the hashing function, as well as the chosen size of the fingerprint bit vector, it is not guaranteed that different fingerprints would not be assigned to the same bit. This, in turn, means that the interpretability of such fingerprints can be lost. Nevertheless, these descriptors still contain important information for ADMET modeling and may, many times, be quite useful for deriving statistical models with significant predictability [7–11, 20–22].

The general problem with using fingerprints resides in their binary nature, that is, either present or absent in a particular (sub)structure. Unlike a continuous variable where inter- or extrapolations are possible for new structures to be predicted by an existing statistical model, a new structure to be predicted by a model based on the binary fingerprints may contain a large number of unrecognized fragments. This, in turn, may, at worst, mean that the new compound is poorly predicted by the latter fingerprint-based models.

14.2.1.3 **Topological Descriptors**

Again, there are a large number of topological descriptors available (Table 14.1) that can be calculated from the 2D structure (graph) of a compound.

Some of the most widely used topological descriptors are the so-called Kier–Hall indices [23] that describe connectivity (${}^{m}\mathrm{Chi}$, $m = 1$–3, where m represents the summation over atoms, bond paths, or bond fragments) and shapes (${}^{m}\mathrm{Kappa}$, $m = 1$–3, where m represents the topological paths of length m). These indices are based on two parameters δ and δ^{v}, respectively, where the former parameter is the difference between the number of sigma electrons and the count of hydrogen atoms, while the latter is the difference between the number of valence electrons and the count of hydrogen atoms for the particular atom in question. Other indices that have been frequently used in ADMET modeling are the Wiener, Balaban, and Zagreb indices [24]. The Wiener index, for instance, is related to the half-sum of the bond path lengths between each atom in a molecule (the sum of all off-diagonal elements of the path distance matrix in a molecule).

Another set of topological descriptors that have been found to be important in ADMET modeling is the BCUT (Burden, Cas, University of Texas) descriptors [25], which are eigenvalue-based parameters. They are computed as the highest and lowest eigenvalues from the hydrogen-depleted 2D connectivity matrix of the structure, where the diagonal elements of the original BCUT parameters have information regarding atomic charge, polarizability, and hydrogen-bonding ability, respectively. A large number of different BCUT-type descriptors have been developed over the years and, for instance, the Dragon software [1] computes over a hundred of these kinds of indices. The disadvantage with these descriptors is that they, in many cases, are difficult to interpret in terms of how should the current structures be modified to obtain a compound with better properties for the investigated target, for example, absorption or solubility. However, the topological descriptors are quite useful for computational screening of large virtual libraries (brute force approach) when a good statistical model has been developed.

Another set of topological descriptors is the electrotopological state indices (E-state indices) developed by Kier and Hall [26, 27]. These descriptors are based on the topological state of a particular atom with corrections for electronic interactions due to other atoms in the structure. This methodology originally devised for nonhydrogen atoms only has been extended to also include E-state indices for hydrogen atoms [28] and to also include atom-type E-state indices, for

Table 14.1 Selected topological descriptors.

Topological descriptor
Information index on molecular size
Total information index of atomic composition
Mean information index on atomic composition
First Zagreb index M_1
First Zagreb index by valence vertex degrees
Second Zagreb index M_2
Second Zagreb index by valence vertex degrees
Quadratic index
Narumi simple topological index (log)
Narumi harmonic topological index
Narumi geometric topological index
Total structure connectivity index
Pogliani index
Ramification index
Polarity number
Logarithm of product row sums (PRSs)
Average vertex distance degree
Mean square distance index (Balaban)
Schultz molecular topological index (MTI)
Schultz MTI by valence vertex degrees
Gutman molecular topological index
Gutman MTI by valence vertex degrees
Xu index
Superpendentic index
Wiener W index
Mean Wiener index
Reciprocal distance Wiener-type index
Harary H index
Quasi-Wiener index (Kirchhoff number)
First Mohar index T_{I1}
Second Mohar index T_{I2}
Hyperdistance-path index
Reciprocal hyperdistance-path index
Detour index
Hyperdetour index
Reciprocal hyperdetour index
Distance/detour index
All-path Wiener index
Wiener-type index from *Z*-weighted distance matrix (Barysz matrix)
Wiener-type index from mass-weighted distance matrix
Wiener-type index from van der Waals-weighted distance matrix
Wiener-type index from electronegativity-weighted distance matrix
Wiener-type index from polarizability-weighted distance matrix
Balaban J index
Balaban-type index from *Z*-weighted distance matrix (Barysz matrix)
Balaban-type index from mass-weighted distance matrix
Balaban-type index from van der Waals-weighted distance matrix
Balaban-type index from electronegativity-weighted distance matrix

Table 14.1 (*Continued*)

Balaban-type index from polarizability-weighted distance matrix
Connectivity index chi-0
Connectivity index chi-1 (Randic connectivity index)
Connectivity index chi-2
Connectivity index chi-3
Connectivity index chi-4
Connectivity index chi-5
Average connectivity index chi-0
Average connectivity index chi-1
Average connectivity index chi-2
Average connectivity index chi-3

example, for methyl groups, hydroxy and keto oxygens, respectively, and the corresponding atom-type E-state sums for various groups, as well as for different hydrogens, for example, hydrogen-bond donors and acceptors. The type E-state sums related to groups of hydrogen atoms have been found to correlate well with hydrogen-bonding properties [29].

One significant difference between many other topological descriptors and the E-state parameters is that the latter indices are much easier to interpret and, thus, capable of answering the question "Which are the next molecules to make?" in a relatively straightforward manner. These two aspects (computational speed and interpretability) make these descriptors quite attractive both for e-screening purposes and for having an interpretable model with which to focus the virtual library generation and for further pharmaceutical investigations or work. The sum of hydrogen-bonding donor- and acceptor-related E-state descriptors is well correlated with the corresponding HYBOT parameters (see Section 14.2.3.2 for further details) with r^2 values between 0.8 and 0.95 [30].

14.2.2
3D Descriptors

The 3D descriptors described in this section are the weighted holistic invariant molecular (WHIM) descriptors, the Jurs descriptors, and the GRID-based VolSurf and Almond descriptors, as well as pharmacophore fingerprints.

14.2.2.1 WHIM Descriptors

The WHIM descriptors are based on statistical indices calculated on the projections of atoms along principal axes [31–34]. There are different types of WHIM descriptors with the aim to incorporate 3D information regarding size, shape, symmetry, and atom distributions independent of molecular alignments.

The WHIM algorithm performs a principal component analysis (PCA) on the mean centered Cartesian coordinates of the molecule from a weighted covariance matrix of the atomic coordinates. The weights of this matrix are such properties as atomic mass, van der Waals volume, Sanderson atomic electronegativity, atomic

Table 14.2 List of examples of WHIM descriptors.

L1u: first component size directional WHIM index/unweighted
L2u: second component size directional WHIM index/unweighted
L3u: third component size directional WHIM index/unweighted
L1m: first component size directional WHIM index/weighted by atomic masses
L2m: second component size directional WHIM index/weighted by atomic masses
L3m: third component size directional WHIM index/weighted by atomic masses
L1v: first component size directional WHIM index/weighted by atomic van der Waals volumes
L2v: second component size directional WHIM index/weighted by atomic van der Waals volumes
L3v: third component size directional WHIM index/weighted by atomic van der Waals volumes
L1e: first component size directional WHIM index/weighted by atomic Sanderson electronegativities
L2e: second component size directional WHIM index/weighted by atomic Sanderson electronegativities
L3e: third component size directional WHIM index/weighted by atomic Sanderson electronegativities
L1p: first component size directional WHIM index/weighted by atomic polarizabilities
L2p: second component size directional WHIM index/weighted by atomic polarizabilities
L3p: third component size directional WHIM index/weighted by atomic polarizabilities
P1p: first component shape directional WHIM index/weighted by atomic polarizabilities
P2p: second component shape directional WHIM index/weighted by atomic polarizabilities
Tu: *T* total size index/unweighted
Tm: *T* total size index/weighted by atomic masses
Tv: *T* total size index/weighted by atomic van der Waals volumes
Te: *T* total size index/weighted by atomic Sanderson electronegativities
Tp: *T* total size index/weighted by atomic polarizabilities
Ts: *T* total size index/weighted by atomic electrotopological states

polarizability, and electrotopological state indices (for a list of selected WHIM descriptors see Table 14.2).

14.2.2.2 Jurs Descriptors

The so-called Jurs descriptors are 3D surface descriptions related to various total and fractional defined surfaces. They can be divided into to two parts: one electronic [35] and one hydrophobic [36]. The former set of descriptors is generated from partial positive and negative surface areas, total charge as well as atomic positively and negatively charged weighted surface areas, and various differential and fractional charged partial surface areas of the molecule (see Table 14.3).

The second set of descriptors describes hydrophobic surface properties of a molecule. As with the first set, the second set contains similar partial hydrophobic and partial hydrophilic surface area descriptors (PPHS-*x* and PNHS-*x*, respectively), differences in partial surface area descriptors (FPHS-*x* and FNHS-*x*), as well as total surface area weighted descriptors (WPHS-*x* and WNHS-*x*). In addition, two descriptors assessing the most hydrophobic atom and the most hydrophilic atom on the overall lipophilicity are also described (RPH and RNH). The atom-based fractional log *P* contributions used for calculations are those of Wildman and Crippen [37] and

Table 14.3 List of selected electronic Jurs descriptors.

PPSA-1: partial positive surface area
PNSA-1: partial negative surface area
PPSA-2: total charge weighted PPSA
PNSA-2: total charge weighted PNSA
PPSA-3: atomic charge weighted PPSA
PNSA-3: atomic charge weighted PNSA
DPSA-1: difference in charged partial surface areas [(PPSA-1) – (PNSA-l)]
DPSA-2: difference in charged partial surface areas [(PPSA-2) – (PNSA-2)]
DPSA-3: difference in charged partial surface areas [(PPSA-3) – (PNSA-3)]
FPSA-1: fractional charged partial surface areas
FNSA-1: fractional charged partial surface areas
FPSA-2: fractional charged partial surface areas
FNSA-2: fractional charged partial surface areas
FPSA-3: fractional charged partial surface areas
FNSA-3: fractional charged partial surface areas
WPSA-1: surface-weighted charged partial surface areas
WNSA-1: surface-weighted charged partial surface areas
WPSA-2: surface-weighted charged partial surface areas
WNSA-2: surface-weighted charged partial surface areas
WPSA-3: surface-weighted charged partial surface areas
WNSA-3: surface-weighted charged partial surface areas
RPCG: relative positive charge
RNCG: relative negative charge
RPCS: relative positive charged surface area
RNCS: relative negative charged surface area

computed on the solvent-accessible surface area using the SAVOL program [38] with a probe radius of 1.5 Å.

The Jurs descriptors have been found useful in modeling ADMET properties such as human intestinal absorption [25, 39] and toxicity [40].

14.2.2.3 VolSurf and Almond Descriptors

VolSurf and Almond descriptors are based on results from molecular interaction fields (MIFs) but do not explicitly require the alignment of the structures under investigation as a first step in the analysis. Although VolSurf and Almond descriptors use the same source of information, that is, the computed GRID MIFs, they differ significantly with respect to their underlying approaches. The VolSurf method [41,42] is created with the aim of predicting pharmacokinetic properties, for example, blood–brain barrier permeation [43]. VolSurf descriptors summarize the MIF information related to the size and shape of the molecule under investigation as well as to the size and shape of the hydrophilic and hydrophobic regions and the balance between the two regions. Almond descriptors are designed to characterize pharmacodynamic properties such as protein–ligand interactions. Almond descriptors are primarily aimed at the identification of optimal interaction sites and the description of the geometrical relationship between such sites by using a default set of GRID probes: DRY (hydrophobic), O (carbonyl oxygen, hydrogen-bond acceptor), and N1 (amide nitrogen, hydrogen-bond donor). A fixed number of GRID points from each MIF with respect to the GRID energy level and the internode distance between the two points are used. An autocorrelogram is generated via the MACC-2 (maximum auto- and cross-correlation) algorithm by storing only the highest pair-wise product of interaction energies between all pairs. The three default auto-correlograms are DRY–DRY (hydrophobic); O–O (hydrogen-bond donor); N1–N1 (hydrogen-bond acceptor). The three default cross-correlograms are DRY–O (hydrophobic–hydrogen-bond donor); DRY–N1 (hydrophobic–hydrogen-bond acceptor); O–N1 (hydrogen-bond donor–hydrogen-bond acceptor). These auto- and cross-correlograms are then used as descriptors. Almond descriptors have, so far, in the ADMET area primarily been used in P450 modeling [44–46].

14.2.2.4 **Pharmacophore Fingerprints**

Pharmacophores have been used for many years to derive models for understanding the common interaction patterns of ligands or their subsets. These pharmacophores have, subsequently, been used not only to design new structures with better target properties or devoid of such activities if so being the desired case, for example, hERG pharmacophores [47], but also to search 3D databases for new interesting entities or core structures. For a recent review on pharmacophores, see Ref. [48].

Pharmacophore fingerprints have also been used to assess the similarity of molecules from an interaction property point of view [49–51].

Pharmacophore fingerprints have typically so far been generated as three- or four-point pharmacophores spanning 0–15 (16) Å edges with increments of typically 2–3 Å. However, there are some issues to the generation of these fingerprints. First, how important is information of chirality, that is, is a three-point pharmacophore triangle sufficient or is the need for a four-point pharmacophore important for the set of studied compounds? This choice has considerable implications with respect to the number of fingerprints generated and stored for subsequent use for a typical set of structures. Second, is a conformational analysis required so that many conformations for a particular compound may map to the pharmacophores or will a single conformation be sufficient? For the latter case, should that conformation be selected from the conformational analysis and, if so, which conformations should be used? Perhaps, the lowest energy conformation should be used or some conformation of choice, for example, the one more closely related to the proposed bioactive conformation, is the best choice. Perhaps, a single conformation generated with a 3D generation program, for example, CORINA, is good enough for the problem at hand. If so, should that conformation be subjected to an energy minimization? Three-point pharmacophore triangles using conformational analysis ($\sim$10 k pharmacophores) with subsequent support vector machine (SVM)-based classification modeling for lead hopping purposes have been published by Saeh and coworkers [52]. Pharmacophore fingerprints have also been used for modeling the efflux transporter P-glycoprotein (P-gp) by Penzotti and coworkers [53]. In this work, a huge ensemble of fingerprints from all two-, three-, and four-point pharmacophores present in the conformers of the investigated compounds was generated. Even though the authors imposed a limit of at most two hydrophobic features for each generated pharmacophore, the resulting pharmacophore bit length was approximately 12 million bits! Recently, an interesting approach using a combination of pharmacophore fingerprints and GRID MIFs (see Section 14.2.2.3) called FLAPs (fingerprints for ligands and proteins) has been developed where the combined knowledge of protein and ligand profiles is used [54].

14.2.3
Property-Based Descriptors

This section will cover a variety of descriptors related more to experimental physicochemical properties such as lipophilicity and hydrogen bonding.

14.2.3.1 log *P*

The calculated water/octanol partition coefficient (log *P*) is probably the most commonly used descriptor in structure–activity modeling. However, this well-known and widely used descriptor is not without computational difficulties. There exist quite a number of software programs for the prediction of log *P*, for example, CLOGP [55], KOWWIN [56], SciLogP/ULTRA [57], and ACD/logP [58], that employ different algorithms including experimental values of parent structures, that is, a substructure of the compound to be predicted, coupled with perturbation equations, that is, equations for corrections due to additional fragments in the investigated structure, and special correction factors due to, for instance, the proximity of other fragments, to more fragment-based approaches, where each fragment has a certain set of log *P* factors, and to rule-based approaches. This, in turn, means that a log *P* prediction for a particular compound may vary markedly using different programs. For two recent investigations of this issue, see Refs [59, 60]. In addition, a fact to be remembered is that log *P* is a composite variable constituted by the three underlying properties of molecular size, polarity/polarizability, and hydrogen bonding. This natural partitioning of the three factors, as it occurs in log *P*, may not always be optimal for deriving the best statistical model [61]. Instead, models employing separate descriptors for molecular size, polarity/polarizability, and hydrogen bonding may indeed have more reliable forecasting abilities.

Although log *P* is a scalar, there are 3D protocols designed to calculate log *P*. These approaches utilize the concept of molecular lipophilicity potentials (MLPs). Testa and coworkers introduced MLPs using distance-dependent functions calculated on the solvent-accessible surface area molecules [62]. They used fragment-based lipophilicity factors from Broto and coworkers [63] as well as from Ghose and Crippen [64] to compute the MLPs. Lately, Du and coworkers have introduced the concept of heuristic molecular lipophilicity potentials (HMLPs) [65] that describe certain aspects of molecular solvation. The HMLPs are based on quantum mechanical electrostatic potentials (ESPs) that are calculated on a formal molecular surface of a compound. The corresponding molecular lipophilicity potential for a particular point on the surface is then constructed by comparing the local electron density at that point with the ESP on the surrounding atoms. Du and coworkers have applied the HMLP approach to calculate log *P* values for some alcohols [65].

14.2.3.2 HYBOT Descriptors

Another set of property-based descriptors that have been quite useful in ADMET modeling is the HYBOT parameters. Raevsky and coworkers have collected a large database of thermodynamic data related to hydrogen bonding with which they have developed the HYBOT program [66]. HYBOT will compute hydrogen-bond donor (ΣC_d) and acceptor (ΣC_a) factors that describe the donor and acceptor strengths, respectively, of a compound. By using these two descriptors, many significant statistical models related to areas such as water solubility, log P estimations, Caco-2 permeability, human intestinal absorption, and human skin permeability have been developed [67–69]. The HYBOT parameters represent another interesting aspect of computational descriptors containing relevant information for calculating not only a qualitative measure of a particular property, for example, hydrogen bonding, but also a more quantitative one [66]. By using HYBOT parameters, it is possible to obtain information regarding the possible importance of various hydrogen-bonding patterns, for example, whether a few but strong hydrogen-bonding groups are more important for the investigated property than perhaps many but weaker such entities.

14.2.3.3 Abraham Descriptors

Abraham and coworkers have developed a general solvation equation

$$\mathrm{SP} = c + eE + sS + aA + bB + vV, \tag{14.1}$$

where the dependent variable, SP, is the target property in question and E is an excess molar refraction, S represents the dipolarity/polarizability solute–solvent interactions, A and B are the hydrogen-bond acidity and basicity, respectively, and represent the strength and number of H bonds formed by donor and acceptor groups, respectively, in solute–solvent interactions, and V is the McGowan characteristic volume. These five parameters (E, S, A, B, and V) constitute the Abraham descriptors. The solute – descriptors A and B are based on the theoretical cavity model of solute-solvent interactions and are widely applied in the prediction of a variety of properties, such as solubility [70], blood–brain partitioning [71], and skin permeability [72]. Again, the use of the Abraham descriptors allows a more detailed understanding of possible hydrogen-bonding patterns.

14.2.3.4 Polar Surface Area

A very useful property for predicting absorption is the polar surface area (PSA), usually defined as those parts of the van der Waals or solvent-accessible surface of a molecule that are associated with hydrogen-bond-accepting capability (e.g., N or O atoms) and hydrogen-bond-donating capability (e.g., NH or OH groups). Three types of PSAs have been used in ADME studies:

1. dynamic PSA ($\mathrm{PSA_d}$) [73];
2. static PSA (PSA) [74]; and
3. two-dimensional (or topological) PSA (TPSA) [75].

The dynamic PSA, $\mathrm{PSA_d}$, was developed by Palm *et al.* [73]. $\mathrm{PSA_d}$ is calculated by a Monte Carlo conformational search with subsequent energy minimization.

This generates a set of low-energy conformers where the van der Waals surface-based PSAs for all conformers are within 2.5 kcal/mol of the "global" minimum, that is, the lowest energy conformer found, are computed. The Boltzmann-weighted average of the calculated PSAs are then used as the PSA_d. Palm and coworkers found a good sigmoidal correlation ($r^2 = 0.94$) between PSA_d and percentage human absorbed ($A\%$) for 20 well-characterized drugs [73].

A major drawback of the PSA_d is, however, the rather time-consuming calculation, particularly the Monte Carlo conformational search, which makes PSA_d inappropriate for computational screening (e-screening) of large virtual libraries.

This prompted further development of the static PSA, originally proposed by van de Waterbeemd and Kansy [76], based on only one conformer. Although this simplification would save considerable computational time, it is not without complications since it raises the question: Which conformation should be used? Probably a low-energy conformation could be considered as a good estimation of the bioactive conformation. However, in some cases, some sort of conformational search, although short, has to be employed, and most of the advantage of using PSA instead of PSA_d would be lost. Fortunately, a single conformer generated directly from the 2D molecular structure without minimization can be used. This approximation does not compromise the excellent correlation with absorption previously found [77, 78]. This approach reduces the computational time to such a level so as to make PSA useful for *in silico* screening of virtual libraries. However, still a slight drawback of PSA is the generation of the 3D conformation. The problem is not related to the computational time but from a conformational point of view. No matter how well 2D and 3D conversion programs, such as CORINA, perform on an overall basis, the generation may, in some cases, result in unreasonable 3D structures. Thus, it would be even more favorable if this step could be circumvented or eliminated in some manner.

Ertl and coworkers [75] have developed such a method for generating a topological PSA (TPSA) based on 3D PSA values for 43 fragments resulting from an analysis of 34 810 compounds taken from the WDI database. The correlation between PSA and TPSA is very high ($r^2 = 0.98$).

A further simplification, avoiding even the use of 3D fragments, has been developed by Sherbukhin [79]. This method uses a 2D projection technique whereby the TPSA (TPSA-2D) is computed. The algorithm employed sums up atomic contributions of 2D-generated atom-based van der Waals spheres and subtracts buried surfaces where two atomic spheres intersect to make a bond.

One thing to bear in mind here is that conformational dependencies may bury parts of the PSA, thus resulting in an overestimation of the computed TPSA.

14.3 Statistical Methods

There are a large number of statistical techniques available to the researcher to relate the independent variables (descriptors) computed for the objects (structures) under investigation to the corresponding dependent variable (target value).

These techniques span the entire field from multiple linear regression (MLR)-type methods and various forms of neural network architectures to rule-based techniques of different kinds. These approaches also span from single models to multiple models, that is, consensus or ensemble modeling. Terms like machine learning and data or information fusion are also frequently encountered in this area of research, as well as the concepts of applicability domain and validation.

This section attempts to present some of the most common statistical techniques used today to derive statistically sound structure–activity or structure–property models with good predictive ability.

There are a number of important issues and possible trade-offs to be discussed in this section that, in principle, do not stand against each other but, in reality, often do, such as the balance between the interpretability versus robustness or predictability of the derived model, that is, transparent versus opaque models, white versus black box models, a well as whether to derive local versus more global models for a particular target in question. For a recent review of statistical methods, see Ref. [80].

There are many ways of characterizing different statistical machine-learning methods and protocols, but in this section, they will be organized into linear and nonlinear methods (even though the descriptor matrix they operate on may contain higher order terms and cross-terms) as well as rule-based and Bayesian methods.

14.3.1 Linear and Nonlinear Methods

14.3.1.1 Multiple Linear Regression

Multiple linear regression is a classic mathematical multivariate regression analysis technique [81] that has been applied to quantitative structure–property relationship (QSPR) modeling. There are a few aspects, with respect to statistical issues, that the researcher must be aware of when using MLR:

1. A general prerequisite that affects all statistical multivariate data analysis techniques is that each of the variables should be given equal chance to influence the outcome of the analysis. This can be achieved by scaling the variables in an appropriative way. One popular method for scaling variables is autoscaling whereby the variance of each variable is adjusted to 1.
2. MLR assumes each variable to be exact and relevant.
3. Strong co-linear variables must be eliminated by removing all but one of the strongly correlated variables. Otherwise, spurious chance correlation may result.
4. Some sort of estimation of the statistical “distance” to the overall model should be reported for each compound to provide an estimate of how much an intra- or extrapolation in multivariate descriptor space the prediction actually constitutes.

MLR has been applied extensively to problems related to various aspects predicting ADMET properties such as solubility, Caco-2 cells permeability, human intestinal

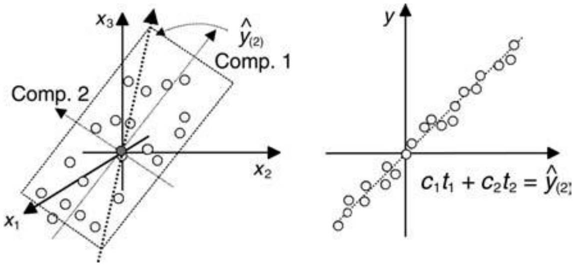

Figure 14.1 PLS of two components (picture reproduced with permission from the authors [125] and Umetrics, Inc.).

absorption, and blood–brain permeability, as well as for predicting metabolism. For a recent review, see Ref. [82].

14.3.1.2 **Partial Least Squares**

Partial least square projection to latent structures (PLS) [83] is a multivariate data analysis tool that has gained much attention during the past 10 years starting with the introduction of the 3D-QSAR method CoMFA [84]. PLS is a projection technique that uses latent variables (linear combinations of the original variables) to construct multidimensional projections while focusing on explaining as much as possible the information in the dependent variable and not among the descriptors used to describe the objects (compounds) under investigation (the independent variables) (Figure 14.1).

PLS differs from MLR in a number of ways:

1. The descriptors are not treated as exact and relevant but as consisting of two parts: one part related to the dependent variable and the other part unrelated (noise).
2. Strong correlations between relevant variables are not a problem in PLS, and all such variables can be kept in the analysis. In fact, the models derived using PLS become more stable with the inclusion of strongly correlated and relevant parameters.
3. The number of original descriptors may vastly exceed the number of compounds in the analysis (as opposed to MLR) since PLS uses only a few (usually less than 5–10) latent variables for the actual statistical analysis.
4. In PLS analysis, a “distance” to the overall model (distance-to-model), defined as the variance in the descriptors remaining after the analysis (residual standard deviation, RSD), is given for each predicted compound. This is an important piece of information that is presented to the researcher.

There are of course also some difficulties faced when using the PLS technique:

1. The number of latent variables (PLS components) has to be determined by some sort of validation technique, for example, cross-validation (CV) [85]. The PLS solution will coincide with the corresponding MLR solution when the number of latent variables becomes equal to the number of descriptors used in the analysis.

The validation technique, at the same time, also serves the purpose of avoiding overfitting of the model.

2. The possibility to use a very large number of descriptors, where many of them may not be particularly correlated with the dependent variable and thus represent large amounts of noise, must be considered with great care or otherwise the signal-to-noise ratio becomes too low for PLS to be able to create useful projections (latent variables).

14.3.1.3 **Artificial Neural Networks**

Artificial neural networks (ANNs) represent, as opposed to PLS and MLR, a nonlinear statistical analysis technique [86]. The most commonly used NN is of the feed-forward back-propagation type (Figure 14.2). As is the case of both PLS and MLR, there are a few aspects of NN to be considered when using this type of analysis technique:

1. The number of middle layers, hidden nodes, in an NN must be identified either through a particular choice or through an optimization procedure with careful monitoring of the predictive behavior of the derived model (see point 2).

2. NNs are well-known to overtrain, that is, to be able to explain a large portion of the variance of the dependent variable for the training set but to fail grossly to be able to predict a correct answer for the objects that are not part of the model (external test set). Overtraining of NNs can be avoided by setting aside a fixed number of compounds to validate the predictive ability of the NN model (validation set) as part of the NN training and stop when the predictive ability starts to deteriorate.

3. The interpretability of the derived NN model may be difficult to understand even though the influences of the descriptors on the derived model can be simulated. Guha and coworkers [87] have developed a two-step method for understanding the weights and biases in neural networks, in which first the neuron transform is linearized followed by a ranking scheme for the neurons.

NN methods have been used by Wessel and coworkers [39], Agatonovic-Kustrin *et al.* [8], and Ghuloum *et al.* [88] to model intestinal absorption.

14.3.1.4 **Bayesian Neural Networks**

Bayesian neural networks (BNNs) are an alternative to the more traditional ANNs. The main advantage with BNNs is that they are less prone to overtraining compared to ANNs. BNNs are based on Bayesian probabilistics for the network training. Network weights are determined by Bayesian inference. BNNs have been successfully used together with automatic relevance determination (ARD) for the selection of relevant descriptors to model aqueous solubility [89]. For a good review on BNNs, see Ref. [90].

14.3.1.5 **Support Vector Machines**

The Support vector machine technique is a relatively new method in the field of structure–property relationships. SVMs originated from the work of Vapnik *et al.* [91] and were originally applied to image analysis, text categorization, and

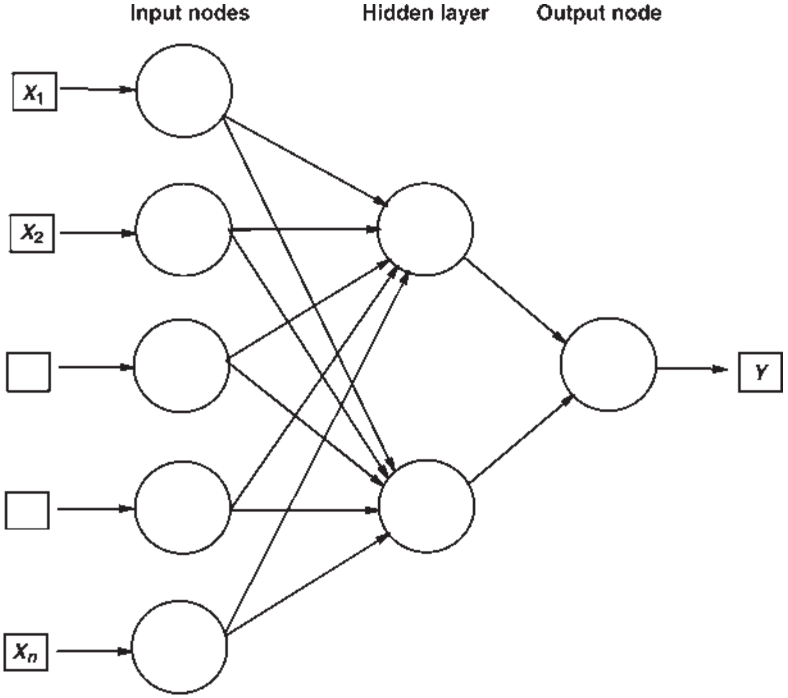

Figure 14.2 Simple scheme of an artificial neural network with one hidden layer.

character recognition [92]. Support vector machines have gained considerable interest in modeling ADMET properties during the last 5–6 years for, among other things, their robustness and forecasting abilities with respect to noisy data [93]. For a compilation of SVM applications in ADMET modeling, see Refs [94–96]. The basic idea of SVM technology is to construct a hyperplane that discriminates between the two classes of objects under investigation (binary SVM). The SVM algorithm maximizes the construction of a margin between the classes. SVMs use transformations of the original data for the successful construction of the margin (Figure 14.3).

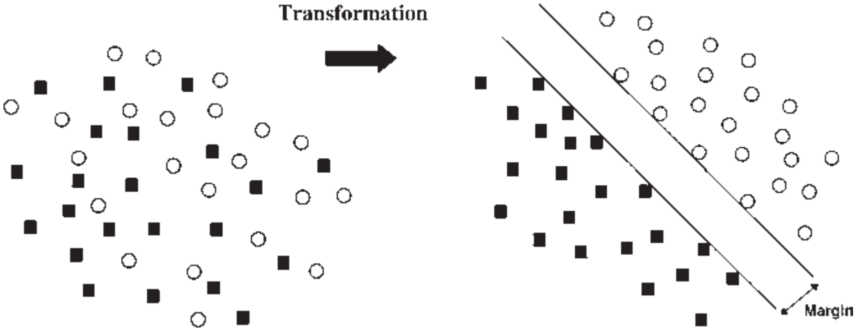

Figure 14.3 Basic principle of a support vector machine transformation for a two-class problem.

These transformations are executed by using so-called kernel functions. The kernel functions can be both linear and nonlinear in nature. The most commonly used kernel function is of the latter type and called the radial basis function (RBF). There are a number of parameters, for example, cost functions and various kernel settings, within the SVM applications that will affect the statistical quality of the derived SVM models. Optimization of those variables may prove to be productive in deriving models with improved performance [97]. The original SVM protocol was designed to separate two classes but has later been extended to also handle multiple classes and continuous data [80].

14.3.1.6 *k*-Nearest Neighbor Modeling

k-Nearest neighbor (kNN) modeling is based on the assumption of similarity, that is, similar compounds have similar target properties. In its simplest form, the method uses an unweighted distance measure, usually Euclidian distance, in chemical property space and from the *k*-nearest objects determines the target property or which class the object in question can be assigned to. There are, however, some aspects to highlight when using this approach:

1. Euclidian distances can, from a strict perspective, only be used for determining the distance between orthogonal variables. This is most often not the case for chemical descriptors. The problem of orthogonality can be handled in two ways: either compensate for the nonorthogonal behavior within the distance calculation, for example, use Mahalanobis distance instead [98] of an Euclidian distance or orthogonalize the variables, for example, by principal component analysis, prior to the Euclidian distance calculation.

2. All variables are treated equally importantly. It is unlikely that all the computed chemical properties for the compounds in the data set are of equal importance for the target property, for example, solubility, absorption, or metabolism.

3. The *k*-nearest neighbors are treated with equal weight with respect to determining the target property. This is of particular importance when estimating continuous properties. This aspect has been investigated by Shen *et al.* [99] where they used weighted distances to obtain better predictions for the target property.

14.3.1.7 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is aimed at finding a linear combination of descriptors that best separate two or more classes of objects [100]. The resulting transformation (combination) may be used as a classifier to separate the classes. LDA is closely related to principal component analysis and partial least square discriminant analysis (PLS-DA) in that all three methods are aimed at identifying linear combinations of variables that best explain the data under investigation. However, LDA and PLS-DA, on one hand, explicitly attempt to model the difference between the classes of data whereas PCA, on the other hand, tries to extract common information for the problem at hand. The difference between LDA and PLS-DA is that LDA is a linear regression-like method whereas PLS-DA is a projection technique (see Sections 14.3.1.1 and 14.3.1.2 for further details). Thus, for a two-class, two-descriptor problem (X_1 and X_2), the LDA description becomes

$$Y = c_1 \cdot X_1 + c_2 \cdot X_2. \tag{14.2}$$

The object of the LDA is to find values of the two constants c_1 and c_2, respectively, that separate the two classes expressed through the variable Y as, for instance, 1 and 2, respectively.

14.3.2 Partitioning Methods

14.3.2.1 Traditional Rule-Based Methods

The basic underlying idea with partitioning methods is to split, usually in a recursive, that is, repetitive, manner, the data set at hand into two or more groups, branches, thus creating a decision tree. The object is to create more and more homogeneous groups in the respective branches. There are several methods available for the construction of decision trees, for example, FIRM [101], CART [102], RDS [103], and C4.5 [104] (Figure 14.4).

As always, there are certain aspects to consider when developing a decision tree:

1. Overtraining: As with many other methods, decision trees are prone to overtraining if not monitored. The forecasting ability of the tree must be estimated by some, usually, internal validation method such as a validation set or through cross-validation. This will determine the depth and degree of branching of the derived tree.
2. Forecasting ability: Most decision tree methods are "greedy," that is, they split on the variable giving the best enhancement in group homogeneity according to

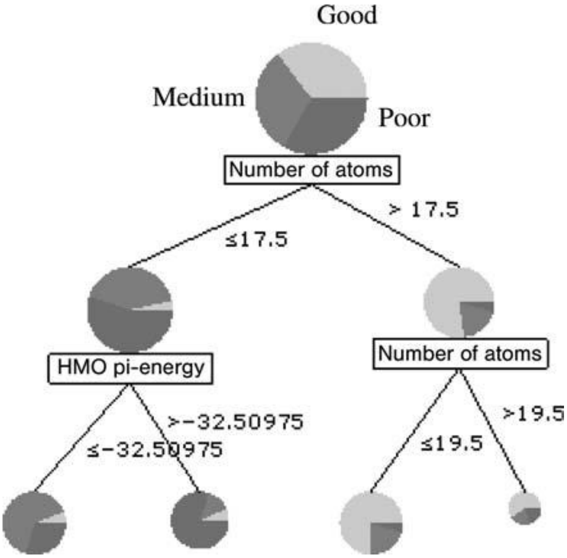

Figure 14.4 Simple decision tree with split points and terminal leaves.

some statistical test, such as the *t*- or *F*-test, at each split point. This may, however, not produce the model with the best predictive performance after model construction. Random selection of variables, that is, a subset, available for each split has been devised to overcome this situation.

Recursive partitioning has successfully been used to develop models for various ADMET properties, see Ref. [80], as well as for the elucidation of toxicological modes of action [103].

14.3.2.2 Rule-Based Methods Using Genetic Programming

Neural networks and genetic algorithms (GAs) have been used in QSAR applications for some time [105]. The main idea of genetic programming (GP) closely resembles that of the GA. Most GP applications use a tree-based representation and normally a genetic program has a tree-like construction consisting of functional nodes and terminal leaves (see Figure 14.5).

The algorithm G-REX uses crossover and mutation (see Figure 14.6), and the extraction strategy is based on genetic programming [106–109].

Crossover is very common, that is, around 85% of each new generation is created by crossover, whereas mutation is used less than 2% for generating new offsprings. One key property of G-REX is the option to directly balance accuracy against comprehensibility by using an appropriate fitness function. G-REX modeling often results in rather short and transparent models (see Chapter 15, Section 15.3.2.2: an example using genetic programming-based rule extraction).

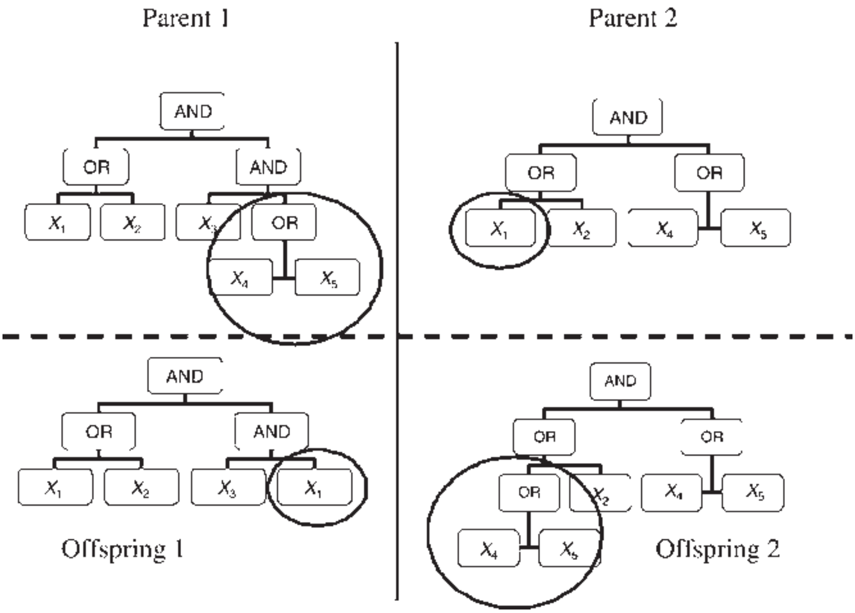

Figure 14.5 Principle of crossover in genetic programming (picture reproduced with permission from the author [109]).

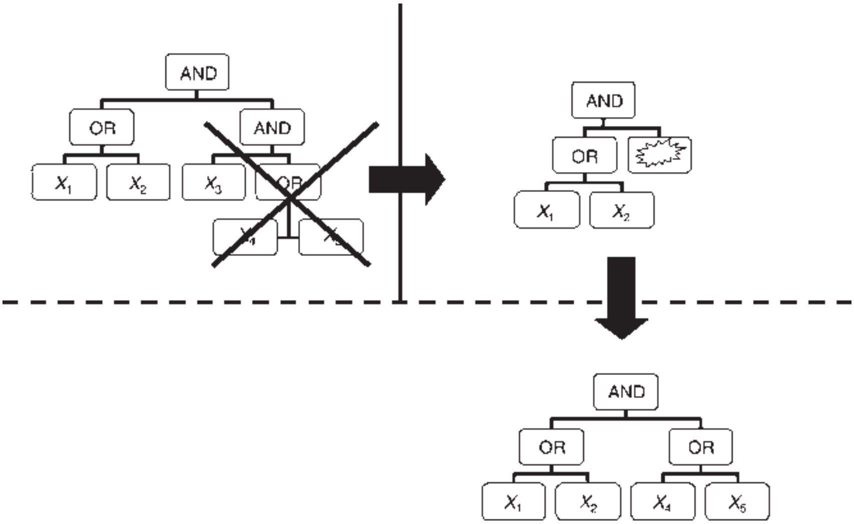

Figure 14.6 Principle of mutation in genetic programming (picture reproduced with permission from the author [109]).

14.3.3 Consensus and Ensemble Methods

Many times there is a trade-off between the transparency and the accuracy of the derived model for a particular target. In some cases, the transparency is the most important aspect as long as the derived model has acceptable predictive ability. These models are better suited to answer the question, "Which are the next compounds to investigate?" by understanding the underlying properties of deriving compounds with improved target properties. Other times the most important aspect of a derived *in silico* model is the accuracy and robustness with respect to its forecasting performance. For the latter cases, consensus, often also called ensemble, models may offer some attractive properties. Many times these kinds of models offer improved forecasting performance and robustness compared to an individual model for the target in question. There are several ways to construct consensus models with respect to the input variables, the statistical methods, and how the final prediction from the battery of models is derived. It should be noted that the two kinds of models, transparent versus more opaque, are not necessarily in conflict with each other. On the contrary, they may benefit from each other's existence in the following manner: One may use the more transparent models to focus the attention around certain areas (subspaces) in property space indicated to result in promising new entities. Once these areas have been identified, the more complex consensus model, sometimes requiring considerably longer time for computation, is employed to predict the target property in question at a higher level of accuracy and precision. However, it may sometimes be debatable whether or not the increase in performance really outweighs the added complexity [110].

The simplest and most straightforward way to employ consensus modeling is to use the same set of descriptors and statistical method (single method-descriptor set methods). Considerable time saving with respect to descriptor generation, that is, using only one set and not several different sets, can be achieved. In addition, using the same statistical approach may have a favorable effect on implementation of the consensus approach with respect to issues such as licenses and data format compatibility. Single method-descriptor set methods have been implemented in decision tree programs such as TreeNet [111] and RDS [112]. The approach actually consists of two selection parts. The training set for the various models of contributing to the ensemble is often selected using bagging (bootstrap aggregating). Then, at each split point for decision tree programs, the variables to be included in the model are also randomly chosen. Thus, a certain variation of the models of the ensemble is achieved that promotes predictive performance and robustness of the final ensemble. There are also approaches available to monitor and ensure a certain diversity of the derived ensembles with acceptable statistical quality. One such an approach is DECORATE (diverse ensemble creation by oppositional relabeling of artificial training examples) [113].

The obvious extension to single method-descriptor set methods is of course some combination of single or multiple methods using single or multiple descriptor sets.

Various forms of these combinations with an accuracy of the ensemble models better than the corresponding single reference have been reported [90, 114–119].

14.4 Applicability Domain

A very important aspect of statistical modeling is to determine the domain in which the model is defined with high significant reliability, called the applicability domain.

It is important to do this for several reasons:

1. To make the users of the model aware of the applicability and limitations of the present model.
2. To avoid the misuse of a model for forecasting of compounds outside the model's present statistical limits, which renders the model (and/or statistical technique as well as the parameterization) a false "bad reputation."
3. To be able to use extrapolations from the present model in a constructive manner to expand the model to cover a larger domain space.

However, many statistical modeling techniques do not, in an easy and straightforward way, by default, enable the estimation of whether a prediction is an interpolation to the model, thus rendering the prediction more credibility or an extrapolation to the model in which case the prediction must be evaluated with greater care. Furthermore, there are two aspects to the extrapolation problem: one structural and the other statistical. Considerable research has been devoted to the problem of ADs. For a recent compilation on this issue, see Ref. [120].

The structurally focused methods for defining ADs are related to a large extent with the independent variable (descriptor) side. These methods comprise techniques, such as

1. Range-based methods whereby the AD is defined solely on the ranges that the investigated descriptors of the training set, that is, the objects used to derive the model in question, span. If a new object to be predicted is within the range of all the model descriptors, then the object is within the AD of the model.
2. Distance-based methods whereby some kind of distance measure between the object to be predicted and the closest neighbor or neighbors of the training set defines the AD. Typical methods for distance-based measures are Euclidian and Mahalanobis distances. The difference between the two is that the former distance can only be used for determining the distance between orthogonal variables, whereas the latter method compensates for the nonorthogonal behavior (see further discussion in Section 14.3.1.6). In its simplest form, all descriptors are treated with equal importance while some more advanced methods use some kind of weighting scheme to increase accuracy and relevance of the distance measure, for example, by using the coefficients of the derived statistical model for modulating the influence of the descriptors.
3. Geometric methods where most definitions rest on defining the smallest convex area that covers the training set compounds in descriptor space. This method is also known as the convex hull method (Figure 14.7).

The statistically focused methods for defining ADs are related to information content of the investigated descriptors, for example, the variance of the descriptor matrix and calculate the amount of an unexplained variance for the training set objects (the model) and compare it with the corresponding amount for the new objects to be predicted. If the amount of unexplained variance for the new objects is much greater, typically more than the two standard deviations from the training set compounds (~95% confidence interval), the former objects are designated to be outside the AD of the model.

A constructive way of using the estimations of AD, and particularly extrapolations thereof, as mentioned under point 3, would be to include some of the predicted

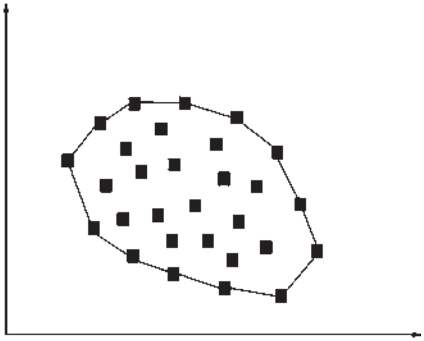

Figure 14.7 Schematic representation of the convex hull method for a two-parameter description.

compounds that are identified as outliers in later updates of the model thus increasing the AD of the new model.

14.5
Training and Test Set Selection and Model Validation

An integral part of deriving a statistically valid and predictive *in silico* model is the choice of training and test set, as well as the model validation. Without proper validation of the derived model, it is difficult to assess its statistical qualities and forecasting ability.

14.5.1
Training and Test Set Selection

There are several methods for designing the training set and test set, respectively. Most of them are dissimilarity based, that is, their aim is to select a training set as diverse as possible for the studied descriptor and target space. Some include the target variable, for example, biological activity, as part of the selection process so that a good spread in target space is also achieved. Lately, however, approaches focusing on local rather than global models have suggested the opposite strategy for the selection of the training set; that is, selecting a small set of similar compounds with respect to the new compound to be predicted and for each prediction, deriving a local statistical model on-the-fly, also known as lazy model [121–124]. The potential issues with lazar (lazy structure–activity relationships) methods are related to the underlying basic assumption of structure–activity relationships, namely, that similar molecules have similar activities. It is sometimes difficult to select a good representative set of similar training set compounds for which there is a sufficient spread in target value, for example, biological activity, and some investigations report that the local lazy model predicts worse than the corresponding global model [122].

For the purpose of selecting a diverse training set, one may use experimental design methods, for example, fractional factorial designs (FFDs) [125–128]. Using these methods, a training set with a good spread in the investigated properties (descriptors) can usually be selected. Since the nature of the FFDs is to select compounds at the edges of the investigated descriptor space and the user has considerable freedom of choice with respect to which compounds to include in the proposed FFD, the user must be aware that too extreme compounds should not form the majority of the compounds selected for the final training set. It is therefore also common to combine the FFDs with some compounds close to the center of the FFD. To avoid some of the above-mentioned problems with classical FFDs, a new type of designs called onion designs (ODs) has recently been developed [129]. The purpose of ODs is to achieve efficiency as well as controlled coverage of both the outer and the inner region of the descriptor space. The OD approach is based on combining several designs in layers. Thus, the compounds available are divided into subsets, or layers, and a separate selection is performed on each subset. The selection technique in each layer can be of different kind, for example, FFDs or some space-filling design (SFD).

SFDs are aimed at creating a uniform distribution of compounds in descriptor space by selecting compounds as dissimilar as possible. A particular form of SFDs is the maximin technique designed by Marengo and Todeschini [130] whereby the desired number of training set compounds are selected by maximizing the shortest (minimum) distance in descriptor space between the chosen compounds. Again, if using Euclidian distance as the distance measure, orthogonality among the descriptors must be ensured. The authors have found that using a principal component analysis prior to the maximin selection addresses not only the descriptor orthogonality problem but also reduces the number of variables to be considered during the selection process. The latter is of interest since the method involves repetitive distance calculations between the compounds presently chosen and potential new compounds to be exchanged in order to maximize the shortest distance among the training set compounds. There are also sphere exclusion algorithms available for training and test set selection [131].

14.5.2 Model Validation

Some kind of model validation is necessary to determine the statistical quality with respect to forecasting target values of new compounds. There are typically three validation tests that should be performed:

- internal validation, for example, cross-validation [85];
- randomization of the target variable; and
- external test set predictions.

The internal validation is often performed through cross-validation whereby one, leave-on-one-out cross validation (LOO-CV), or several, leave-multiple-out cross validation (LMO-CV), compounds are removed from the training set. The remaining compounds of the training set are then used to derive a model with which the left-out compounds are predicted. Another set of compounds is then removed from the training set, a new model derived, and the new set of left-out compounds predicted. This procedure is continued until all compounds have been left out once. The computed measure of quality is the cross-validation squared correlation coefficient (q^2). While the normal squared correlation coefficient (r^2) can only assume values between 0 and 1, the cross-validation squared correlation coefficient can be both positive and negative. In fact, a value of zero for q^2 merely indicates that the model has used the average experimental value for all of the training set compounds as the predicted value for each test compound. Normally, values greater than 0.3 are recommended for a model to be considered as statistically sound, but it has been shown that values lower than 0.3 may be acceptable depending on the size of the data set [132]. The problem with LOO-CV methods is that they tend to overestimate the forecasting ability of the model when presented with new external compounds to be predicted. Also, LMO-CV methods have the same tendency, albeit to a lower extent. A real danger with cross validation is through the combination with variable selection. When variable selection is applied and the computed q^2 is used to drive the variable selection, then the validation aspect of cross validation is lost and q^2 becomes an optimization function instead. By using this kind of approach, it is possible to fit random (white) noise with excellent statistical "quality" and respectable q^2 values (>0.5, unpublished result by the author). For further information regarding the predictive ability overestimation by the q^2 metric, see Ref. [133]. Variable randomization is another method for ensuring the reliability of models. In this method, the values of the target are randomly reassigned for training set compounds and then used to derive a new "model" [134]. After performing the randomization procedure sufficiently (more than 50–100 times), there should be clear difference between the model derived using true target values and the model derived using randomized target data. The most rigorous validation of a derived model is, however, through the use of an external test set, that is, data that have not been used for deriving the models. The use of external test sets is not without problems either. It is important that the test set also covers the applicability domain of the model to be evaluated in a good manner. There should be sufficient difference between training and test set compounds so that near-neighbor compounds are not in both sets. If that is the case, then the predictive ability of the derived model will most likely be significantly overestimated. The use of an external test set may, in some cases, be the only way, of the three validation procedures described here, to realize that the derived model is without any forecasting ability. For more details regarding model validation, see Ref. [135].

14.6
Future Outlook

The application of SAR and QSAR in modern discovery research to predict important properties, for example, solubility, absorption, and toxicity, of both small and large collections of (virtual) libraries, also known as "frontloading," forces not only the development of both more informative and more easily computed, for example, faster computed, molecular descriptors but also necessitates new statistical techniques to be used. For instance, the need for robust and predictive methods and models in virtual screening may infer the use of rather opaque consensus or ensemble methods, whereas transparent models are of value to understand the most important properties for the target in question and perhaps, at the same time, learn something about the mechanisms and/or processes at hand. Thus, one may envision the intertwined use of both these approaches, that is, transparent and opaque models, to enable better understanding as well as final precision and quality of predictions. The transparent models, with acceptable statistical quality, may then be utilized to drive virtual library generation to the most promising areas in chemical space, whereas the more complex and opaque models may then be applied for the final predictions to obtain the extra precision and robustness offered by these latter techniques.

Another important area of research is the “T” in *in silico* ADMET, that is, *in silico* toxicology, on which a lot of effort is already spent both in academia and in industry. The ADME area has seen substantial efforts during the past 10 years to obtain important understanding as well as to derive good models for solubility and absorption. The coming years will most likely see the same kind of attention for *in silico* toxicology. However, this new area of research is more demanding from a mechanistic point of view than that of solubility and absorption. Several mechanisms may come into play, even within analogous series of compounds, depending on what chemical functionalities are present in the molecule. Also, overall physicochemical parameters describing the entire structure may be less useful for modeling toxicology. Probably, structural fragments, toxicophores, will prove to be more important and useful descriptors and as descriptors capturing electronic properties of the investigated compounds. The use of such descriptors is most probably not a straightforward exercise either, to a certain extent depending upon the constitution of the structural fragments employed, since the chemical surrounding of a substructure, fragment, is important to ascertain whether the compound is likely to be toxic or not. Thus, a particular fragment may be identified as inducing toxicity in one compound but not in another. Many of the statistical methods employed today will have problems with such descriptors and will be unable to derive good statistical models. A possible way in the future is therefore to both derive new descriptor sets and use other statistical tools that take context dependency better into account.

References

1 http://www.talete.mi.it.

2 http://www.edusoft-lc.com/molconn.

3 http://www.chemcomp.com.

4 http://www.tripos.com.

5 http://www.molecular-networks.com.

6 Liu, H.X., Hu, R.J., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T. (2005) The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *Journal of Computer-Aided Molecular Design*, **19**, 33–46.

7 Bai, J.P.F., Utis, A., Crippen, G., He, H.-D., Fischer, V., Tullman, R., Yin, H.-Q., Hsu, C.-P., Jiang, L. and Hwang, K.-K. (2004) Use of classification regression tree in predicting oral absorption in humans. *Journal of Chemical Information and Computer Sciences*, **44**, 2061–2069.

8 Agatonovic-Kustrin, S., Beresford, R., Pauzi, A. and Yusof, M. (2001) Theoretically-derived molecular descriptors important in human intestinal absorption. *Journal of Pharmaceutical and Biomedical Analysis*, **25**, 227–237.

9 Klon, A.E., Lowrie, J.F. and Diller, D.J. (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *Journal of Chemical Information and Modeling*, **46**, 1945–1956.

10 Zmuidinavicius, D., Didziapetris, R., Japertas, P., Avdeef, A. and Petrauskas, A. (2003) Classification structure–activity relations (C-SAR) in prediction of human intestinal absorption. *Journal of Pharmaceutical Sciences*, **92**, 621–633.

11 Klopman, G., Stefan, L.R. and Saiakhov, R.D. (2002) ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences*, **17**, 253–263.

12 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23**, 3–25.

13 http://www.scitegic.com.

14 Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J.G. (2002) Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, **42**, 1273–1280.

15 http://www.digitalchemistry.co.uk.

16 http://www.leadscope.com.

17 http://www.mdl.com.

18 Kazius, J., McGuire, R. and Bursi, R. (2005) Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, **48**, 312–320.

19 http://www.daylight.com.

20 Niwa, T. (2003) General regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *Journal of Chemical Information and Computer Sciences*, **43**, 113–119.

21 Pérez, M.A.C., Sanz, M.B., Torres, L.R., Ávalos, R.G., González, M.P. and Díaz, H.G. (2004) A topological sub-structural approach for predicting human intestinal absorption of drugs. *European Journal of Medicinal Chemistry*, **39**, 905–916.

22 Sun, H. (2004) A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *Journal of Chemical Information and Computer Sciences*, **44**, 748–757.

23 Hall, L.H. and Kier, L.B. (1991) *The molecular connectivity chi indices and kappa shape indices in structure–property modelling, in Reviews of Computational Chemistry*, **Vol. 2** (eds D.B. Boyd and K. Lipkowitz), VCH Publishers, USA, pp. 367–422.

24 Downs, G.M. (2004) Molecular descriptors, in *Computational Medicinal Chemistry for Drug Discovery* (eds P. Bultinck, J.P. Tollenaere and H. de Winter), Marcel Dekker, USA, pp. 515–537.

25 Gunturi, S.B. and Narayanan, R. (2007) *In silico* ADME modeling 3: computational models to predict human intestinal absorption using sphere exclusion and kNN QSAR methods. *QSAR & Combinatorial Science*, **26**, 653–668.

26 Hall, L.H. and Kier, L.B. (1999) *Molecular Structure Description: The Electrotopological State*, Academic Press, USA.

27 Hall, L.H., Mohney, B. and Kier, L.B. (1991) The electrotopological state: structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Computer Sciences*, **31**, 76–82.

28 Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H. (1996) The e-state fields. Applications to 3D QSAR. *Journal of Computer-Aided Molecular Design*, **10**, 513–520.

29 Rose, K., Hall, L.H. and Kier, L.B. (2002) Modeling blood–brain barrier partitioning using the electrotopological state. *Journal of Chemical Information and Computer Sciences*, **42**, 651–666.

30 Stenberg, P., Norinder, U., Luthman, K. and Artursson, P. (2001) Experimental and computational screening models for the prediction of intestinal drug absorption. *Journal of Medicinal Chemistry*, **44**, 1927–1937.

31 Todeschini, R. and Grammatica, P. (1997) 3D-modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies. *Quantitative Structure–Activity Relationships*, **16**, 120–125.

32 Todeschini, R., Grammatica, P., Marengo, E. and Provenzani, R. (1996)

Modeling and prediction by using WHIM descriptors in QSAR studies: submitochondrial particles (SMP) as toxicity biosensors of chlorophenols. *Chemosphere*, **33**, 71–79.

33 Todeschini, R., Lasagni, M. and Marengo, E. (1994) New molecular descriptors for 2D and 3D structures. Theory. *Journal of Chemometrics*, **8**, 263–272.

34 Todeschini, R., Vighi, M., Provenzani, R., Finzio, A. and Grammatica, P. (1996) Modeling and prediction by using whim descriptors in QSAR studies: toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere*, **32**, 1527–1545.

35 Stanton, D.T. and Jurs, P.C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Analytical Chemistry*, **62**, 2323–2329.

36 Stanton, D.T., Mattioni, B.E., Knittel, J.J. and Jurs, P.C. (2004) Development and use of hydrophobic surface area (HSA) descriptors for computer-assisted quantitative structure–activity and structure–property relationship studies. *Journal of Chemical Information and Computer Sciences*, **44**, 1010–1023.

37 Wildman, S.A. and Crippen, G.M. (1999) Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, **39**, 868–873.

38 Pearlman, R.S. (1980) Molecular surface area and volumes and their use in structure/activity relationships, in *Physical Chemical Properties of Drugs* (eds S.H., Yalkowsky, A.A. Sinkula and S.C. Valvani), Marcel Dekker, USA, pp. 321–347.

39 Wessel, M.D., Jurs, P.C., Tolan, J.W. and Muskal, S.M. (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences*, **38**, 726–735.

40 Mazzatorta, P., Cronin, M.T.D. and Benfenati, E. (2006) A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. *QSAR & Combinatorial Science*, **25**, 616–628.

41 Cruciani, G., Pastor, M. and Guba, W. (2000) VolSurf: a new tool for the pharmacokinetic optimization of lead compound. *European Journal of Pharmaceutical Sciences*, **11** (Suppl. 2), S29–S39.

42 Cruciani, G., Meniconi, M., Carosati, E., Zamora, I. and Mannhold, R. (2003) VOLSURF: a tool for drug ADME–properties prediction, in *Drug Bioavailability*, Vol. 18 (eds H. van de Waterbeemd, H. Lennernäs and P. Artursson), *Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Germany, pp. 406–419.

43 Crivori, P., Cruciani, G., Carrupt, P.A. and Testa, B. (2000) Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry*, **43**, 2204–2216.

44 Crivori, P., Zamora, I., Speed, B., Orrenius, C. and Poggesi, I. (2004) Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *Journal of Computer-Aided Molecular Design*, **18**, 155–166.

45 Afzelius, L., Zamora, I., Masimirembwa, C.M., Karlen, A., Andersson, T.B., Mecucci, S., Baroni, M. and Cruciani, G. (2004) Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. *Journal of Medicinal Chemistry*, **47**, 907–914.

46 Afzelius, L., Masimirembwa, C.M., Karlen, A., Andersson, T.B. and Zamora, I. (2002) Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *Journal of Computer-Aided Molecular Design*, **16**, 443–458.

47 Aronov, A.M. and Goldman, B.B. (2004) A model for identifying hERG K^+ channel blockers. *Bioorganic and Medicinal Chemistry*, **12**, 2307–2315.

48 Drie, J.H. (2004) Pharmacophore discovery: a critical review, in *Computational Medicinal Chemistry for Drug Discovery* (ed. P. Bultinck), Marcel Dekker, USA, pp. 437–460.

49 Mason, J.S. and Pickett, S.D. (2003) Combinatorial library design, molecular similarity and diversity applications, in *Burger's Medicinal Chemistry and Drug Discovery* 6 (ed. D.J. Abraham), John Wiley & Sons, Inc., USA, pp. 187–242.

50 Good, A.C., Mason, J.S. and Pickett, S.D. (2000) Pharmacophore pattern application in virtual screening, library design and QSAR, in *Methods and Principles in Medicinal Chemistry*, Vol. 10 (eds H.J. Bohm and G. Schneider), John Wiley & Sons, Inc., USA, pp. 131–159.

51 Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C.R. and Labaudiniere, R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry*, **42**, 3251–3264.

52 Saeh, J.C., Lyne, P.D., Takasaki, B.K. and Cosgrove, D.A. (2005) Lead hopping using SVM and 3D pharmacophore fingerprints. *Journal of Chemical Information and Modeling*, **45**, 1122–1133.

53 Penzotti, J.E., Lamb, M.L., Evensen, E. and Grootenhuis, P.D.J. (2002) A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *Journal of Medicinal Chemistry*, **45**, 1737–1740.

54 Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. and Mason, J.S. (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *Journal of Chemical Information and Modeling*, **47**, 279–294.

55 http://www.biobyte.com.

56 http://www.syrres.com/esc/kowwin.htm.

57 http://www.scivision.com.

58 www.acdlabs.com.

59 Eros, D., Kövesdi, I., Orfi, L., Takács-Novák, K., Acsády, G. and Kéri, G. (2002) Reliability of logP predictions based on calculated molecular descriptors: a critical review. *Current Medicinal Chemistry*, **9**, 1819–1829.

60 Machatha, S.G. and Yalkowsky, S.H. (2005) Comparison of the octanol/water partition coefficients calculated by ClogP®, ACDlogP and KowWin® to experimentally determined values. *International Journal of Pharmaceutics*, **294**, 185–192.

61 Norinder, U. and Österberg, T. (2001) Theoretical calculation and prediction of drug transport processes using simple parameters and partial least squares projections to latent structures (PLS) statistics. The use of electrotopological state indices. *Journal of Pharmaceutical Sciences*, **90**, 1076–1084.

62 Testa, B., Carrupt, P.-A., Gaillard, P., Billois, F. and Weber, P. (1996) Lipophilicity in molecular modeling. *Pharmaceutical Research*, **13**, 335–343.

63 Broto, P., Moreau, G. and Van Dycke, C. (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies. *European Journal of Medicinal Chemistry*, **19**, 61–70.

64 Ghose, A.K. and Crippen, G.M. (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. 1. Partition coefficients as a measure of hydrophobicity. *Journal of Computational Chemistry*, **7**, 565–577.

65 Du, Q., Liu, P.-J. and Mezey, P.G. (2005) Theoretical derivation of heuristic molecular lipophilicity potential: a quantum chemical description for molecular solvation. *Journal of Chemical Information and Modeling*, **45**, 347–353.

66 Raevsky, O.A. (1997) Hydrogen bond estimation by means of HYBOT, in *Computer-Assisted Lead Finding and Optimisation* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Verlag Helvetica Chimica Acta, Switzerland, pp. 367–378.

67 Raevsky, O.A., Schaper, K.-J., van de Waterbeemd, H. and McFarland, J. (1999) Hydrogen bond contribution to properties and activities of chemicals and drugs, in *Molecular Modeling and Prediction of Bioactivity* (eds K. Gundertofte and K. Jorgensen), Kluwer Academic/Plenum Publishers, USA, pp. 221–228.

68 McFarland, J.W., Raevsky, O.A. and Wilkerson, W.W. (1999) Hydrogen bond acceptor and donor factors, C_a and C_d: new QSAR descriptors, in *Molecular Modeling and Prediction of Bioactivity* (eds K. Gundertofte and K. Jorgensen), Kluwer Academic/Plenum Publishers, USA, pp. 280–281.

69 Raevsky, O.A., Schaper, K.-J., Artursson, P. and McFarland, J.W. (2002) A novel approach for prediction of intestinal absorption of drugs in humans based on hydrogen bond descriptors and structural similarity. *Quantitative Structure–Activity Relationships*, **20**, 402–413.

70 Abraham, M.H. and Le, J. (1999) The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *Journal of Pharmaceutical Sciences*, **88**, 868–880.

71 Abraham, M.H., Ibrahim, A., Zhao, Y. and Acree, W.E., Jr. (2006) A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *Journal of Pharmaceutical Sciences*, **95**, 2091–2100.

72 Abraham, M.H. and Martins, F. (2004) Human skin permeation and partition: general linear free-energy relationship analyses. *Journal of Pharmaceutical Sciences*, **93**, 1508–1523.

73 Palm, K., Stenberg, P., Luthman, K. and Artursson, P. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research*, **14**, 568–571.

74 Clark, D.E. (1999) Rapid calculation of polar surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *Journal of Pharmaceutical Sciences*, **88**, 807–814.

75 Ertl, P., Rohde, B. and Selzer, P. (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, **43**, 3714–3717.

76 van de Waterbeemd, H. and Kansy, M. (1992) Hydrogen-bonding capacity and brain penetration. *Chimia*, **46**, 299–303.

77 Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of ‘drug-likeness’. *Drug Discovery Today*, **5**, 49–58.

78 Clark, D.E. (2001) Prediction of intestinal absorption and blood–brain barrier penetration by computational methods. *Combinatorial Chemistry & High Throughput Screening*, **4**, 477–496.

79 Sherbukhin, V.V. (2002) Personal communication.

80 Fox, T. and Kriegl, J.M. (2006) Machine learning techniques for *in silico* modeling of drug metabolism. *Current Topics in Medicinal Chemistry*, **6**, 1579–1591.

81 Livingstone, D. (1995) *Data Analysis for Chemists Applications to QSAR and Chemical Product Design*, Oxford University Press, United Kingdom.

82 Lombardo, F., Gifford, E. and Shalaeva, M.Y. (2003) *In silico* ADME prediction: data, models, facts and myths. *Mini Reviews in Medicinal Chemistry*, **3**, 861–875.

83 Wold, S., Johansson, E. and Cocchi, M. (1993) PLS – partial least-squares projections to latent structures, in *3D QSAR in Drug Design* (ed. H. Kubinyi), ESCOM Science Publishers B.V., The Netherlands, pp. 523–550.

84 Cramer, R.D., Patterson, D.E. and Bunce, J.D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, **110**, 5959–5967.

85 Wold, S. (1979) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 379–405.

86 Rojas, R. (1996) *Neural Networks – A Systematic Introduction*, Springer-Verlag, Germany.

87 Guha, R., Stanton, D.T. and Jurs, P.C. (2005) Interpreting computational neural network quantitative structure–activity relationship models: a detailed interpretation of the weights and biases. *Journal of Chemical Information and Modeling*, **45**, 1109–1121.

88 Ghuloum, A.M., Sage, C.R. and Jain, A.N. (2000) Molecular hashkeys: a novel method for molecular characterisation and its application for predicting important pharmaceutical properties of molecules. *Journal of Medicinal Chemistry*, **42**, 1739–1748.

89 Bruneau, P. (2001) Search for predictive generic model of aqueous solubility using Bayesian neural nets. *Journal of Chemical Information and Computer Sciences*, **41**, 1605–1616.

90 Gola, J., Obrezanova, O., Champness, E. and Segall, M. (2006) ADMET property prediction: the state of the art and current challenges. *QSAR & Combinatorial Science*, **25**, 1172–1180.

91 Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*, Springer, USA.

92 Hearst, M.A., Schölkopf, B., Dumais, S., Osuna, E. and Platt, J. (1998) Trends and controversies: support vector machines. *IEEE Intelligent Systems*, **13**, 18–28.

93 Czerminski, R., Yasri, A. and Hartsough, D. (2001) Use of support vector machine in pattern classification: application to QSAR studies. *Quantitative Structure–Activity Relationships*, **20**, 227–240.

94 Trotter, M.W.B. and Holden, S.B. (2003) Support vector machines for ADME property classification. *QSAR & Combinatorial Science*, **22**, 533–548.

95 Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S. and Lemmen, C. (2003) Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, **43**, 667–673.

96 Barrett, S.J. and Langdon, W.B. (2006) Advances in the application of machine learning techniques in drug discovery, design and development, in *Applications of Soft Computing: Recent Trends* (*Advances in Soft Computing*) (eds A. Tiwari, J. Knowles, E. Avineri, K. Dahal and R. Roy), Springer, USA.

97 Norinder, U. (2003) Support vector machine models in drug design – applications to drug transport processes and QSAR using simplex optimisations and variable selection. *Neurocomputing*, **55**, 337–346.

98 De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. (2000) Tutorial. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, **50**, 1–18.

99 Shen, M., Xiao, Y., Golbraikh, A., Gombar, V.K. and Tropsha, A. (2003) Development and validation of *k*-Nearest-Neighbor QSPR models of metabolic, stability of drug candidates. *Journal of Medicinal Chemistry*, **46**, 3013–3020.

100 Li, Y., Jiang, J.-H., Chen, Z.-P., Xu, C.-J. and Yu, R.-Q. (1999) Robust linear discriminant analysis for chemical pattern recognition. *Journal of Chemometrics*, **13**, 3–13.

101 Hawkins, D.M. and Kass, G.V. (1982) Automatic interaction detection, in *Topics* *in Applied Multivariate Analysis* (ed. D.M. Hawkins), Cambridge University Press, United Kingdom.

102 Breimann, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth, New York.

103 Norinder, U., Lidén, P. and Boström, H. (2006) Discrimination between modes of toxic action of phenols using rule based methods. *Molecular Diversity*, **10**, 207–212.

104 Quinlan, J.R. (1992) *C45 Programs for Machine Learning*, Morgan Kaufmann Publishers, USA.

105 Niculescu, S.P. (2003) Artificial neural networks and genetic algorithms in QSAR. *Journal of Molecular Structure (THEOCHEM)* , **622**, 71–83.

106 Johansson, U., König, R. and Niklasson, L. (2003) Rule extraction from trained neural networks using genetic programming. 13th International Conference on Artificial Neural Networks, Istanbul, Turkey, supplementary proceedings, pp. 13–16.

107 Johansson, U., Sönströd, C., König, R. and Niklasson, L. (2003) Neural networks and rule extraction for prediction and explanation in the marketing domain. The International Joint Conference on Neural Networks, IEEE Press, USA, Portland, OR, pp. 2866–2871.

108 Johansson, U., König, R. and Niklasson, L. (2004) The truth is in there – rule extraction from opaque models using genetic programming. 17th Florida Artificial Intelligence Research Symposium (FLAIRS) 04, AAAI Press, USA, Miami, FL, pp. 658–662.

109 Johansson, U. (2007) Obtaining accurate and comprehensible data mining models, PhD thesis, Institute of Technology, Linköping University.

110 Hewitt, M., Cronin, M.T.D., Madden, J.C., Rowe, P.H., Johnson, C., Obi, A. and Enoch, S.J. (2007) Consensus QSAR models: do the benefits outweigh the complexity? *Journal of Chemical Information and Modeling*, **47**, 1460–1468.

111 http://www.salfordsystems.com.

112 http://www.compumine.com.

113 Melville, P. and Mooney, R.J. (2005) Creating diversity in ensembles using artificial data. *Journal of Information Fusion (Special Issue on Diversity in Multiple Classifier Systems)*, **6**, 99–111.

114 O'Brien, S.E. and de Groot, M.J. (2005) Greater than the sum of its parts: combining models for useful ADMET prediction. *Journal of Medicinal Chemistry*, **48**, 1287–1291.

115 Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., Xie, Q. and Tong, W. (2004) Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis*, **19**, 365–377.

116 Manallack, D.T., Tehan, B.G., Gancia, E., Hudson, B.D., Ford, M.G., Livingstone, D.J., Whitley, D.C. and Pitt, W.R. (2003) A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *Journal of Chemical Information and Computer Sciences*, **43**, 674–679.

117 Merkwirth, C., Mauser, H., Schulz-Gasch, T., Roche, O., Stahl, M. and Lengauer, T. (2004) Ensemble methods for classification in cheminformatics. *Journal of Chemical Information and Computer Sciences*, **44**, 1971–1978.

118 Agrafiotis, D.K., Cedeno, W. and Lobanov, V.S. (2002) On the use of neural network ensembles in QSAR and QSPR. *Journal of Chemical Information and Computer Sciences*, **42**, 903–911.

119 van Rhee, A.M. (2003) Use of recursion forests in the sequential screening process: consensus selection by multiple recursion trees. *Journal of Chemical Information and Computer Sciences*, **43**, 941–948.

120 Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.M., Tong, W., Veith, G. and Yang, C. (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM workshop 521. *Alternatives to Laboratory Animals*, **33**, 155–173.

121 Zhang, H., Ando, H.Y., Chen, L. and Lee, P.H. (2007) On-the-fly selection of a training set for aqueous solubility prediction. *Molecular Pharmacology*, **4**, 489–497.

122 Guha, R., Dutta, D., Jurs, P.C. and Chen, T. (2006) Local lazy regression: making use of the neighborhood to improve QSAR predictions. *Journal of Chemical Information and Modeling*, **46**, 1836–1847.

123 Helma, C. (2006) Lazy structure–activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Molecular Diversity*, **10**, 147–158.

124 Zhang, S., Golbraikh, A., Oloff, S., Kohn, H. and Tropsha, A. (2006) A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *Journal of Chemical Information and Modeling*, **46**, 1984–1995.

125 Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. and Wold, S. (2006) Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications, 2nd edn, Umetrics.

126 Linusson, A., Gottfries, J., Olsson, T., Örnskov, E., Folestad, S., Nordén, B. and Wold, S. (2001) Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors. *Journal of Medicinal Chemistry*, **44**, 3424–3439.

127 Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, John Wiley & Sons, Inc., USA.

128 Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, A., Pettersen, J. and Bergman, R. (1998) Experimental design and optimisation. *Chemometrics and Intelligent Laboratory Systems*, **42**, 3–40.

129 Gottfries, J. and Wold, S. (2004) D-optimal onion designs in statistical molecular design. *Chemometrics and Intelligent Laboratory Systems*, **73**, 37–46.

130 Marengo, E. and Todeschini, R. (1992) A new algorithm for optimal distance-based experimental design. *Chemometrics and Intelligent Laboratory Systems*, **16**, 37–44.

131 Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H. and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, **17**, 241–253.

132 Clark, M. and Cramer, R.D., III (1993) The probability of chance correlation using partial least squares (PLS). *Quantitative Structure–Activity Relationships*, **12**, 137–145.

133 Golbraikh, A. and Tropsha, A. (2002) Beware of q2! *Journal of Molecular Graphics & Modelling*, **20**, 269–276.

134 van der Voet, H. (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, **25**, 313–323.

135 Tropsha, A., Gramatica, P. and Gombar, V.K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, **22**, 69–77.

15
Computational Absorption Prediction

Christel A.S. Bergström, Markus Haeberlein, and Ulf Norinder

Abbreviations

ADME	Absorption, distribution, metabolism, and excretion
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
Alog$P$	Ghose–Crippen–Viswanadhan octanol–water partition coefficient
BCS	Biopharmaceutics Classification System
BCUT	Burden, CAS, University of Texas descriptors
CART	Classification and regression tree
Clog$P$	Calculated partition coefficient between octanol and water
CMR	Calculated molar refractivity
F_a	Fraction absorbed
FPSA	Fractional polar surface area
GP	Genetic programming
G-REX	Genetic rule extraction
GSE	General solubility equation
HBA	Hydrogen-bond acceptors
HBD	Hydrogen-bond donors
HIA	Human intestinal absorption
LDA	Linear discriminant analysis
LOO-CV	Leave-one-out cross-validation
Mlog$P$	Moriguchi log P
MV	Molar volume
MW	Molecular weight
nHBA	Number of hydrogen-bond acceptors
nHBD	Number of hydrogen-bond donors
NN	Neural network
NPSA	Nonpolar surface area
nRB	Number of rotatable bonds
PDR	Physician's desk reference
P-gp	P-glycoprotein

PLS	Partial least square projection to latent structures
PSA	Polar surface area
QSAR	Quantitative structure–activity relationship
RMSE	Root mean square error
ROC	Receiver-operating characteristic
RP	Recursive partitioning
SVM	Support vector machine
TOPS-MODE	TOPological Substructural Molecular Design
TPSA	Topological polar surface area

Symbols

2D	Two-dimensional
3D	Three-dimensional
A	Abraham hydrogen-bond acidity parameter
log $D_{6.5}$	Logarithm of apparent partition coefficient at pH 6.5
log P	Logarithm of partition coefficient between octanol and water
log S	Logarithm of intrinsic solubility
N rule-of-5	Number of violations of the four rule-of-5 rules developed by Lipinski
r^2	Coefficient of determination (correlation coefficient)
q^2	Cross-validated coefficient of determination

15.1 Introduction

During the past 10 years starting with the publications of Lipinski and coworkers [1] and Palm and coworkers [2], a considerable amount of research has been performed to develop predictive computational models for intestinal absorption in humans. The purpose of these investigations has been to develop computationally fast and accurate models for *in silico* electronic screening of large virtual compound libraries.

This chapter will give a theoretical background of the oral absorption and then discuss the computational models that are based on the publicly available data sets. A short overview of the software for absorption prediction is also included in the discussion.

15.2 Descriptors Influencing Absorption

The intestinal wall is optimized to absorb fluids and nutrients while keeping away different xenobiotics. Which factors, from a theoretical point of view, are the most influential in intestinal absorption?

Let us assume passive diffusion as the main driving force for absorption. Passive diffusion can be calculated by applying Fick's first law to the flux through the intestinal wall. At each point i on the intestinal surface, the flux J_i is:

$$J_i = c_i P_i, \tag{15.1}$$

where c_i is the concentration of the drug at a point i and P_i is the permeability of the drug at the same point. Hence, the total mass m of absorbed drug at a time t can be written as:

$$m(t) = \int_0^t \iint_A c_i P_i \mathrm{d}A\mathrm{d}t, \tag{15.2}$$

where A is the total area of the intestinal tract. The fraction absorbed (F_a) is defined by the total mass absorbed divided by the given dose of the drug:

$$F_\mathrm{a} = \frac{m(\infty)}{\mathrm{Dose}}. \tag{15.3}$$

This brief simplified analysis shows that the absorption mainly depends on the concentration at the intestinal wall, the permeability of the drug, and the given dose [3]. Let us analyze these three factors further.

15.2.1 Solubility

The concentration of the drug at a point i at the intestinal wall depends on the dissolution rate and the gastrointestinal transit. The dissolution rate of a drug molecule is affected by the energy difference that arises when the compound dissolves with similar molecules in the crystal or the molecules of the formulation and instead forms bonds with the components in the intestinal fluid. If this process is related to a high-energy penalty, the dissolution rate of the compound will be low, whereas if the process releases energy, the dissolution rate will be high. The most common factor for drug molecules is that the dissolution process is related to an energy penalty of some extent. Some of the factors influencing the dissolution rate and the maximum solubility obtained in the intestinal fluid are the formulation, particle size, particle aggregation, pH in different segments of the intestine, food content, and physicochemical properties of the drug molecule. Considering these factors one can expect a large variation in the solubility for the same drug with different formulations in different subjects, for example, humans.

If we look at the physicochemical factors governing solubility, among the first identified were log P [4] and melting point [5, 6]. The lipophilicity is often calculated theoretically using, among other techniques, fragment-based approaches. It has lately become apparent that the log P is not always correctly calculated for new drug-like compounds, and for new AstraZeneca and Pfizer compounds, the root mean square error (RMSE) has been reported to be 0.84–1.46 on a log scale [7, 8]. Hence, the calculated (Clog$P$) value for such compounds becomes 7–29-fold falsely calculated.

Several approaches have been applied to predict the melting point, but all of them result in prediction errors of 35–45 °C [9, 10] and can therefore not be regarded as accurate enough to be included in solubility calculations. Hence, the prediction of solubility from the general solubility equation (GSE) established by Yalkowsky and coworkers [6] still requires the experimentally determined melting point. Other typical molecular descriptors included in solubility predictions are molecular size, hydrogen bonding, nonspecific van der Waals interactions, aromaticity, flexibility, and dipole moment [11–17].

15.2.2
Membrane Permeability

The other mechanistically important component for intestinal absorption is the actual passage over the cell membrane. Before reaching the cell membrane, the drug molecule needs to diffuse through the unstirred water layer. However, theoretical considerations suggest that in most cases this diffusion is not the rate-limiting step for permeability. When at the cell wall, there are a number of different mechanisms for which a compound can be transported across the cell barrier. The most important mechanisms are transcellular passive diffusion, paracellular diffusion, active transport with a transporter, and transcytosis (Figure 15.1). In addition, the drug can be metabolized in close connection to the luminal cell membrane by CYP3A4.

For a theoretical model, each mechanism has to be described in a different manner.

If we restrict ourselves to discuss transcellular diffusion, there are numerous theoretical approaches, which differ depending on the underlying assumptions. In general, permeability mainly depends on lipophilicity (log P or Clog$P$), molecular

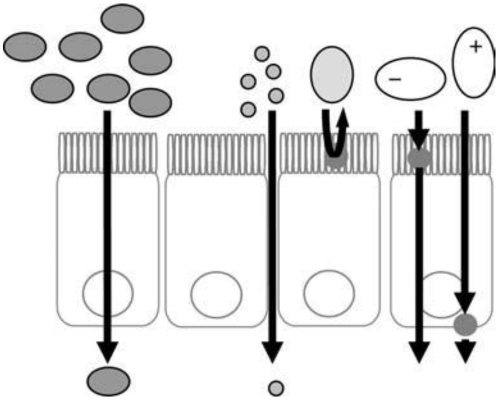

Figure 15.1 The following routes are available for permeating the intestinal wall (from the left-hand side): the transcellular route, mainly used by nonpolar and medium-sized molecules; the paracellular route, mainly used by polar and small molecules often bearing a net charge; and energy-dependent active transport processes, which efflux (secret) transporter substrates or influx (take up) transporter substrates. Each transport protein has its own substrate specificity.

weight (MW), and measures of hydrogen-bonding capacity or polarity [18]. If we use the findings from computational models on permeability measurements of cell lines, for example, Caco-2 cells, the following factors are among the most important: polar surface area (PSA), nonpolar surface area (NPSA) and/or lipophilicity, hydrogen-bond acceptors (HBA), hydrogen-bond donors (HBD), polarity (charge distribution), MW, size, shape, and degree of ionization [19–22].

Amidon *et al.* [3] devised a Biopharmaceutics Classification System (BCS), where they divided drugs into four different classes based on their solubility and permeability: class 1 (high solubility, high permeability), class 2 (low solubility, high permeability), class 3 (high solubility, low permeability), and class 4 (low solubility, low permeability); see also Chapter 19. The rate-limiting step to drug absorption and hence the factors affecting drug absorption will differ depending on which class the drug belongs to. For class 2, the rate-limiting step is dissolution, and the permeability plays a minor role. For class 3, however, the permeability is rate limiting and the dissolution has very little influence on the absorption. Given the above-mentioned considerations, it is difficult to believe that it would be possible to fit drugs from all four classes into one single model. However, it is worth to note that several molecular descriptors highly influence both permeability and solubility. For example, it has been suggested that the four BCS classes can be divided solely by considering the MW and PSA [23].

15.3 Computational Models of Oral Absorption

15.3.1 Quantitative Predictions of Oral Absorption

To date, a large number of models aiming at quantitative prediction of oral absorption are available, either published in scientific journals (Table 15.1) or included in commercial software (Table 15.2). These models are often based on human F_a data, also known as the human intestinal absorption (HIA), extracted from the literature. Original sources are generally clinical studies, the physician's desk reference (PDR) or product specifications. Occasionally, substitute parameters are used for absorption, one of the most commonly applied being the permeability in Caco-2 cell monolayers [24]. The Caco-2 cells originate from colon carcinoma, which when cultured *in vitro* easily form an intact monolayer mimicking the intestinal epithelium. The advantages with using such a system are obvious; for instance, a large number of compounds can easily be screened for their intestinal permeability at a low cost and without facing ethical restrictions. However, the disadvantages are also clear – the permeability obtained in the *in vitro* cell system reflects only the F_a after oral administration if the absorption of the compound is limited by permeability. Hence, for compounds that have poor solubility (BCS class 2 and 4) and/or stability issues and are subjected to active transport, such an *in vitro* surrogate marker for F_a is not applicable.

Table 15.1 Publications on quantitative models of human fraction absorbed.

Publication	Title	Data set (n)	Technique	RMSE test set	Descriptors
Raevsky *et al.*, *Quantitative Structure–Activity Relationships*, 2000 [34]	Quantitative estimation of drug absorption in humans for passively transported compounds on the basis of their physicochemical parameters	32	MLR, MNLR	n.a.	MW, log D, hydrogen-bond descriptors ($n = 5$)
Verma *et al.*, *Journal of Computer-Aided Molecular Design*, 2007	Comparative QSAR studies on PAMPA/modified PAMPA for high-throughput profiling of drug absorption potential with respect to Caco-2 cells and human intestinal absorption	68	MLR	n.a.[a]	Clog P, hydrogen-bond descriptors ($n = 3$)
Wessel *et al.*, *Journal of Chemical Information and Computer Sciences*, 1998 [25]	Prediction of human intestinal absorption of drug compounds from molecular structure	86	GANN	16% ($n = 10$)	Topological and electronical 2D, geometrical 3D ($n = 6$)
Agatonovic-Kustrin *et al.* *Journal of Pharmaceutical and Biomedical Analysis*, 2001	Theoretically derived molecular descriptors important in human intestinal absorption	86	GANN	17% ($n = 10$)	Constitutional, topological, chemical, geometrical, and quantum mechanical ($n = 15$)
Niwa, *Journal of Chemical Information and Computer Sciences*, 2003 [35]	Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures	86	GRNN	23% ($n = 10$)	CMR, Clog P, 2D topological descriptors ($n = 7$)

Deretey *et al.*, *Quantitative Structure–Activity Relationships*, 2002	Rapid prediction of human intestinal absorption	124	MNLR	17% ($n=31$)	Clog P, hydrogen-bond descriptor ($n=2$)
Abraham *et al.*, *European Journal of Medicinal Chemistry*, 2002	On the mechanism of the human intestinal absorption	127	MLR	n.a.	Abraham descriptors ($n=5$)
Deconinck *et al.*, *Journal of Pharmaceutical and Biomedical Analysis*, 2005	Prediction of gastrointestinal absorption using multivariate adaptive regression splines	140	MARS	n.a.	2D and 3D descriptors from Dragon ($n=9$)
Deconick *et al.*, *Journal of Pharmaceutical and Biomedical Analysis*, 2007	Exploration of the linear modeling techniques and their combination with multivariate adaptive regression splines to predict gastrointestinal absorption of drugs	141	PLS-MARS	n.a.	2D and 3D descriptors from Dragon ($n=9$ PCs from the PLS analysis)
Zhao *et al.*, *Journal of Pharmaceutical Sciences*, 2001 [26]	Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity (QSAR) with Abraham descriptors	169	MLR	14% ($n=131$)	Abraham descriptors ($n=5$)
Liu *et al.*, *Journal of Computer-Aided Molecular Design*, 2005	The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine	169	SVM	14% ($n=56$)	CODESSA descriptors: constitutional, topological, electrostatic, geometrical, and quantum mechanical ($n=5$)
Gunturi and Narayanan, *QSAR & Combinatorial Science*, 2007 [41]	*In silico* ADME modeling. 3. Computational models to predict human intestinal absorption using sphere exclusion and kNN QSAR methods	174	kNN-QSAR	9% ($n=49$)^{b}	Structural, physicochemical, geometrical, and topological ($n=4$)

(*Continued*)

Table 15.1 (*Continued*)

Publication	Title	Data set (n)	Technique	RMSE test set	Descriptors
Iyer *et al.*, *Molecular Pharmaceutics*, 2007	Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis	188	MLR	n.a.	2D and 3D solute descriptors ($n = 7$)
Zhao *et al.*, *Journal of Pharmaceutical Sciences*, 2002	Rate-limited steps of human oral absorption and QSAR studies	238	MLR, MNLR	n.a.[c]	Abraham descriptors ($n = 5$)
Klopman *et al.*, *European Journal of Pharmaceutical Sciences*, 2002 [31]	ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans	467	MLR	12% ($n = 50$)	Physicochemical descriptors group contribution approach ($n = 37$)
Votano *et al.*, *Molecular Diversity*, 2004 [27]	New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors	612	ANN	16% ($n = 195$)	Clog *P*, PSA, and topological descriptors ($n = 10$)

Abbreviations used: multiple linear regression (MLR), multiple nonlinear regression (MNLR), genetic algorithm neural network (GANN), general regression neural network (GRNN), multivariate adaptive regression splines (MARS), partial least square projection to latent structures (PLS), support vector machine (SVM), *k*-nearest neighbor quantitative structure–activity relationship (*k*NN-QSAR), artificial neural network (ANN), root mean square error (RMSE).n.a., not applicable; the fraction-absorbed model has not been validated with a test set or the result of a validation is not given.

[a] The model was tested on 11 compounds and log F_a was used as response. Since several of the compounds were severely falsely predicted, our calculation of the RMSE of the test set was more than 100% indicating that the model cannot be used for predictions. The authors do not comment this in their paper.

[b] Based on model 1.

[c] Only a qualitative validation has been performed.

Table 15.2 Examples of commercial software available for computational prediction of human fraction absorbed and related properties.

Software	Company	Dissolution	Sol	Perm	Trp	Oral bioavailability	F_a	Other PK
ADME boxes/batches	Pharma Algorithms		●	●	●	●	●	●
Cerius^2	Accelrys		●				●	●
Chem Silico modules	ChemSilico		●	●			●	●
KnowItAll ADME/Tox	Bio-Rad Laboratories		●	●		●	●	●
QikProp	Schrödinger		●	●				●
QMPRPlus	Simulations Plus		●	●			●	●

Bullets show properties predicted in each of the reported software. The following abbreviations are used: solubility (Sol), membrane permeability (Perm), transporters (Trp), human intestinal absorption (F_a), pharmacokinetic properties (PK).

15.3.1.1 Responses: Evaluations of Measurement of Fraction Absorbed

To obtain the responses, that is, the experimentally measured value for F_a, several different techniques have been applied. We will briefly go through the procedure performed in the establishment of three different data sets, namely, the "Wessel" data set [25], the "Zhao" data set [26], and the "Votano" data set [27]. These data sets were selected based on their repeated use in model development and/or their large size.

The first large data set for F_a prediction was created by Wessel and coworkers [25], who compiled F_a data for 86 compounds based on results found in 151 studies. Each reference was carefully reviewed to ensure that the value used was indeed the F_a data and not the absolute oral bioavailability, since the latter can be lower than the F_a. Furthermore, the data were controlled to not be dose-dependent or disease-dependent, that is, only results based on healthy volunteers were used. The 86 compounds were divided into a training set of 76 compounds and a test set of 10. The authors claimed that they included all poorly absorbed compounds available at the time for the construction of the data set. Furthermore, they did not include all highly absorbed compounds available with the intention to not let the highly absorbed compounds skew the data set and thereby affect the results of the modeling. Even though these precautions were taken, the final training set ($n = 76$) consisted of 49 compounds with more than 80% absorbed and 7 compounds below 20%. It has lately become apparent that some of the compounds included in the "Wessel" data set are substrates for active transporters and therefore the data set may not be optimal for F_a modeling. The clinical relevance of such active transport has though been debated in the literature, and many claim that active transporters in most cases do not affect the absorption rate in the intestine due to high concentrations of the drug available. However, for specific molecules, the active transport is the dominating uptake mechanism and is of pharmaceutical importance to some peptides, β-lactam antibiotics, and ACE inhibitors [28]. Hence, the generalization that active transport does not have a significant effect on the uptake from the intestine can lead to significant false predictions of such molecules.

Zhao and colleagues [26] published a quantitative structure–activity relationship (QSAR) for F_a based on a data set of 241 compounds. From 244 papers, the following properties were recorded for the compounds:

- the absorption data;
- the oral or absolute bioavailability;
- the percentage of cumulative urinary excretion of unchanged drug and metabolites following oral and intravenous administration;
- the percentage of metabolites in urine or first-pass effect following oral and intravenous administration;
- the percentage of unchanged drug in urine following oral and intravenous administration;
- the percentage excretion of drug in bile following oral and intravenous administration;
- the percentage of cumulative excretion of drug in feces following oral and intravenous administration;
- total recovery of drug in urine and feces following oral and intravenous administration.

The information was thereafter used to sort the response into classes that depend on the quality of the data. This resulted in 169 compounds in the group sorted as having good or OK quality of the response data, which were divided into a training set of 38 compounds and a test set of 131 compounds. Out of these, 23 compounds of the training set had an F_a larger than 80%, and only 6 compounds had an F_a of less than 20%. The number of compounds for the test set was 96 displaying more than 80% absorbed, but only 2 compounds with fraction absorbed data less than 20%. This, together with the histogram of F_a for the data set, clearly shows that the "Zhao" data set is heavily skewed toward compounds with high fraction absorbed. The clear reason for this is that most marketed drugs already have been optimized for absorption, and hence troublemakers have failed during the development process. From a computational model development viewpoint, this results in the tools developed to be good at identifying a compound with good absorption, whereas it will be difficult to identify poorly absorbed compounds since this chemical space has not been well represented in the model development. The skewness of data sets used for prediction of oral absorption has also been identified and treated recently [29].

The largest data set we have found published for quantitative prediction of F_a is the data set treated by Votano and coworkers [27], who used a training set of 417 compounds and a test set of 195 compounds for model development and validation, respectively. The data came from several different sources [26, 30, 31], the PDR [32], and therapeutic drugs [33], and the compounds included were scrutinized to remove substances reported to be actively transported across the intestinal membrane. A true objective validation of this data set, however, cannot be performed, since the authors do not reveal the compounds included in the study. However, the authors state that a large fraction of the compounds showed a high F_a. Only 25% of the compounds displayed an F_a value less than 60%, whereas 54% had an F_a value more than 80%. The authors divided the complete data set into three groups: two groups were formed through the use of a molecular weight cut-off rule, to handle paracellular ($\leq$251 Da) and transcellular ($\geq$252 Da) transport separately. Again, the compounds included in each cluster are not publicly available, making this effort of mechanistic modeling difficult to evaluate. Finally, a subset of 23 compounds carrying a formal positive charge was excluded from the two groups and modeled separately. The results from the three different models were thereafter combined, resulting in an RMSE of the training set of 11.5% and an $\text{RMSE}_{\text{test set}}$ of 15.9%. Of these compounds, 10 were not well predicted by the model, and these were probenecid, gilbornuride, indomethacin, meropenem, cymarin, piretanide, lodoxamide, etretinate, exemestane, and carbenoxolone. The authors could not find any chemical reason for the bad predictions (27–49% falsely predicted), but they speculate that the solubility may be a limiting factor for the absorption *in vivo*. The obtained "transcellular" model was based on lipophilicity, PSA, and other hydrogen-bond descriptors. Unfortunately, the authors do not reveal which descriptors were most important for the prediction of the "paracellular" data set and the 23 charged compounds, and therefore conclusions regarding absorption mechanisms based on molecular descriptors cannot be drawn.

15.3.1.2 **Model Development: Data sets, Descriptors, Technologies, and Applicability**

Quantitative predictions of oral absorption aim at returning an accurate percentage absorbed from the prediction. Going through the models published for prediction of F_a reveals that the size of the data sets used differs tremendously, from 32 compounds [34] to more than 600 [27] (Table 15.1). Depending on the size of the data set, and hence the volume and the density of the chemical space investigated, the obtained model will be more or less generally applicable. Small data sets as well as data sets including a large series of homologous structures are often generally less applicable than models based on larger and structurally diverse data sets.

Most commonly applied descriptors for the development of F_a models have different 2D and 3D properties. These are physicochemical, topological, electrostatical, or geometrical. Several different software programs for the calculation of these descriptors are available, which are rapid and allow several hundreds of descriptors to be calculated. Typical descriptors are discussed in detail in Section 14.2.

In Table 15.1, quantitative predictions of F_a published during the past 10 years are compiled. As can be seen, the problem of predicting F_a has been investigated using quite different statistical techniques, and a variety of linear and nonlinear methodologies have been applied. In general, the models predict the training sets within 10–15% range of the experimental value, even though the true accuracy is difficult to evaluate. To do so, the obtained models must be challenged with test sets composed of compounds that have not been included in the model development. This is not performed in all studies, sometimes due to the limitation of compounds available or selected for the study. However, when test sets have been used, the range of accuracy for the test set is within 9–23%. This indicates that there is a large uncertainty in the value of absorption obtained from the prediction, as a result of which a compound can easily be falsely predicted by as much as 20% in absorption. Thus, there is a tendency to perform qualitative predictions, in which the percentage absorption is binned into classes such as low, intermediate, and high F_a. These investigations will be discussed in Section 15.3.2. However, we also note that quantitative models are sometimes recommended to be used more as a sorting tool than for the actual value resulting from the prediction. This is exemplified in the study performed by Niwa [35], who treated the "Wessel" data set with a general regression neural network (NN) and a probabilistic NN based on calculated molar refractivity (CMR), Clog*P*, and 2D topological descriptors. As a result of the general regression NN, the test set was predicted with an RMSE of 22.8%, indicating that the model is not really quantitatively reliable. When the same data set was used for classification purposes, the results improved, and 80% of the test set was correctly predicted (see further description of this study in Section 15.3.2). In this study, the effects of skewed data sets also became clear. A large majority of the responses had F_a values of more than 80%, as a result of which all the well-absorbed compounds were correctly sorted by the model whereas the poorly absorbed were partly misclassified.

15.3.2
Qualitative Predictions of Oral Absorption

15.3.2.1 Model Development: Data sets, Descriptors, Technologies, and Applicability

Owing to large uncertainties in measured F_a values as well as the uneven distribution of the poorly and well-absorbed compounds, it is rather common to derive qualitative *in silico* models for F_a instead of quantitative models (Table 15.3). For this purpose, the F_a (0–100%) is split (binned) into two or more classes. As always, there is a potential danger with binning continuous data since poor binning may disrupt the underlying data structure of a continuous variable.

Zmuidinavicius *et al.* [30] have used a compilation of compounds both from the "Zhao" and "Wessel" data sets and from some additional sources such as therapeutic drugs [33, 36, 37]. The data set covered over 1000 compounds. After questionable data and compounds influenced by active transport were removed, the data set consisted of 977 compounds. Unfortunately, the authors do not disclose more than a sample data set of some 200 compounds in their publication. The structures were described by properties such as Abraham descriptors (see Section 14.2.3.3), hydrogen-bonding parameters, log *P*, PSA, and the number of rotatable bonds (nRB). Structural descriptors of fragment-type were also used to characterize the investigated compounds. The authors divided the F_a absorption into two classes with "good" absorption defined as $F_a > 15\%$ and "poor" absorption defined as $F_a < 10\%$, respectively. A recursive partitioning (RP) approach was employed by the authors to derive a small set of rules, less than 10, which correctly explained ~94.2% of the data. The data set is, however, rather skewed with ~90% of the compounds belonging to the "good" class of compounds. Important parameters for determining the correct class were log *P*, PSA, and Abraham alpha (A) hydrogen-bond acidity parameter. From the publication, it is not possible to determine how the authors validated their model with respect to both internal cross-validation and external validation (see Section 14.5 for details of

Table 15.3 Publications on classification models of human fraction absorbed.

Publication	Title	Data set	Technique	Correct test set	Descriptors
Clark, *Journal of Pharmaceutical Sciences*, 1999	Rapid calculation of polar surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption	94	SR	91% ($n = 74$)	PSA
Subramanian and Kitchen, *Journal of Molecular Modeling*, 2006	Computational approaches for modeling human intestinal absorption and permeability	121	SR	70%[a] ($n = 46$)	PSA
Deconinck *et al.*, *Journal of Pharmaceutical and Biomedical Analysis*, 2005 [46]	Classification of drugs in absorption classes using the classification and regression trees (CART) methodology	141	CART	85% ($n = 27$)	2D and 3D descriptors from Dragon, HyperChem, and ACDlabs ($n_{\text{final}} = 9$)
Sun, *Journal of Chemical Information and Computer Sciences*, 2004 [42]	A universal molecular descriptor system for prediction of log P, log S, log BB, and absorption	169	PLS-DA	n.a.	Atom types ($n_{\text{final}} = 3$ PCs from the PLS-DA analysis)
Wegner *et al.*, *Journal of Chemical Information and Computer Sciences*, 2004	Feature selection for descriptors based classification models. 2. Human intestinal absorption (F_a)	196	GA-SEC	n.a.	2D and 3D from several sources ($n_{\text{final}} = 10$–245)
Cabrera Perez *et al.*, *European Journal of Medicinal Chemistry*, 2004	A topological sub-structural approach for predicting human intestinal absorption of drugs	209	LDA	93%	TOPS-MODE ($n_{\text{final}} = 3$)
Egan *et al.*, *Journal of Medicinal Chemistry*, 2000	Prediction of drug absorption using multivariate statistics	234	Pattern recognition	n.a.	Clog P, PSA, MW

(*Continued*)

Table 15.3 (*Continued*)

Publication	Title	Data set	Technique	Correct test set	Descriptors
Klon *et al.*, *Journal of Chemical Information and Modeling*, 2006 [49]	Improved näve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction	264	Näve Bayes with continuous numerical	81%	2D and 3D descriptors from Dragon ($n_{\mathrm{final}} = 3$)
Zmuidinavicius *et al.*, *Journal of Pharmaceutical Sciences*, 2003 [30]	Classification structure–activity relations (C-SAR) in prediction of human intestinal absorption	977	HC and RP	n.a.	2D and 3D descriptors ($n_{\mathrm{final}} = 2$–3)
Bai *et al.*, *Journal of Chemical Information and Computer Sciences*, 2004 [44]	Use of classification regression tree in predicting oral absorption in humans	1260	CART	65%[b]	2D and 3D descriptors ($n_{\mathrm{final}} =$ n.s.)
Hou, *et al.*, *Journal of Chemical Information and Modeling*, 2007	Use of support vector machines in predicting oral absorption in humans	578	SVM	96.9% ($n = 98$)	log $D_{6.5}$, TPSA, nHBD, MW, MV, *N* rule-of-5 and N+

Abbreviations used: sigmoidal regression (SR), classification and regression trees (CART), partial least square projection to latent structure discrimination analysis (PLS-DA), genetic algorithms based on Shannon entropy cliques (GA-SEC), linear discriminant analysis (LDA), hierarchical clustering (HC), recursive partitioning (RP), topological substructural molecular design (TOPS-MODE). n.a., not applicable; the fraction-absorbed model has not been validated with a test set or the result of a validation is not given. n.s., not stated.

[a] Using a cutoff limit between poor and good absorption of 80%, and Equation 12.

[b] Six classes were used in this evaluation: 0–19, 20–31, 32–43, 44–59, 60–75, and 76–100%.

the importance of model validation). This makes the results obtained somewhat unreliable with respect to the forecasting ability of the derived model.

Pérez and coworkers [38] have used linear discriminant analysis (LDA) to develop a classification model for F_a. The data set studied is, in part, based on the "Zhao" data set but with additional compounds from Benet *et al.* [39] and consists of 209 compounds, which the authors divided into a training set and a test set of 82 and 127 compounds, respectively. The F_a was divided into three classes: highly ($F_a > 70\%$), moderately (F_a between 30 and 70%), and poorly ($F_a < 30\%$) absorbed compounds. The authors used a methodology called TOPS-MODE [40], which is based on the calculation of spectral moments of a bond matrix, whose entries are ones or zeros if the corresponding bonds are adjacent or nonadjacent. The diagonal elements of the bond matrix in this study were weighted by PSA, hydrophobicity, molar refraction, atomic charge, and atomic mass. This weighting aspect of the descriptor matrix (bond matrix) makes computed descriptors of the TOPS-MODE method similar to the ones obtained using the BCUT methodology (see Ref. [41] for further description of the BCUTmethod). The authors actually derived two models to be used sequentially. The purpose of the first model was to distinguish the poorly absorbed compounds from the highly and moderately absorbed ones while the second model was designed to do the opposite, that is, distinguish the highly absorbed compounds from the moderately and poorly absorbed ones. Pérez *et al.* performed extensive validation of their model apart from the training and test set selection mentioned above. They also conducted leave-one-out cross-validation (LOO-CV) on their training set and tested the derived model with an additional external test set of some 100 compounds. The predictive ability of the derived models with respect to both external and internal validation is impressive with accuracies of between 80 and 94% for the various validation sets. The authors found variables related to log P, PSA, the number of bonds in the molecules, and the size of the molecules to be important for discriminating the three absorption classes.

Sun [42] has also investigated the "Zhao" data set using atom-type descriptors, as the author derived two models – a 2-class model and a 3-class model. The three classes were defined as class 1 $F_a > 80\%$, class 2 $F_a = 20–80\%$, and class 3 $F_a < 20\%$. For the 2-class model, the division between classes was set at 20%. The atom-type classification employed in this study was based upon identifying a particular type depending upon several factors, namely, its element, its aromaticity, its neighboring atoms, and whether the atom is in a ring or not. This atom classification scheme resulted in 218 different descriptors. Sun used partial least square projection to latent structures (PLS) [43] (see Section 14.3 for further details) as statistical engine for deriving the relationship and cross-validation as internal validation technique. For the 3-class model, the analysis resulted in a five-component model with a coefficient of determination (r^2) of 0.92 and a cross-validated coefficient of determination (q^2) of 0.79. The corresponding values for the 2-class case were 0.94 and 0.86, respectively. Unfortunately, Sun neither reports the accuracy of the predictions nor does the investigation use external validation for determining the forecasting ability of the derived model.

Bai *et al.* [44] investigated approximately 1260 drugs from the OraSpotter human pharmacokinetic database [45] using CART rule-based modeling. They divided the F_a into six classes (0–0.19, 0.2–0.31, 0.32–0.43, 0.44–0.59, 0.6–0.75, and 0.76–1) and used 28 different molecular descriptors that included variables such as log P, number of HBDs and HBAs, MW, and PSA, as well as counts of some functional groups. The data set was randomly split into a training set and a test set consisting of 899 and 362 compounds, respectively. The accuracy was 65% for the prediction of the correct class and 80.4% accuracy within one class error. Furthermore, Bai and coworkers additionally tested three more diverse data sets that consisted of 67, 90, and 37 compounds and resulted in 85.1, 74.4, and 86.4% accuracy, respectively, within one class error. From the investigations, the authors concluded that the CART model performed better for high and low absorption but performed not so well for the intermediate classes between 0.32 and 0.59. As with most data sets, the data set used by Bai and coworkers was also skewed and had relatively few compounds in the intermediate range between 0.32 and 0.59. This may, in part, explain the somewhat poor predictive ability for this kind of compounds.

Deconinck *et al.* [46] have also used CART to model F_a for the "Zhao" data set. They investigated 141 compounds using both Dragon [47] and Hyperchem [48] descriptors (>1400 descriptors). The authors divided the F_a range into five classes: class 1, 0–25%; class 2, 26–50%; class 3, 51–70%; class 4, 71–90%; and class 5, more than 90%. For internal validation, a 10-fold cross-validatory procedure was employed. Deconinck and coworkers developed three models: the first, second, and third models were based on all available descriptors, all available 2D descriptors, and all available 3D descriptors, respectively. They found that the first model based on all descriptors performed best. Furthermore, the authors also found that the first five splits in the CART tree were defined by 2D descriptors. Thus, the investigation indicated that the rough classification of the compounds was performed by 2D descriptors and then refined, by additional splits in the model further down the tree, by 3D descriptors. The predictive power of the three models was tested with an external test set consisting of 27 compounds, that is, ~20% of the size of the training set. The three models predicted the test with accuracies of 88.9, 85.2, and 77.8%, respectively. The data set used by Deconinck and coworkers is well documented so that other researchers may investigate the same data set.

Another study on the "Zhao" data set with some additional compounds was performed by Klon and coworkers [49]. After removing P-glycoprotein (P-gp) substrates and compounds for which human intestinal absorption was either not reported or could not be related to passive intestinal absorption, the data set consisted of 264 structures. The authors randomly assigned 75% of the compounds to the training set (205 entries) while the remaining compounds constituted the external test set (59 entries). Unfortunately, Klon *et al.* do not reveal the names or structures of the compounds included in the training and test sets, which makes it difficult for other researchers to verify or reinvestigate the data set in question. The authors used three different implementations of naïve Bayesian classifiers – one in-house-developed method and two commercially available [50, 51]. The former method uses a Gaussian approach while the latter two are based on a Laplacian implementation. The authors treated the F_a investigation as a binary classification with three different cut-offs (90, 80, and 70%, respectively) for defining the high (above the cut-off) and the low (below the cut-off) absorbed compounds. The measure of performance by the derived models was also estimated in a somewhat different fashion compared to what is usually the case. Normally, accuracy is used as the criteria of how well the model performs. In this investigation, a well-known measure within the field of machine learning was used, namely, the receiver-operating characteristic (ROC) curve. The ROC curve is a measure of the models' sensitivity, that is, the ability to identify true positives, and specificity, that is, the ability to avoid false negatives. The area under the ROC curve serves as a measure of the predictive ability of the derived model. A value of 1.0 represents a perfect model that is able to discriminate perfectly between true positives and true negatives, while 0.5 is indicative of a model with random performance, that is, no predictive ability. The structures in the data set were described by three sets of descriptors: Dragon descriptors [47], Pipeline Pilot descriptors [51], and the ADME Profiler descriptors FPSA (a polar surface area descriptor) and Alog*P* (a calculated log *P* descriptor) [52]. Finally, the Dragon descriptors were selected so that only variables with an absolute correlation with F_a of 0.7 were retained. Also, pairwise highly correlated variables were removed keeping only one of the descriptors. After redundant descriptors were removed, only the hydrophilic factor (Hy), TPSA(NO), and the Moriguchi log *P* (Mlog*P*) remained. The Pipeline Pilot descriptors were an extended connectivity fingerprint with a neighborhood size of six bonds (FCFP_6), Alog*P*, MW, the number of hydrogen-bond donors (nHBD), the number of hydrogen-bond acceptors (nHBA), the number of rotatable bonds, and PSA defined by nitrogen and oxygen atoms (PSA(NO)). The authors found that the Gaussian implementation outperformed both the Pipeline Pilot and the binary QSAR implementations. The area under ROC curve varied from 0.70 using the FPSA and AlogP98 descriptors at the 90% cut-off for good absorption to 0.91 using the selected Dragon descriptors at 70% cut-off.

Hou and coworkers [53] have also investigated HIA using a data set of 578 compounds and support vector machine (SVM) technology. The data set studied in this work is a compilation from the Palm, Wessel, and Zhao data sets. Eleven different descriptors were used (topological polar surface area (TPSA), the octanol–water partitioning coefficient (log *P*), the apparent partition coefficient at pH 6.5 (log $D_{6.5}$), the number of violations of the four rule-of-5 rules developed by Lipinski (*N* rule-of-5), the number of hydrogen-bond donors and acceptors, the intrinsic solubility (log *S*), the number of rotatable bonds, the molar volume (MV), the molecular weight, and a binary indicator (N+) representing the existence of a positively charged N atom). The authors divided the data set into a 480-molecule training set and a 98-molecule test set. Ten SVM classification models were developed to investigate the impact of different individual molecular properties on F_a. The final model consisted of the seven parameters: log $D_{6.5}$, TPSA, nHBD, MW, MV, *N* rule-of-5, and N+. The overall correctness of the model is quite impressive: 97.8 and 94.5% of the good and poor classes, respectively, were correctly classified for the training set while for the test set, the model achieved corresponding accuracies of 97.8 and 100%, respectively.

15.3.2.2 An Example Using Genetic Programming-Based Rule Extraction

The example described here employs genetic programming (GP) (see further description of the method in Section 14.3.2.2) and the genetic rule extraction (G-REX) algorithm [54, 55].

We have used the data set published by Hou and coworkers [53]. The data set consists of 578 compounds. The compounds were divided into two classes depending upon the measured F_a value. Compounds with an F_a higher than 30% were assigned to class "high" while the remaining compounds were designated "low." The distribution of classes was, as is unfortunately the case for data sets of this kind, rather skewed with 407 compounds belonging to class "high" while only 73 compounds belonging to class "low." The training data were randomly divided into a training set and a validation set consisting of 380 and 100 compounds, respectively, and the classes were balanced internally by adding multiple copies of each object. The data set is available through Ref. [53]. Crossover and mutation for the genetic algorithm were set to 0.8 and 0.001, respectively. G-REX was applied and a rather simple model emerged (see Figure 15.2) with good fit and predictive ability. The actual model consists of four rules using parameters N+, *N* rule-of-5, log $D_{6.5}$, and MV (Figure 15.2).

In the external test set, the "high" and the "low" compounds were predicted with accuracies of 96.8 and 100%, respectively. The corresponding accuracies for the training set and the validation set are 93.1%, 93.3% and 96.6%, 100%, respectively. The G-REX technique applied in this study thus performed as good as the SVM model from the original study in Ref. [53] with respect to the external predictive ability. A possible advantage of the G-REX-derived model is the simplicity and transparency of the model that makes it quite attractive for further use.

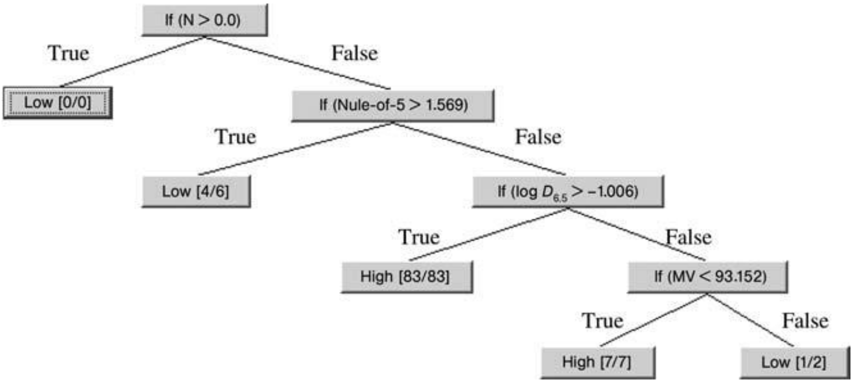

Figure 15.2 Genetic programming F_a classification model for the data set in Ref. [53]. Compounds with F_a higher than 30% were assigned to class "high" while the remaining compounds were designated "low."

15.3.3 Repeated Use of Data Sets

The available literature within the field of prediction of oral absorption reveals that the same literature data sets are repeatedly used. One of the most used data sets to date is the "Palm" data set [56], which is included in almost all large data sets, the "Wessel" data set [25], and the "Zhao" data set [26]. All of these are based on marketed drugs and hence are heavily skewed toward data reflecting good absorption. The scientific interest has been strongly directed toward which technique or which descriptor space is the best for obtaining models with good external predictivity, but it can be questioned how much more we can learn from these data sets. Can we use them to produce predictive models applicable to the drug discovery process of today? Can we use the results to predict the oral absorption of new lead structures? One way to improve the predictions and to make them applicable also for new drug structures is to go back to the experimental settings and produce response data for such compounds. Instead of using F_a data from traditionally used drugs, which are sometimes of intermediate quality, it may be more successful to model the major underlying mechanisms for absorption, for example, solubility and membrane permeation. These investigations can be performed on new chemicals and proof of concept can be performed in animal studies. To virtually predict human F_a of new chemical entities based on such *in vitro* and *in vivo* data with high accuracy for both poorly and well-absorbed compounds is one of the future scientific challenges in this area.

15.4 Software for Absorption Prediction

A large number of software programs are available for prediction of oral absorption and other pharmacokinetically relevant properties (Table 15.2). Is it possible to know beforehand which one to use? Evaluations to test the performance of the software are performed with the help of a standard data set. One such evaluation compared the performance of GastroPlus and iDEA, two simulation software, among others, for predicting oral absorption, and found them to perform quite equally [57]. One reason for this can be the issues discussed in Section 15.3, that is, the repeated use of data sets. It is likely that the training sets used for the model development are similar and hence the performance becomes similar. However, the training set used and the applicability domain for the models incorporated in the software are generally not stated and thus it becomes difficult to know beforehand which one to use. Therefore, the best way to decide which software to choose for future use is to evaluate several software programs for a selected test set representing the typical compounds that are to be predicted. By doing so, not only the accuracy of the software but also the user-friendliness of each program is included in the evaluation and the decision. Often the most predictive models are established in-house since these models are based on the chemical space of interest. However, the commercial software can be a good complement to such in-house models in terms of investigating interactions between different processes and allow visualization of the complete absorption process.

15.5 Future Outlook

Frontloading of assessing ADME properties early in the discovery process has gained much importance during recent years and is now considered a routine. These efforts have significantly reduced the ADME-related attrition in the clinical phase.

For the frontloading process in early drug discovery to have an impact, the ADME properties have to be assessed on a large number of compounds already in the early lead generation phase. This has led to the development of high-capacity *in vitro* assays to model different aspects of the *in vivo* situation; for example, permeability is being approximated with the Caco-2 assay. The information gained from such assays has played an important role in the design of molecules with good ADME properties.

To have a higher impact on the decision-making process, the next step has been the heavy use of prediction models in the design stage before the molecules are synthesized. Ideally, the prediction models are based on more complex measures, such as F_a. The largest limitation of this initiative is the availability of such data, which is not likely to increase much in the near future. However, *in vitro* permeability and solubility values are being routinely measured on thousands of compounds. This gives the opportunity to generate more elaborate models and also to fine-tune them using technologies such as correction libraries [58]. Combining solubility and permeability models (and possibly models for active transport) can give a good estimate of oral absorption for a large number of compounds.

For the use of these models in the drug discovery phase, one can identify two scenarios:

1. *The lead generation (or hit-to-lead) phase:* In this phase, it is desirable to obtain predictions of a large number of molecules before any experimental measurements are feasible from a practical point of view. The predictions do not need to be limited to the hits found in a high-throughput screen but can also comprise large virtual libraries of possible follow-up compounds. This requires very fast models for which the prediction of each molecule is done in a fraction of a second. A necessary requirement for future models is therefore speed, which has to be combined with a quality that is acceptable. Another future requirement is the generation of good data analysis programs, which can handle and judge the impact of each prediction and make intelligent selection of which compounds will be most successful. In this step, it is most likely that other absorption, distribution, metabolism, elimination/excretion, and toxicity (ADMET) components will be included such as predictions of transporter interactions, distribution, enzymatic degradation, and toxicity.

2. *The lead optimization phase:* For the lead optimization phase, one can afford slightly more elaborate (and also probably slower) models if they have a significant

increase in predictability. However, calculation times above minutes for each molecule are still not desirable. Experimental data are more generally available in this phase of the drug discovery process and the compounds of interest can normally be assigned to one or several series. Therefore, local prediction models can be advantageous to use.

An increased use of models predicting oral absorption will not only reduce the ADME-related attrition in the clinic but also increase the speed of the discovery process. Even though the models can give a good estimate of the ADME properties of molecules, it is most likely that the *in silico* models of the future will be used in concert with *in vitro* and *in vivo* models to predict the complex ADME profile of compounds that have advanced to a later phase in the drug discovery process.

References

1 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Drug Delivery Reviews*, **23**, 3–25.

2 Palm, K., Stenberg, P., Luthman, K. and Artursson, P. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research*, **14**, 568–571.

3 Amidon, G.L., Lennernäs, H., Shah, V.P. and Crison, J.R. (1995) A theoretical basis for a biopharmaceutic drug classification: the correlation of *in vitro* drug product dissolution and *in vivo* bioavailability. *Pharmaceutical Research*, **12**, 413–420.

4 Hansch, C., Quinlan, J.E. and Lawrence, G.L. (1968) Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *The Journal of Organic Chemistry*, **33**, 347–350.

5 Yalkowsky, S.H. and Valvani, S.C. (1980) Solubility and partitioning I: Solubility of nonelectrolytes in water. *Journal of Pharmaceutical Sciences*, **69**, 912–922.

6 Jain, N. and Yalkowsky, S.H. (2001) Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *Journal of Pharmaceutical Sciences*, **90**, 234–252.

7 Tetko, I.V. and Bruneau, P. (2004) Application of AlogPS to predict 1-octanol/water distribution coefficients, log *P* and log *D*, of AstraZeneca in-house database. *Journal of Pharmaceutical Sciences*, **93**, 3103–3110.

8 Tetko, I.V. and Poda, G.I. (2004) Application of ALOGPS 2.1 to predict log *D* distribution coefficient for Pfizer propriety compounds. *Journal of Medicinal Chemistry*, **47**, 5601–5604.

9 Bergström, C.A.S., Norinder, U., Luthman, K. and Artursson, P. (2003) Molecular descriptors influencing melting point and their role in classification of solid drugs. *Journal of Chemical Information and Computer Sciences*, **43**, 1177–1185.

10 Karthikeyan, M., Glen, R.C. and Bender, A. (2005) General melting point prediction based on a diverse compound data set and artificial neural networks. *Journal of Chemical Information and Modeling*, **45**, 581–590.

11 Huuskonen, J., Salo, M. and Taskinen, J. (1997) Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *Journal of Pharmaceutical Sciences*, **86**, 450–454.

12 Bruneau, P. (2001) Search for predictive generic model of aqueous solubility using Bayesian neural nets. *Journal of Chemical*

Information and Computer Sciences, **41**, 1605–1616.

13 Liu, R.F. and So, S.S. (2001) Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *Journal of Chemical Information and Computer Sciences*, **41**, 1633–1639.

14 Livingstone, D.J., Ford, M.G., Huuskonen, J.J. and Salt, D.W. (2001) Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *Journal of Computer-Aided Molecular Design*, **15**, 741–752.

15 Bergström, C.A.S., Wassvik, C.M., Norinder, U., Luthman, K. and Artursson, P. (2004) Global and local computational models for aqueous solubility prediction of drug-like molecules. *Journal of Chemical Information and Computer Sciences*, **44**, 1477–1488.

16 Wassvik, C.M., Holmen, A.G., Bergström, C.A.S., Zamora, I. and Artursson, P. (2006) Contribution of solid-state properties to the aqueous solubility of drugs. *European Journal of Pharmaceutical Sciences*, **29**, 294–305.

17 Bergström, C.A.S., Wassvik, C.M., Johansson, K. and Hubatsch, I. (2007) Poorly soluble marketed drugs display solvation limited solubility. *Journal of Medicinal Chemistry*, **50**, 5858–5862.

18 Camenisch, G., Folkers, G. and van de Waterbeemd, H. (1996) Review of theoretical passive drug absorption models: historical background, recent developments and limitations. *Pharmaceutica Acta Helvetiae*, **71**, 309–327.

19 Abraham, M.H., Chadha, H.S. and Mitchell, R.C. (1995) The factors that influence skin penetration of solutes. *The Journal of Pharmacy and Pharmacology*, **47**, 8–16.

20 Palm, K., Luthman, K., Ungell, A.L., Strandlund, G. and Artursson, P. (1996) Correlation of drug absorption with molecular surface properties. *Journal of Pharmaceutical Sciences*, **85**, 32–39.

21 Norinder, U., Osterberg, T. and Artursson, P. (1997) Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parameterisation and PLS statistics. *Pharmaceutical Research*, **14**, 1786–1791.

22 Stenberg, P., Luthman, K., Ellens, H., Lee, C.P., Smith, P.L., Lago, A. and Elliott, J.D. (1999) Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharmaceutical Research*, **16**, 1520–1526.

23 van de Waterbeemd, H. (1998) The fundamental variables of the biopharmaceutics classification system (BCS): a commentary. *European Journal of Pharmaceutical Sciences*, **7**, 1–3.

24 Artursson, P. (1990) Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. *Journal of Pharmaceutical Sciences*, **79**, 476–482.

25 Wessel, M.D., Jurs, P.C., Tolan, J.W. and Muskal, S.M. (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences*, **38**, 726–735.

26 Zhao, Y.H., Le, J., Abraham, M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Butina, D., Beck, G., Sherborne, B., Cooper, I., Platts, J.A. and Boutina, D. (2001) Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences*, **90**, 749–784.

27 Votano, J.R., Parham, M., Hall, L.H. and Kier, L.B. (2004) New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Molecular Diversity*, **8**, 379–391.

28 Yang, C.Y., Dantzig, A.H. and Pidgeon, C. (1999) Intestinal peptide transport systems

and oral drug availability. *Pharmaceutical Research*, **16**, 1331–1343.

29 Matsson, P., Bergström, C.A.S., Nagahara, N., Tavelin, S., Norinder, U. and Artursson, P. (2005) Exploring the role of different drug transport routes in permeability screening. *Journal of Medicinal Chemistry*, **48**, 604–613.

30 Zmuidinavicius, D., Didziapetris, R., Japertas, P., Avdeef, A. and Petrauskas, A. (2003) Classification structure–activity relations (C-SAR) in prediction of human intestinal absorption. *Journal of Pharmaceutical Sciences*, **92**, 621–633.

31 Klopman, G., Stefan, L.R. and Saiakhov, R.D. (2002) ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences*, **17**, 253–263.

32 Physicians' Desk Reference (2003) 57th edn, Thomson Healthcare, USA.

33 Dollery, C. (1999) Therapeutic Drugs, 2nd edn, Churchill Livingstone, UK.

34 Raevsky, O.A., Fetisov, V.I., Trepalina, E.P., McFarland, J.W. and Schaper, K.-J. (2000) Quantitative estimation of drug absorption in humans for passively transported compounds on the basis of their physicochemical parameters. *Quantitative Structure–Activity Relationships*, **19**, 366–374.

35 Niwa, Y. (2003) Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *Journal of Chemical Information and Computer Sciences*, **43**, 113–119.

36 Physicians' Desk Reference (1994) 48th edn, Medical Economics Data Production Company, USA.

37 Physicians' Desk Reference (1997) 51th edn, Medical Economics Data Production Company, USA.

38 Pérez, M.A.C., Sanz, M.B., Torres, L.R., Ávalos, R.G., Pérez González, M. and Díaz, H.G. (2004) A topological sub-structural approach for predicting human intestinal absorption of drugs. *European Journal of Medicinal Chemistry*, **39**, 905–916.

39 Benet, L.Z., Øie, S. and Schwartz, J.B. (1996) Appendix II. Design and optimization of dosage regimens; pharmacokinetic data, in *Goodman and Gilman's The Pharmacological Basis of Therapeutics* (eds J.G. Hardman, L.E. Limbird, P.B. Molinoff and R.W. Ruddon), McGraw-Hill, New York, pp. 1712–1792.

40 Estrada, E. (1997) Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *Journal of Chemical Information and Computer Sciences*, **37**, 320–328.

41 Gunturi, S.B. and Narayanan, R. (2007) *In silico* ADME modeling. 3. Computational models to predict human intestinal absorption using sphere exclusion and kNN QSAR methods. *QSAR & Combinatorial Science*, **26**, 653–668.

42 Sun, H. (2004) A universal molecular descriptor system for prediction of log *P*, log *S*, log BB, and absorption. *Journal of Chemical Information and Computer Sciences*, **44**, 748–757.

43 Wold, S., Johansson, E. and Cocchi, M. (1993) PLS: Partial least-squares projections to latent structures, in *3D QSAR in Drug Design* (ed. H. Kubinyi), ESCOM Science Publishers B.V., The Netherlands, pp. 523–550.

44 Bai, J.P.F., Utis, A., Crippen, G., He, H.-D., Fischer, V., Tullman, R., Yin, H.-Q., Hsu, C.-P., Jiang, L. and Hwang, K.-K. (2004) Use of classification regression tree in predicting oral absorption in humans. *Journal of Chemical Information and Computer Sciences*, **44**, 2061–2069.

45 ZyxBio LLC, Hudson.

46 Deconinck, E., Hancock, T., Coomans, D., Massart, D.L. and Vander Heyden, Y. (2005) Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *Journal of Pharmaceutical and Biomedical Analysis*, **39**, 91–103.

47 Todeschini, R., Consonni, V., Mauri, A. and Pavan, M.Dragon version 4.0, Talete srl.

48 http://www.hyper.com.

49 Klon, A.E., Lowrie, J.F. and Diller, D.J. (2006) Improved naïve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *Journal of Chemical Information and Modeling*, **46**, 1945–1956.

50 MOE Molecular Operating Environment (2005 –2006) Chemical Computing Group, Inc., http://www.chemcomp.com.

51 Pipeline Pilot, version 5.1, SciTegic, Inc, http://www.scitegic.com.

52 Cheng, A., Diller, D.J., Dixon, S.L., Egan, W.J., Lauri, G. and Mertz, K.M., Jr (2002) Computation of the physio-chemical properties and data mining of large molecular collections. *Journal of Computational Chemistry*, **23**, 172–183.

53 Hou, T., Wang, J. and Li, Y. (2007) ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *Journal of Chemical Information and Modeling*, **47**, 2408–2415.

54 Johansson, U., Sönströd, C., König, R. and Niklasson, L. (2003) Neural networks and rule extraction for prediction and explanation in the marketing domain, in *The International Joint Conference on Neural Networks*, IEEE Press, Portland, OR, pp. 2866–2871.

55 Johansson, U. (2007) Obtaining accurate and comprehensible data mining models, PhD Thesis, Institute of Technology, Linköping University, (http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-8881).

56 Palm, K., Luthman, K., Ungell, A.-L., Strandlund, G., Beigi, F., Lundahl, P. and Artursson, P. (1998) Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *Journal of Medicinal Chemistry*, **41**, 5382–5392.

57 Parrott, N. and Lave, T. (2002) Prediction of intestinal absorption: comparative assessment of GASTROPLUS and IDEA. *European Journal of Pharmaceutical Sciences*, **17**, 51–61.

58 Rodgers, S.L., Davis, A.M., Tomkinson, N.P. and van de Waterbeemd, H. (2007) QSAR modeling using automatically updating correction libraries: application to a human plasma protein binding model. *Journal of Chemical Information and Modeling*, **47**, 2401–2407.

16
In Silico Prediction of Human Bioavailability

David J. Livingstone and Han van de Waterbeemd

Abbreviations

ADME	Absorption, distribution, metabolism, and excretion
AFP	Adaptive fuzzy partitioning
ANN	Artificial neural network
CoMFA	Comparative molecular field analysis
MLR	Multiple linear regression
NCE	New chemical entity
PBPK	Physiologically-based pharmacokinetics
PK	Pharmacokinetics
QSAR	Quantitative structure–activity relationship
R&D	Research and development
SIMCA	Soft independent modeling of class analogy

Symbols

$A\%$	Percentage absorbed
AUC	Area under the curve
C_0	Concentration at time zero
Caco-2	Human colon adenocarcinoma cell line (used as absorption model)
CL	Clearance
CL_u	Unbound clearance
Dose	Administered dose
F	Bioavailability (expressed as fraction)
$F\%$	Percentage bioavailable
f_a	Fraction absorbed
f_g	Fraction escaping gut wall intestinal metabolism
f_u	Fraction unbound to plasma proteins
$\log D$	Logarithm of the distribution coefficient D (for ionized species)

$\log P$	Logarithm of the partition coefficient P
$t_{1/2}$	Half-life
V_d	Volume of distribution
V_{du}	Unbound volume of distribution

16.1 Introduction

To make it convenient to the patients and to increase compliance, most drugs are given orally. Therefore, high bioavailability is a key quest in most drug discovery projects. Low bioavailability usually results in undesired variability due to population differences. *In vitro* ADMET and safety profiling is now well established in drug discovery [1]. Often, oral bioavailability is assessed in the rat. However, this is not always predictive for bioavailability in human. Factors that influence oral drug bioavailability can be divided into physicochemical/biopharmaceutical and physiological/biological factors. The first group tends to be essentially species independent. Indeed, species differences in pH values change percentage ionization and hence molecular behavior. Biological factors are often different between species [2]. Early estimates of oral bioavailability can help to focus on most promising lead series and clinical candidates. To address bioavailability issues in a drug discovery project, a road map of experimentation and prediction has been proposed [3].

This chapter reviews some of the *in silico* attempts to predict oral bioavailability. However, bioavailability is a complex property, and various pros and cons of current quantitative structure–activity relationship (QSAR) based approaches will be discussed here. As an alternative, physiologically-based pharmacokinetic (PBPK) modeling is discussed as a promising approach to predict and simulate pharmacokinetics (PK), including estimating bioavailability.

In silico models of biological activity have been constructed in the discovery research departments of the pharmaceutical companies since the 1970s. Early adopters of the technologies built QSAR models of pharmacological response and even wrote molecular modeling software before integrated packages became commercially available. Expectations of the results that might be delivered by these approaches were high, partly as a result of the enthusiasm of the computational chemists and partly because of the acceptance of the ideas by medicinal chemists. The failure of these early attempts to deliver dramatic changes in the rate of discovery of new chemical entities (NCEs) led to disappointment and a decrease in popularity of these approaches over the next few years. Expectations have now reached a mature level, and computer-aided molecular design is an accepted part of the discovery process with the advantages and limitations of *in silico* modeling well understood [4].

The next major technological advance in drug discovery was the development of combinatorial chemistry [5] and high-throughput screening [6], which increased the number of compounds synthesized and tested by factors of hundreds or even thousands. A decade after the widespread adoption of this approach, however, has seen little or no increase in the number of NCEs submitted to the regulatory authorities, despite the ever-increasing expenditure on pharmaceutical R&D. There has been much debate about the reasons for this apparent failure, but a general consensus of opinion at that time was that the primary cause is poor ADME (absorption, distribution, metabolism, and excretion) properties. Indeed, widely quoted reports conclude that as much as 40% of the failures of NCEs can be attributed to this cause [7–9]. In fact, these figures are now old data and the truer situation is probably as little as 10–15% [10, 11] but, nevertheless, there is still a considerable effort focussed on the optimization of ADME properties, and high-throughput methods are now being applied to generate ADME and toxicity information [12–14].

Various attrition analysis studies appeared in the literature, the one by Pfizer's Chris Lipinski being one of the most cited. He investigated a range of easily computable properties of compounds in the World Drug Index (WDI). Since these compounds are either marketed drugs or are currently in clinical trial, the properties common to these compounds define what we call now "drug-like" properties [86]. It was found that compounds tend to have poor oral absorption if their molecular weight (MW) is more than 500, calculated partition coefficient (Clog *P*) is more than 5, the number of hydrogen-bond acceptors (HBAs) is more than 10, and number of hydrogen-bond donors (HBDs) is more than 5 [15]. This was called the rule-of-5, since all key numbers are a multiple of 5. Many drugs are ionized at physiological pH values (pH 5.5–7.4) and to reflect this, a proposal was made to use log *D* instead of log *P* in drug-like filters [16]. Unfortunately, quite often the rule-of-5 is wrongly linked to bioavailability [17, 18]. Poor bioavailability can occur even for compounds with excellent oral absorption, if they have high first-pass liver clearance (CL). In Figure 16.1, the difference between oral absorption and bioavailability is

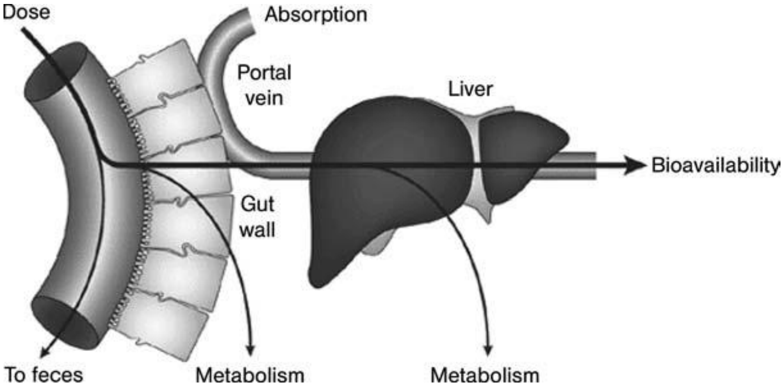

Figure 16.1 Definition of oral absorption (percentage of dose reaching the portal vein) and bioavailability (percentage of dose reaching the systemic circulation [50].

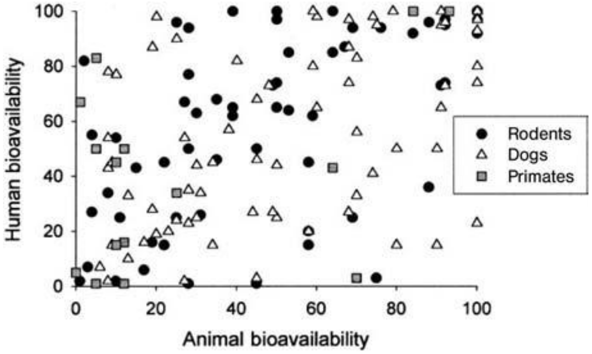

Figure 16.2 Plot of the absolute human bioavailability of various drugs versus their absolute bioavailability in primates, dogs, and rodents [19].

schematically presented. The key difference is metabolism (or clearance) and, for some compounds, the interaction with transporters.

Human bioavailability is often estimated by measuring bioavailability *in vivo* in several animal species and assuming humans are similar. Unfortunately, the direct use of measured animal bioavailability data is unlikely to be a good model of the human situation as shown in Figure 16.2, where data are plotted from Sietsema [19]. At first sight, it might appear that these experimental bioavailability measurements in animals are completely irrelevant for the prediction of human bioavailability, but the situation is not as bad as it first appears since what is plotted here are absolute bioavailability measurements. The lack of correlation between animal and human data is likely to be due to differences in physiology between the species, that is, to say differences in absorption, metabolism, plasma protein binding, and so on [20]. The use of absolute bioavailability data from animals to model the human data is equivalent to attempting to correlate all of the independent processes in the animal and simultaneously relate them to the corresponding processes in the humans [21]. This lack of direct correlation simply highlights the fact that bioavailability is a complex property as discussed in the next section. It should be noted that oral absorption compares often, but not always, much better between species [20, 51].

It should also be borne in mind that a complex biological property such as bioavailability is influenced by many factors as discussed in this chapter. The result is the considerable interindividual variability of about 15% standard deviation in observed bioavailability. The consequence of this is that any modeling approach cannot be better than this (Figure 16.3).

Another approach to human bioavailability estimation is based on *in vitro* data using Caco-2 as a measure of permeability and human liver microsomes for metabolism estimates. These data are combined in a graphical method to get a rough estimate of human oral bioavailability [22]. In principle, but not yet proven, this method could also be applied by using calculated permeability and metabolic stability.

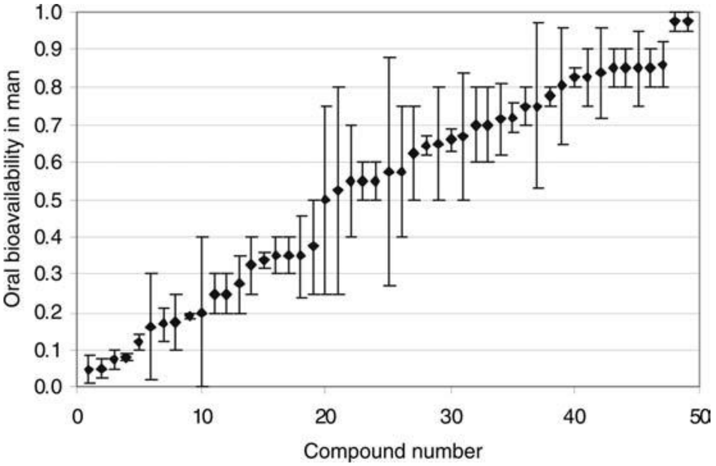

Figure 16.3 Interindividual variability in bioavailability [32].

16.2 Concepts of Pharmacokinetics and Role of Oral Bioavailability

Most drugs are given orally for reasons of convenience and compliance. Typically, a drug dissolves in the gastrointestinal tract, is absorbed through the gut wall, and then passes the liver to get in the blood circulation. The percentage of the dose reaching the systemic circulation is called bioavailability. From there, the drug will get distributed to various tissues and organs in the body. The extent of distribution will depend on the compound's structural and physicochemical properties. Some drugs may enter the brain and central nervous system (CNS) via the blood–brain barrier (BBB). Finally, the drug will bind to its molecular target, for example, a receptor or ion channel, and exert its desired action. A short summary of the key pharmacokinetic parameters is given here [23].

The volume of distribution (V_d) is a theoretical concept that connects the administered dose with the actual initial concentration (C_0) present in the circulation:

$$V_\mathrm{d} = \frac{\mathrm{Dose}}{C_0}. \tag{16.1}$$

Most drugs will bind to various tissues and in particular to proteins such as albumin in the blood. Since only the free (unbound) drug will bind to the molecular target, the concept of unbound volume of distribution (V_du) is used:

$$V_\mathrm{du} = \frac{V_\mathrm{d}}{f_\mathrm{u}}, \tag{16.2}$$

where f_u is the fraction unbound to plasma proteins. Clearance of the drug from the body mainly takes place via the liver (hepatic clearance or metabolism and biliary excretion) and the kidney (renal excretion). By plotting the plasma concentration

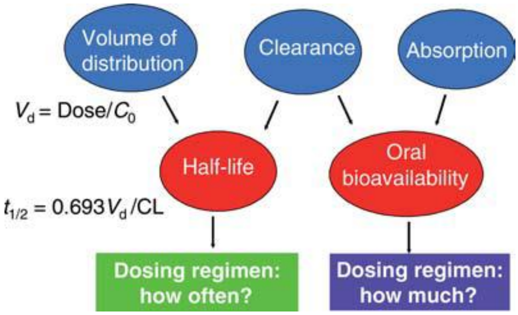

Figure 16.4 The key pharmacokinetic properties and their role in setting dose size and dose regimen [50].

against time, the area under the curve (AUC) relates to dose, bioavailability, and clearance [24]:

$$\mathrm{AUC} = \frac{F \times \mathrm{Dose}}{\mathrm{CL}}. \tag{16.3}$$

The required dose can be estimated from the potency (e.g., IC_{50}) of the compound and the unbound clearance (CL_u):

$$\text{Dose} = \text{therapeutic concentration} \times \text{dose interval} \times \text{oral unbound clearance}. \tag{16.4}$$

The daily dose size is determined by the free (unbound) concentration of drug required for efficacy and not by plasma protein or tissue binding. Protein or tissue binding is important in the actual dosage regimen or frequency per day. The greater the binding the lower and more sustained the free drug concentrations are [23].

Half-life ($t_{1/2}$), the time taken for a drug concentration in the plasma to reduce by 50%, is a function of the clearance and volume of distribution and reflects how often a drug needs to be administered as shown in Figure 16.4:

$$t_{1/2} = \frac{0.693\,V_d}{\mathrm{CL}}. \tag{16.5}$$

16.3 *In Silico* QSAR Models of Oral Bioavailability

16.3.1 Prediction of Human Bioavailability

Quantitative structure–activity relationships have been used since the 1960s to model receptor and enzyme affinity, as well as physicochemical properties. The renewed interest in QSAR [25] arises from the recognition that an early prediction of ADMET properties ensures compound quality and avoids early development failure related to

ADME and safety/toxicity issues. Predictive models are therefore widely used in library design and profiling.

One of the earliest *in silico* models of human bioavailability was reported by Hirono and coworkers [26]. This study employed a set of 188 compounds that were classified as low (<50%), medium (50–89%), or well (>90%) absorbed and used a classification routine, fuzzy adaptive least squares, to generate discriminant functions. The molecules were described by their physicochemical properties and substructural descriptors, which meant that functional groups or substructures that enhanced bioavailability (e.g., saturated carbon atoms in side chains) or reduced it (e.g., aliphatic hydroxyl groups) could be identified. The performance varied between the three classes with the lowest success for the well-absorbed compounds, perhaps the most important group of the three.

Yoshida and Topliss published another classification study in 2000 [27]. This used a larger set of compounds ($n = 232$), classified into four classes, described by 15 substructural descriptors expected to be related to metabolism. The authors used in their work on bioavailability prediction a descriptor $\Delta \log D$ defined as the difference between the distribution coefficient at pH 6.5 (taken as pH of the small intestine) and at pH 7.4 (blood) for an ionizable species to reflect drug transport. The efficiency of this *in silico* system on a test set of 40 compounds was around 60%.

A model of bioavailability using a continuous measure was generated through stepwise regression and recursive partitioning to optimize the regression equations [28]. This study employed 591 compounds that were characterized by a large set ($\sim$600) of simple chemical substructure descriptors. Model efficiency was quite poor since the average R^2 value from 2000 random splits of the data into 80/20% training and test sets was 0.58. The model was also judged by comparison with predictions from the "rule-of-5" [15] (although this rule refers to oral absorption) and was shown to give a slight improvement over the false negative, 3 versus 5%, and false positive, 46 versus 53%, predictions. Some insights into the difficulty of modeling bioavailability may be gained by the complexity of the regression model, which involved 85 terms. Another linear regression approach using 169 compounds led to a regression model containing eight terms [29]. This study involved more complex descriptors, including some calculated by quantum mechanics, and gave a slight improvement in fit compared to the model reported by Andrews and coworkers.

These later two models of bioavailability as a continuous variable are linear since they used stepwise multiple linear regression (MLR) as the modeling tool. An obvious alternative, which may offer improved performance, is a nonlinear technique and such a model using an artificial neural network (ANN) was reported by Turner and colleagues [30]. This study employed 167 compounds characterized by several descriptor types, 1D, 2D, and 3D, and resulted in a 10-term model. Although the predictive performance was judged adequate, it was felt that the model was better able to differentiate qualitatively between poorly and highly bioavailable compounds.

Given the relatively poor performance of quantitative models, it is not surprising that other attempts to build *in silico* models of human bioavailability have concentrated on classification. Adaptive fuzzy partitioning (AFP) was applied for two sets of bioavailability data subdivided into four ranges of activity [31]. The best models using the Yoshida and Topliss data [27] were able to predict correctly 75% of the validation set compounds. It was also shown that the predictive power increases when including more chemical diversity in the training set.

A genetic programming algorithm has been used to build models based on an automatic generation of substructural descriptors [32]. These models performed as good as other models based on classified data, so given the variety of descriptors tried and modeling techniques employed, this perhaps indicates that the problem in modeling human bioavailability lies in the data. This is not to say that there is anything wrong with the data but that it represents a summation of many different processes as discussed further in the next section.

Another approach is based on the combination of molecular interaction fields using the 3D-QSAR technique CoMFA and soft independent modeling of class analogy (SIMCA) [33]. Predictions were made for *F*% ranges by using the data sets from Refs [19, 27], with about 60% correctly classified.

Hou *et al*. compiled a database of human bioavailability for 768 compounds, which is publicly available [34]. These authors used a cutoff of 20% as acceptable. This can be questioned as *F*% up to about 30–40% can show considerable interindividual variability. It was concluded that *F*% of highly metabolized compounds cannot be well predicted from simple molecular descriptors as these do not encode for metabolism.

Martin proposed a "bioavailability score" based on several molecular properties including polar surface area (PSA), rule-of-5, and molecular charged state. With the descriptors used, this is an example aiming to estimate oral absorption and not bioavailability [19]; hence, the title of this work is misleading. A score was developed to assign the probability that a compound has an *F* more than 10% in the rat. We do not consider this as a meaningful cutoff. Better would be *F* more than 30% in man [30].

A cascade method was proposed using recursive partitioning and descriptors generated with a program called Algorithm Builder were used with a 800-compound training set [35]. Their predictions are 2-class models with *F*s less than and more than 30%, respectively. As parameters, they use a combination of solubility, pK_a, fractions ionized, human permeability, P-gp substrate specificity, physicochemical properties, and various structural descriptors. This is an attempt to model the components of bioavailability and then to integrate them into an overall prediction (see further in Section 16.4).

Using a data set of 577 compounds with experimental human bioavailability, a set of 42 bioavailability-boosting fragments was derived [36], although the general validity of these can be questioned without further proof of concept. These fragments were combined with other descriptors and with a genetic algorithm (GA), a set of 20 models for *F*% was obtained, and the final prediction was based on a consensus score ($r^2 = 0.55$, RMSE = 21.9%). In addition, an HQSAR (hologram QSAR) model (see also below) was derived from the same data set ($r^2 = 0.35$, RMSE = 26.4%). The combined consensus GA and HQSAR model works best ($r^2 = 0.62$, RMSE = 20.2%). This is a reasonable result in view of the fact that the standard error of the experimental data is 14.5%.

Molecular holograms are an extended form of fingerprints based on the 2D structures. An HQSAR model was derived for 250 compounds ($r^2 = 0.93$, $q^2 = 0.70$) and tested with 52 compounds ($r^2 = 0.85$) [37], which is a good result. The authors correctly point out some limitations of the model. Training is based on drugs, most compliant to the rule-of-5, and has no real solubility issues. The question therefore arises whether such model would be predictive for and pick out nondrug-like compounds.

A recent review gives a comprehensive survey of the state of the art in modeling human bioavailability [38]. Commercial software for the prediction of bioavailability using QSAR approaches include ADME Boxes [www.ap-algorithms.com], truPK [trupk.strandgenomics.com], and KnowItAll [www.knowittallcom].

16.3.2
Prediction of Animal Bioavailability

These models have all attempted to explain human bioavailability, which of course is our primary interest in drug design. Much data, however, have been measured in animals, particularly in the rat. Veber and colleagues at GSK have studied a set of 1100 compounds for which oral bioavailability in the rat was measured in-house [39]. The most important properties favorable for high oral bioavailability (in the rat) appear to be reduced molecular flexibility as measured by the number of rotatable bonds and low polar surface area. They conclude that these properties are in fact independent of the molecular weight. This would contradict the MW less than 500 rule as proposed by Lipinski in developing his rule-of-5 [15]. However, it was also found by others at Pharmacia that these results could not be generalized [40]. These authors reminded that property calculations can be algorithm dependent and that conclusions can be drug-class dependent; therefore, generalizations must be used with caution.

16.4
Prediction of the Components of Bioavailability

Oral *bioavailability* is a complex property. Many of the contributing factors are known (see Figure 16.4), but their cooperation is not always fully clear. Modeling of each of the more fundamental properties contributing to oral bioavailability will give more mechanistic insight. This might help the medicinal chemist to fine-tune the properties of a compound, which lead to poor bioavailability, while keeping the others in the right ballpark [41]. Fortunately, *in silico* models have been developed for many of these individual processes and, as the pharmaceutical industry continues to concentrate on these problems, better and more meaningful experimental tests are being developed leading to larger amounts of more accurate and reliable data. Reasonably successful models have been developed for several of the components shown in Figure 16.5. Many chapters in this book detail these approaches to understand properties contributing to bioavailability.

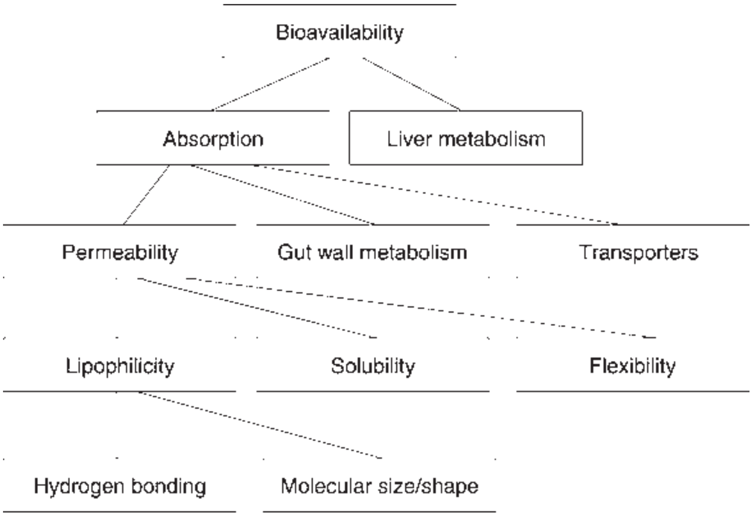

Figure 16.5 Bioavailability is a complex property, which can be unravelled into its more fundamental components [51].

Fundamental *physicochemical properties* (see Chapter 5) such as partition coefficient (log *P*) and distribution coefficient (log *D*) and pK_a are well predicted directly from chemical structure [42, 43]. Aqueous solubility may also be predicted reasonably well [44–46] (see Chapter 4), though there are warnings on the accuracy of solubility predictions for in-house pharmaceutical company compounds since these tend to be quite different chemical structures from those used to develop the commercial models [47].

Permeability is perhaps the most widely studied of the "biological" components of Figure 16.5 and as a result has led to a number of *in silico* models of this component [48]. There are various experimental systems designed to give some measure of permeability ranging in complexity from partition into liposomes to permeability across Caco-2 cells (see Chapter 7). *In silico* models of Caco-2 cell permeability have been constructed [49], but the question may be asked: "why model the model of human absorption?" [50]. It may be better to measure, and make models of, more fundamental factors that affect permeability.

There is considerable literature available on the prediction of oral *absorption* [51, 34] (see Chapter 15). One of the key problems is the lack of sufficient data to build robust predictive models.

Other components of bioavailability are also studied experimentally and in some cases *in silico* models have been developed for them. Examples include plasma protein binding [52], P-glycoprotein [53] (see Chapter 18) and other transporters [54] (see Chapter 10), and metabolism by cytochrome P450s [55, 56] (see Chapter 12).

Metabolism forms an important and probably least well-modeled part of the overall ADME process [57]. While metabolism/clearance may have a significant effect on oral bioavailability, it is clearly ultimately responsible for the fate of xenobiotics, except those few that are excreted unchanged. To be able to produce reliable *in silico* models of the ADME properties of drugs, it will be necessary to understand and model the complex processes involved in metabolism. There is need to answer such questions as which enzyme is involved [58], what extent and rate, regioselectivity, or site of metabolism [59, 60], which metabolites are being formed [54, 55], are some of these reactive metabolites (which can cause toxic effects)? There are two main approaches to this problem: expert systems and computer-aided design. The expert system approach consists of a set of rules based on individual experience of how compounds are "dissected" by metabolism. The rules may be supplemented by physicochemical property calculations to apply some simple QSAR predictions. Examples of programs based on expert systems are MetabolExpert [61] and METEOR [62, 63]. The computer-aided design methods may be based on the properties of the ligands (QSAR) [64, 65] or on the structure of the enzyme (molecular modeling) [66–68].

This brief discussion of the components of bioavailability has shown that many of them are accessible experimentally and that *in silico* models may be built for the majority of them. Some recent reviews discuss the modeling of these components and other ADME processes [69–72]. It is by no means clear, however, and therefore remains a challenge as to how the individual components can easily be integrated into an overall prediction of oral bioavailability.

16.5 Using Physiological Modeling to Predict Oral Bioavailability

The concepts of pharmacokinetics come from classical compartmental modeling as described in Section 16.2. These compartments are not "real" compartments in any physical sense but rather virtual compartments that are required to make the modeling process work. The components of bioavailability and other parts of the whole ADME process are quite well understood, amenable to experimental measurements, and capable of *in silico* modeling as discussed in Section 16.3. The question remains, therefore, whether is it possible to link these pharmacokinetic concepts to the individual components and processes occurring in the body? The simple answer to this question is "yes, in principle," but the way to do this is by no means obvious. There is one approach, known as physiologically-based pharmacokinetic modeling, which is promising and increasingly used [73, 74]. PBPK modeling attempts to produce models that describe a system in physiological terms, in other words, the actual organs, blood flow, partition processes, and so on [75–77]. The information required for PBPK modeling is both chemical and biological as shown in Table 16.1.

This form of modeling is intellectually appealing since it is based on physiology, and thus there is a good scientific rationale to the process. Since it is based on physiology, it is possible to draw mechanistic conclusions and make quantitative predictions of disposition in various tissues. As a result, it being routinely applied in chemical risk assessment and can be judged as a standard methodology in this

Table 16.1 Information needs for PBPK models (adapted from Ref. [57]).

Chemical-specific data	Biological data
Partition coefficients	Anatomical dimensions
Metabolic rate constants	Organ blood flows
Elimination rate constants	Organ volumes
Molecular weight	Cardiac output
Aqueous solubility	Ventilation rate
Vapor pressure	Body mass
Permeability coefficients	Level of physical activity
Diffusion coefficients	Age
Protein-binding constants	Gender

field [78]. In drug research, most applications are performed in drug development where sufficient data are available to feed into the models. The ultimate bioavailability of a new drug considerably depends on formulation. The biopharmaceutical assessment is therefore an important part of the preclinical development program. The Biopharmaceutical Classification System (BCS) is one such tool used (see Chapter 19). Hurdles and critical parameters for oral bioavailability can be studied by using computer simulations, for example, GastroPlus [www.simulations-plus.com] [79] (see Chapter 17).

Unfortunately, the application of PBPK modeling in drug discovery so far has been very limited, and only a few studies have been reported [80, 81]. This is almost certainly due to the high data requirements, both chemical and biological, to produce these models. This sort of data is not normally collected as part of the regular drug discovery process and thus an investment in resources is required to produce it. Fortunately, some of the parameters required for PBPK modeling, for example, solubility, partition coefficients, uptake, and so on, can be estimated from *in silico* models, and so it looks set to become a more routine part of drug research [82].

Another way in which the components of bioavailability can be used, without the complexity of a formal PBPK model, is to link the components in an empirical but logical fashion. Two programs that demonstrated this, iDEA and pkEXPRESS, had modules to estimate oral bioavailability from Caco-2 and microsomal metabolic stability [83, 84]. Unfortunately, both products are no longer commercially available.

The principle for the bioavailability estimate is as follows [84]. Hepatic intrinsic clearance ($\mathrm{CL_{int}}$) is measured in hepatocytes by measuring the half-life ($t_{1/2}$) in the following equation:

$$\mathrm{CL_{int}} = \left(\frac{0.693}{t_{1/2}}\right) \times \left(\frac{\text{g liver}}{\text{kg body}}\right) \times \left(\frac{\text{ml incubation}}{\text{cells incubation}}\right) \times \left(\frac{\text{cells}}{\text{g liver}}\right). \qquad (16.6)$$

The blood clearance can then be obtained through one of the several approaches, for example, the well-stirred model and fraction unbound ($f_{\mathrm{u},}$ obtained from plasma protein binding) and the liver blood flow (Q, which for humans can be taken as 25 ml/min/kg [23]):

$$\mathrm{CL_b} = \frac{Q \times \mathrm{CL_{int}} \times f_\mathrm{u}}{Q + \mathrm{CL_{int}} \times f_\mathrm{u}}. \tag{16.7}$$

The hepatic extraction rate (E_H) is

$$E_H = \frac{\mathrm{CL_b}}{Q}. \tag{16.8}$$

The bioavailability (F), correcting for the fraction absorbed (f_a) and fraction escaping intestinal metabolism (f_g), is

$$F = f_\mathrm{a} \times f_\mathrm{g} \times (1 - E_\mathrm{H}). \tag{16.9}$$

The fraction absorbed (f_a) is obtained from other measurements including solubility and permeability by using, for example, Caco-2 or PAMPA data [83], each of which in principle can also be predicted with a QSAR model (but see comment in 16.4).

More recent PBPK software packages such as SIMCYP [www.simcyp.com], PK-Sim [www.pk-sim.com], GastroPlus [www.simulations-plus.com], and Cloe PK [www.cyprotex.com] offer similar bioavailability estimation. It is clear that these approaches require more data input than just molecular structure as in QSAR models.

Integration of *in vitro* results and pharmacokinetic modeling is also used to assess the bioavailability of nutrients [85] using TNO's gastrointestinal model TIM [www.tno.nl/pharma].

16.6 Conclusions

The properties that are important for drug metabolism and pharmacokinetics (DMPK) are much better understood now than they were some 10 years ago [24]. Good progress has been made in recent years toward robust modeling of a number of pharmacokinetic properties and various aspects of human drug metabolism. More and good quality data have become available for some of the important end points. However, and unfortunately, some end points are by nature very complex. These include clearance and oral bioavailability.

It is possible to build *in silico* models of human bioavailability but while these may work well for certain classes of drugs, possibly because their bioavailability is dominated by one process such as uptake, it is unlikely that they will work well for all drugs. This may be improved by increasing quantities of data, but taking into consideration more classes of drugs may have just the opposite effect. The cause of these problems is clear since a small number of fundamental physicochemical properties determine many of the components of bioavailability. Changes in these physicochemical properties may have quite different effects on individual components as we change from one drug class to another or, indeed, within a single class.

There is a choice of building in-house predictive models starting from literature and in-house bioavailability data. A wide range of QSAR tools are commercially or freely available. Alternatively, commercial packages can be used as discussed in this chapter.

Bioavailability is influenced by many properties, depending on the rate and the extent of absorption and systemic clearance. Each of these properties is impacted by physicochemical properties such as solubility, log *P*, log *D*, and pK_a. Absorption and metabolism are often governed by opposing factors. More lipophilic compounds tend to be more permeable, but solubility may also become a limiting factor. In addition, more lipophilic compounds will be more rapidly and extensively metabolized and will show increased toxicity liabilities [86].

Although formulation variables such as particle size and excipients have not been discussed here, they are highly relevant in practice. In addition, food can play an important role in oral absorption and thus bioavailability. Food may increase blood flow and thus limit the extent of first-pass effect. Bile secretion increases with food intake, which may enhance the solubility of lipophilic compounds. Attempts have been made to predict the effect of food on the extent of drug absorption [87]. Gastric emptying time is another factor, which depends on the type and the amount of food intake and physiopathology, among others.

Another approach to increase bioavailability via better absorption is using prodrugs (see Chapter 20).

It is therefore good to stress that bioavailability predictions can only be ballpark predictions in very early discovery stages, and they get better in later developments as *in silico* data can be mixed with *in vitro* and *in vivo* measurements [1].

References

1 Wang, J., Urban, L. and Bojanic, D. (2007) Maximising use of *in vitro* ADMET tools to predict *in vivo* bioavailability and safety. *Expert Opinion on Drug Metabolism and Toxicology*, **3**, 641–665.

2 Hurst, S., Loi, C.-M., Brodfuehrer, J. and El-Kattan, A. (2007) Impact of physiological, physicochemical and biopharmaceutical factors in absorption and metabolism mechanisms on the drug oral bioavailability of rats and humans. *Expert Opinion on Drug Metabolism and Toxicology*, **3**, 469–489.

3 Thomas, V.H., Bhattachar, S., Hitchingham, L., Zocharski, P., Naath, M., Surendran, N., Stoner, C.L. and El-Kattan, A. (2005) The road map to oral bioavailability: an industrial perspective. *Expert Opinion on Drug Metabolism and Toxicology*, **2**, 591–608.

4 Livingstone, D.J. and van de Waterbeemd, H. (2006) *In silico* models for human bioavailability, in *Virtual ADMET Assessment in Target Selection and Maturation* (eds B. Testa and L. Turski), IOS Press, Amsterdam, pp. 151–161.

5 Beck-Sickinger, A. and Weber, P. (2002) *Combinatorial Strategies in Biology and Chemistry*, John Wiley & Sons, Chichester, UK.

6 Dixon, G.K., Major, J.S. and Rice, M.J. (eds) (2000) *High Throughput Screening – The Next Generation*, Bios, Oxford, UK.

7 Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discovery Today*, **2**, 436–444.

8 Lipper, R.A. (1999) How can we optimize selection of drug development candidates from many compounds at the discovery stage? *Modern Drug Discovery*, **2**, 55–60.

9 Venkatesh, S. and Lipper, R.A. (2000) Role of the development scientist in compound lead selection and optimization. *Journal of Pharmaceutical Sciences*, **89**, 145–154.

10 Forum, General Metrics (2001).

11 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, **3**, 711–715.

12 Bertrand, M., Jackson, P. and Walther, B. (2000) Rapid assessment of drug metabolism in the drug discovery process. *European Journal of Pharmaceutical Sciences*, **11** (Suppl. 2), S61–S62.

13 Thompson, T.N. (2000) Early ADME in support of drug discovery: the role of metabolic stability studies. *Current Drug Metabolism*, **1**, 215–241.

14 Li, A.P. (2001) Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today*, **6**, 357–366.

15 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23**, 3–25.

16 Bhai, S.K., Kassam, K., Peirson, I.G. and Pearl, G.M. (2007) The rule of five revisited: applying log *D* in place of log *P* in drug-likeness filters. *Molecular Pharmacology*, **4**, 556–560.

17 Hou, T., Wang, J., Zhang, W. and Xu, X. (2007) ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *Journal of Chemical Information and Modeling*, **47**, 208–218.

18 Martin, Y.C. (2005) A bioavailability score. *Journal of Medicinal Chemistry*, **48**, 3164–3170.

19 Sietsema, W.K. (1989) The absolute oral bioavailability of selected drugs. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **27**, 179–211.

20 Mandagere, A.K. and Jones, B. (2003) Prediction of bioavailability, in *Drug Bioavailability* (eds H. van de Waterbeemd, H. Lennernäs and P. Artursson), Wiley-VCH Verlag GmbH, Weinheim, pp. 444–460.

21 Grass, G.M. and Sinko, P.J. (2002) Physiologically-based pharmacokinetic simulation modelling. *Advanced Drug Delivery Reviews*, **54**, 433–451.

22 Mandagere, A.K., Thompson, T.N. and Hwang, K.K. (2002) Graphical model for estimating oral bioavailability of drugs in humans and other species from their Caco-2 permeability and *in vitro* liver enzyme metabolic stability rates. *Journal of Medicinal Chemistry*, **45**, 304–311.

23 Smith, D.A., van de Waterbeemd, H. and Walker, D.K. (2006) *Pharmacokinetics and Metabolism in Drug Design*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim.

24 van de Waterbeemd, H., Smith, D.A., Beaumont, K. and Walker, D.K. (2001) Property-based design: optimization of drug absorption and pharmacokinetics. *Journal of Medicinal Chemistry*, **44**, 1313–1333.

25 van de Waterbeemd, H. and Rose, S. (2008) Quantitative approaches to quantitative structure–activity relationships, in *The Practice of Medicinal Chemistry*, 3rd edn (ed. C.G. Wermuth), Elsevier, Amsterdam.

26 Hirono, S., Nakagome, I., Hirano, H., Matsushita, Y., Yoshi, F. and Moriguchi, I. (1994) Non-congeneric structure–pharmacokinetic property correlation studies using fuzzy adaptive least-squares: oral bioavailability. *Biological & Pharmaceutical Bulletin*, **17**, 306–309.

27 Yoshida, F. and Topliss, J.G. (2000) QSAR model for drug human oral bioavailability. *Journal of Medicinal Chemistry*, **43**, 2575–2585.

28 Andrews, C.W., Bennett, L. and Yu, L.X. (2000) Predicting human oral bioavailability of a compound: development of a novel quantitative structure–bioavailability relationship. *Pharmaceutical Research*, **17**, 639–644.

29 Turner, J.V., Glass, B.D. and Agatonovic-Kustrin, S. (2003) Prediction of drug bioavailability based on molecular structure. *Analytica Chimica Acta*, **485**, 89–102.

30 Turner, J.V., Maddalena, D.J. and Agatonovic-Kustrin, S. (2004) Bioavailability prediction based on molecular structure for a diverse series of drugs. *Pharmaceutical Research*, **21**, 68–82.

31 Pintore, M., van de Waterbeemd, H., Piclin, N. and Chrétien, J.R. (2003) Prediction of oral bioavailability by adaptive fuzzy partitioning. *European Journal of Medicinal Chemistry*, **38**, 427–431.

32 Bains, W., Gilbert, R., Sviridenko, L., Gascon, J.L., Scoffin, R., Birchall, K., Harvey, I. and Caldwell, J. (2002) Evolutionary computational methods to predict oral bioavailability QSPRs. *Current Opinion in Drug Discovery & Development*, **5**, 44–51.

33 Wolohan, P.R.N. and Clark, R.D. (2003) Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *Journal of Computer-Aided Molecular Design*, **17**, 65–76.

34 Hou, T., Wang, J., Zhang, W. and Xu, X. (2007) ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling*, **47**, 460–463.

35 Japertas, P., Riauba, L., Zmuidinavicius, D., Didziapetris, R. and Petrauskas, A. (2003) Classification SAR in predicting oral bioavailability, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (eds M. Ford, D. Livingstone, J. Dearden and H. van de Waterbeemd), Blackwell, Oxford, pp. 229–230.

36 Wang, J., Krudy, G., Xie, X.-Q., Wu, C. and Holland, G. (2006) Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *Journal of Chemical Information and Modeling*, **46**, 2674–2683.

37 Moda, T.L., Montanari, C.A. and Andricopulo, A.D. (2007) Hologram QSAR model for the prediction of human oral bioavailability. *Bioorganic and Medicinal Chemistry*, **15**, 7738–7745.

38 Turner, J.V. and Agatonovic-Kustrin, S. (2007) *In silico* prediction of oral bioavailability, in *ADME/Tox Approaches, Vol. 5* (eds B. Testa and H. van de Waterbeemd) in *Comprehensive Medicinal Chemistry*, 2nd edn, Elsevier, Oxford, pp. 699–724.

39 Veber, D.F., Johnson, S.R., Cheng, H.-Y., Smith, B.R., Ward, K.W. and Kopple, K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, **45**, 2615–2623.

40 Lu, J.J., Crimin, K., Goodwin, J.T., Crivori, P., Orrenius, C., Xing, L., Tandler, P.J., Vidmar, Th.J., Amore, B.M., Wilson, A.G.E., Stouten, P.F.W. and Burton, P.S. (2004) Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *Journal of Medicinal Chemistry*, **47**, 6104–6107.

41 Clark, R.D. and Wolohan, P.R.N. (2003) Molecular design and bioavailability. *Current Topics in Medicinal Chemistry*, **3**, 1269–1288.

42 Livingstone, D.J. (2003) Theoretical property predictions. *Current Topics in Medicinal Chemistry*, **3**, 1171–1192.

43 Tetko, I.V. and Livingstone, D.J. (2007) Rule-based systems to predict lipophilicity, in *ADME/Tox Approaches*, Vol. 5 (eds B. Testa and H. van de Waterbeemd), in

Comprehensive Medicinal Chemistry, 2nd edn, Elsevier, Oxford, pp. 649–668.

44 Livingstone, D.J., Ford, M.G., Huuskonen, J.J. and Salt, D.W. (2001) Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *Journal of Computer-Aided Molecular Design*, **15**, 741–752.

45 Jorgensen, W.L. and Duffy, E.M. (2002) Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews*, **54**, 355–366.

46 Manallack, D.T., Tehan, B.G., Gancia, E., Hudson, B.D., Ford, M.G., Livingstone, D.J., Whitley, D.C. and Pitt, W.R. (2003) A consensus neural network-based technique for discriminating soluble and insoluble compounds. *Journal of Chemical Information and Computer Sciences*, **43**, 674–679.

47 Morris, J.J. and Bruneau, P.P. (2000) Prediction of physicochemical properties, in *Virtual Screening for Bioactive Molecules* (eds H.J. Böhm and G. Schneider), John Wiley & Sons, Ltd, Chichester, pp. 33–58.

48 Egan, W.J. and Lauri, G. (2002) Prediction of intestinal permeability. *Advanced Drug Delivery Reviews*, **54**, 273–289.

49 Kulkarni, A., Yi, H. and Hopfinger, A.J. (2002) Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *Journal of Chemical Information and Computer Sciences*, **42**, 331–342.

50 van de Waterbeemd, H. and Gifford, E. (2003) ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery*, **2**, 192–204.

51 van de Waterbeemd, H. (2007) *In silico* models to predict oral absorption, in *ADME/Tox Approaches*, Vol. 5 (eds B. Testa and H. van de Waterbeemd), in *Comprehensive Medicinal Chemistry*, 2nd edn, Elsevier, Oxford, pp. 669–698.

52 Kratochwil, N.A., Huber, W., Muller, F., Kansy, M. and Gerber, P.R. (2002) Predicting plasma protein binding of drugs: a new approach. *Biochemical Pharmacology*, **64**, 1355–1374.

53 Stouch, T.R. and Gudmundsson, O. (2002) Progress in understanding the structure–activity relationships of P-glycoprotein. *Advanced Drug Delivery Reviews*, **54**, 315–328.

54 Zhang, E.Y., Phelps, M.A., Cheng, C., Ekins, S. and Swaan, P.W. (2002) Modeling of active transport systems. *Advanced Drug Delivery Reviews*, **54**, 329–354.

55 Jones, J.P., Mysinger, M. and Korzekwa, K.R. (2002) Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **30**, 7–12.

56 Arimoto, R. (2006) Computational models for predicting interactions with cytochrome P 450 enzyme. *Current Topics in Medicinal Chemistry*, **6**, 1609–1618.

57 Boobis, A., Gundert-Remy, U., Kremers, P., Macheras, P. and Pelkonen, O. (2002) *In silico* prediction of ADME and pharmacokinetics. Report of an expert meeting organized by COST B15. *European Journal of Pharmaceutical Sciences*, **17**, 183–193.

58 Terfloth, L., Bienfait, B. and Gasteiger, J. (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *Journal of Chemical Information and Modeling*, **47**, 1688–1701.

59 Sheridan, R.P., Korzekwa, K.R., Torres, R.A. and Walker, M.J. (2007) Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *Journal of Medicinal Chemistry*, **50**, 3173–3184.

60 Afzelius, L., Hasselgren Arnby, C., Broo, A., Carlsson, L., Isaksson, C., Jurva, U., Kjellander, B., Kolmodin, K., Nilsson, K., Raubacher, F. and Weidolf, L. (2007) State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. *Drug Metabolism Reviews*, **39**, 61–86.

61 Ekins, S., Nikolsky, Y. and Nikolskaya, T. (2005) Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in Pharmacological Sciences*, **26**, 202–209.

62 Greene, N., Judson, P.N., Langowski, J.J. and Marchant, C.A. (1999) Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR and QSAR in Environmental Research*, **10**, 299–314.

63 Testa, B., Balmat, A.L., Long, A. and Judson, P. (2005) Predicting drug metabolism? An evaluation of the expert system METEOR. *Chemistry & Biodiversity*, **2**, 872–885.

64 Crivori, P., Zamora, I., Speed, B., Orrenius, C. and Poggesi, I. (2004) Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *Journal of Computer-Aided Molecular Design*, **18**, 155–166.

65 Ekins, S., de Groot, M.J. and Jones, J.P. (2001) Pharmacophore and three dimensional quantitative structure–activity relationship methods for modelling cytochrome P450 active sites. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **29**, 936–944.

66 Crivori, P. and Poggesi, I. (2006) Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *European Journal of Medicinal Chemistry*, **41**, 795–808.

67 Kuhn, B., Jacobsen, W., Christians, U., Benet, L.Z. and Kollman, P.A. (2001) Metabolism of sirolimus and its derivative everolimus by cytochrome P450 3A4: insights from docking, molecular dynamics and quantum chemical calculations. *Journal of Medicinal Chemistry*, **44**, 2027–2034.

68 Afzelius, L., Raubacher, F., Karlén, A., Jörgensen, F.S., Andersson, T.B., Masimirembwa, C. and Zamora, I. (2004) Structural analysis of CYP2C9 and CYP2C5 and an evaluation of commonly used molecular modelling techniques. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **32**, 1–12.

69 Boyer, S. and Zamora, I. (2002) New methods in predictive metabolism. *Journal of Computer-Aided Molecular Design*, **16**, 403–413.

70 van de Waterbeemd, H. and Jones, B.C. (2003) Predicting oral absorption and bioavailability. *Progress in Medicinal Chemistry*, **41**, 1–59.

71 Stouch, T.R., Kenyon, J.R., Johnson, S.R., Chen, X.-Q., Doweyko, A. and Li, Y. (2003) *In silico* ADME/Tox: why models fail. *Journal of Computer-Aided Molecular Design*, **17**, 83–92.

72 Yamashita, F. and Hashida, M. (2004) *In silico* approaches for predicting ADME properties of drugs. *Drug Metabolism and Pharmacokinetics*, **19**, 327–338.

73 Mager, D.E. (2006) Quantitative structure–pharmacokinetic/pharmacodynamic relationships. *Advanced Drug Delivery Reviews*, **58**, 1326–1356.

74 De Buck, S.S., Sinha, V.K., Fenu, L.A., Nijsen, M.J., Mackie, C.E. and Gilissen, R.A.H.J. (2007) Prediction of human pharmacokinetics using physiologically-based modelling: a retrospective analysis of 26 clinically tested drugs. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **35**, 1766–1780.

75 Theil, F.-P., Guentert, T.W., Haddad, S. and Poulin, P. (2003) Utility of physiologically-based pharmacokinetic models to drug development and rational drug discovery candidate selection. *Toxicology Letters*, **138**, 29–49.

76 Dickins, M. and van de Waterbeemd, H. (2004) Simulation models for drug disposition and drug interactions. *Drug Discovery Today: Biosilico*, **2**, 38–45.

77 Germani, M., Crivori, P., Rocchetti, M., Burton, P.S., Wilson, A.G.E., Smith, M.E. and Poggesi, I. (2007) Evaluation of a basic physiologically-based pharmacokinetic model for simulating the first-time-in-animal study. *European Journal of Pharmaceutical Sciences*, **31**, 190–201.

78 Andersen, M.E. (1995) Physiologically-based pharmacokinetic (PBPK) models in the study of the disposition and biological effects of xenobiotics and drugs. *Toxicology Letters*, **82/83**, 341–348.

79 Kuentz, M., Nick, S., Parrott, N. and Röthlisberger, D. (2006) A strategy for preclinical formulation development using GastroPlus as pharmacokinetic simulation tool and a statistical screening design applied to a dog study. *European Journal of Pharmaceutical Sciences*, **27**, 91–99.

80 Kawai, R., Lemaire, M., Steimer, J.-L., Bruelisauer, A., Niederberger, W. and Rowland, M. (1994) Physiologically-based pharmacokinetic study on a cyclosporine derivative, SDZ IMM 125. *Journal of Pharmacokinetics and Biopharmaceutics*, **22**, 327–365.

81 Charnik, S.B., Kawai, R., Nefelman, J.R., Lemaire, M., Niederberger, W. and Sato, H. (1995) Perspectives in pharmacokinetics. Physiologically-based pharmacokinetic modeling as a tool for drug development. *Journal of Pharmacokinetics and Biopharmaceutics*, **23**, 231–235.

82 Leahy, D.E. (2003) Progress in simulation modelling for pharmacokinetics. *Current Topics in Medicinal Chemistry*, **3**, 1257–1268.

83 Stoner, C.L., Cleton, A., Johnson, K., Oh, D.-M., Hallak, H. Brodfuehrer, Surendran, N. and Han, H.-K. (2004) Integrated oral bioavailability projection using *in vitro* screening data as a selection tool in drug discovery. *International Journal of Pharmaceutics*, **269**, 241–249.

84 Cai, H., Stoner, C., Reddy, A., Freiwald, S., Smith, D., Winters, R., Stankovic, C. and Surendran, N. (2006) Evaluation of an integrated *in vitro* – *in silico* PBPK (physiologically-based pharmacokinetic) model to provide estimates of human bioavailability. *International Journal of Pharmaceutics*, **308**, 133–139.

85 Verwei, M., Freidig, A.P., Havenaar, R. and Groten, J.P. (2006) Predicted serum folate concentrations based on *in vitro* studies and kinetic modelling are consistent with measured folate concentrations in humans. *The Journal of Nutrition*, **136**, 3074–3078.

86 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, **6**, 881–890.

87 Gu, C.H., Li, H., Levons, J., Leatz, K., Gandhi, R.B., Raghavan, K. and Smith, R.L. (2007) Predicting effect of food on extent of drug absorption based on physicochemical properties. *Pharmaceutical Research*, **24**, 1118–1130.

17
Simulations of Absorption, Metabolism, and Bioavailability

Michael B. Bolger, Robert Fraczkiewicz, and Viera Lukacova

Abbreviations

ACAT	Advanced compartmental absorption and transit model
ADMET	Absorption, distribution, metabolism, excretion, and toxicity
Caco-2	Adenocarcinoma cell line derived from human colon
CAT	Compartmental absorption and transit model
EPA	Environmental Protection Agency
FDA	Food and Drug Administration
GI	Gastrointestinal
MRTD	Maximal recommended therapeutic dose
P-gp	P-Glycoprotein
PK	Pharmacokinetics
PBPK	Physiologically-based pharmacokinetics
RBA	Ratio of the estimated estrogen receptor-binding affinities for 17b-estradiol divided by the affinity estimated for the unknown molecule
TEER	Transcellular epithelial electrical resistance

Symbols

C_p	Plasma concentration
$\Delta\log P$	Difference between log P in octanol/water and log D at a given pH
HIA%	Percent human intestinal absorption across apical membrane of the enterocyte.
log D	Logarithm of the distribution coefficient, usually in octanol/water at a specified pH
log P	Logarithm of the partition coefficient, usually in octanol/water (for neutral species)
MW	Molecular weight

P_{app}	Apparent permeability
pK_a	Ionization constant in water
S_w	Solubility
SITT	Small intestinal transit time (3.3 h = 199 min)
V	Volume
V_{ss}	Volume at steady state

17.1 Introduction

Ever since this chapter was first published in 2002, there has been an explosion of awareness and research in the area of *in silico* methods for early assessment of absorption and bioavailability [1–5]. Physiologically-based mechanistic gastrointestinal simulation and physiologically-based pharmacokinetic (PBPK) models of absorption and distribution are now routinely used to identify and rank drug discovery candidates with regard to their absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties [6–12]. Biopharmaceutical property inputs for such simulations can be derived from *in silico* estimations or *in vitro* experiments [13, 14]. The one property that still requires experimental data for quantitative estimation is metabolism. However, computational approaches have advanced rapidly in the last few years [15]. Formulation of development candidates can be enhanced by using this same type of simulation [16]. A new area for computational approaches to absorption and bioavailability is the application of systems biology [17, 18]. Finally, model-based drug development and clinical trial simulation technology have become a mainstay for regulatory agencies [19–22].

The observed oral bioavailability and biological activity of a particular therapeutic agent can be broken down into components that reflect delivery to the intestine, liberation from formulated product (gastric emptying, intestinal transit, pH, and food), absorption from the lumen (dissolution, lipophilicity, particle size, and active transport), intestinal and hepatic first-pass metabolism, distribution into tissues, and subsequent excretion, and toxicity (ADMET) [23]. This chapter will focus on *in silico* approaches that have demonstrated ability to save valuable resources in the drug discovery and development process. We will review some of the recent advances in physiologically-based pharmacokinetics and discuss our results in simulating GI absorption by using the advanced compartmental absorption and transit model (ACAT).

17.2 Background

For the purposes of GI simulation, it is important to distinguish absorption (transfer of drug from the lumen of the intestine across the apical membrane into the enterocyte) from bioavailability (the fraction of administered dose that is available in the systemic circulation for interaction with the target tissue). Simulation of absorption and bioavailability must account for many factors that fall into three classes [24]. The first class represents physicochemical factors including pK_a, solubility, stability, diffusivity, lipophilicity, and salt forms. The second class comprises physiological factors including GI pH, gastric emptying, small and large bowel transit times, active transport and efflux, and gut wall and liver metabolism. The third class comprises formulation factors such as surface area, drug particle size and crystal form, and dosage forms such as solution, tablet, capsule, suspension, and modified release.

An early concept governing oral absorption of organic molecules was called the "pH-partition" hypothesis. Under this hypothesis, only the unionized form of ionizable molecules was thought to partition into the membranes of epithelial cells lining the GI tract [25, 26]. The contribution of pH to permeability and dissolution of solid dosage forms has been proven to be a critical factor, but ionized molecules have now been shown to be absorbed by a variety of mechanisms [27]. Ho and colleagues developed one of the most sophisticated early theoretical approaches to simulating drug absorption based on the diffusional transport of drugs across a compartmental membrane [28–30]. Their physical model consisted of a well-stirred bulk aqueous phase, an aqueous diffusion layer, and a heterogeneous lipid barrier composed of several compartments ending in a perfect sink. Their model represented the first example of the rigorous application of a physical model to the quantitative and mechanistic interpretation of *in vivo* absorption [31]. The simultaneous chemical equilibria and mass transfer of basic and acidic drugs were modeled and compared favorably to *in situ* measurements of intestinal, gastric, and rectal absorption in animals. The pH-partition theory was shown to be a limiting case of the more general model they developed. Because of its complexity, the diffusional mass transit model has not been widely used. In the 1980s, a simple and intuitive alternative approach based on a series of mixing tank compartments was developed [32]. Pharmacokinetic models incorporating discontinuous GI absorption from at least two absorption sites separated by *N* nonabsorbing sites have been used to explain the occurrence of double peaks in plasma concentration versus time (C_p–time) profiles for ranitidine and cimetidine [33]. A similar discontinuous oral absorption model based on two absorption compartments and two transit compartments was developed to explain the bioavailability of nucleoside analogues [34]. Amidon and Yu developed a compartmental absorption and transit model (CAT) of the GI tract based on seven equal transit time compartments [24]. Using a five-compartment GI simulation model, Norris *et al*. were able to estimate C_p–time profiles for ganciclovir [35, 36]. A physiologically-based segregated flow model (SFM) was developed to examine the influence of intestinal transport (absorption and exsorption), metabolism, flow, tissue-partitioning characteristics, and elimination in other organs on intestinal clearance, intestinal availability, and systemic bioavailability [37]. Using a completely different approach, a stochastic simulation of drug molecules moving through a cylinder of fixed radius with random geometric placement of dendritic-type virtual "villi" was able to accurately account for the observed human SI transit time distribution [38, 39]. Ito *et al.* have developed a pharmacokinetic model for drug absorption that includes metabolism by CYP3A4 inside the epithelial cells, P-gp mediated efflux into the lumen, intracellular diffusion from the luminal side to the basal side, and subsequent permeation through the basal membrane [40]. As expected, they demonstrated that the fraction of dose into the portal vein was synergistically elevated by simultaneous inhibition of both CYP3A4 and P-gp. The Simcyp Consortium Project has compiled extensive demographic and physiological data to build virtual human populations and has demonstrated good prediction of *in vivo* pharmacokinetic profiles using *in vitro* data and a simulation approach similar to the CAT model [41]. In contrast to the compartmental absorption and transit model, others have developed a simulation of the GI tract modeled as a continuous tube with spatially varying properties and continuous plug flow with dispersion of drug molecules [42, 43].

We demonstrated the utility of GI absorption simulation based on the ACAT in predicting the impact of physiological and biochemical processes on oral drug bioavailability [44, 45].

The ACAT model is loosely based on the work of Amidon and Yu who found that seven equal transit time compartments are required to represent the observed cumulative frequency distribution for small intestine transit times [24]. Their original CAT was able to explain the oral plasma concentration profiles of atenolol [46].

17.3 Use of Rule-Based Computational Alerts in Early Discovery

17.3.1 Simple Rules for Drug Absorption (Druggability)

In silico ADMET profiling of compound libraries in early discovery has become a valuable addition to the research toolbox of computational and medicinal chemists. A computational alert was developed by Lipinski based on the physicochemical characteristics of approximately 90% of 2245 drugs with USAN names that have had clinical exposure found in the World Drug Index [47]. Most of these drugs have entered at least phase-II clinical trials. The rule-of-5 has had a significant impact on early drug discovery and has stimulated development of similar computational alerts [48–52]. Application of a computational alert to compound libraries prior to synthesis helps limit the requirement of *in vitro* testing to those compounds that are most likely to have "drug-like" characteristics.

We have developed a new set of rules, called "ADMET Risk," that contains cutoffs for human jejunal permeability, pH of a saturated solution of the drug in water, partial charge on H-bond donors and acceptors, an indicator variable for permanent cations, and a low-level cutoff for log *P*. The ADMET Risk rules and two-letter abbreviations are listed as follows:

LP	$S + \log P < -1.006$
Pr	$S + P_{\mathrm{eff}} < 0.314$
pH	$S + \mathrm{pH} < 3.12$
Hd	$\mathrm{HBDCH} > 1.34$
Ha	$\mathrm{HBACH} < -6.6$
PC	$\mathrm{QuaAmine\_{>}}[N+] > 0$,

where $S + \log P$ represents Simulations Plus artificial neural network model of log P (octanol/water); $S + P_{\mathrm{eff}}$ represents Simulations Plus predicted human jejunal effective permeability; $S + \mathrm{pH}$ represents Simulations Plus estimation of the pH of a saturated solution of the drug in pure water; HBDCH represents partial charge on hydrogen-bond donors; HBACH represents partial charge on hydrogen-bond acceptors; and $\mathrm{QuaAmine\_{>}}[N+]$ represents an indicator variable for the presence of quaternary amines, sulfonium cations, or diazonium cations.

All the descriptors and properties necessary to calculate ADMET Risk are generated by the software program ADMET Predictor (formerly called QMPRPlus) (Simulations Plus, Inc.). The current set of ADMET Predictor computational models for biopharmaceutical properties is listed below:

- multiprotic ionization constants (pK_a);
- log P ($\log_{10}$ of octanol–water partition coefficient for unionized molecules);
- log D ($\log_{10}$ of octanol–water distribution coefficient for all molecular species);
- effective permeability (human jejunum) (P_{eff}, cm/s × 10^4);
- average effective permeability (entire small intestine) (P_{avg}, cm/s × 10^4);
- MDCK cell monolayer permeability (P_{app}, nm/s);
- blood–brain barrier permeation (high, low, undecided)
- saturated aqueous solubility in pure water (mg/ml);
- saturated aqueous pH in pure water;
- saturated intrinsic solubility in pure water (mg/ml);
- saturated solubility at user-specified pH (mg/ml);
- salt solubility factor;
- diffusivity (diffusion coefficient, cm^2/s);
- molal volume (cm^3/mol);
- percentage unbound to blood plasma proteins (%);
- pharmacokinetic volume of distribution (l/kg);
- maximum recommended therapeutic dose (mg/kg/day);
- estrogen receptor toxicity;
- lethal acute toxicity against fathead minnow (mg/l/96 h);
- affinity toward hERG K$^+$ channel (a measure of cardiac toxicity);
- carcinogenicity in rats and mice;
- Ames mutagenicity in *Salmonella*;
- metabolism rate constants ($V_{\max}$, K_m) for five main CYP enzymes in human (1A2, 2C19, 2C9, 2D6, and 3A4);
- inhibition of HIV-1 integrase;
- simulated fraction absorbed in human.

These models are based on the calculation of 297 molecular descriptors obtained by parsing the 2D or 3D structures of drug molecules as represented either in SMILES string format or as ISIS-.RDF, .SDF, or .MOL file format (MDL Information Systems, Inc., http://www.mdli.com/). Molecular descriptor values are used as inputs to independent mathematical models to generate estimates for each of the biopharmaceutical properties listed above. Using these property estimates or experimentally determined properties as inputs to the ACAT model, drug molecules may be classified according to their ADMET qualities. While no computer program is able to estimate the exact experimental values for these properties, we have demonstrated that the estimated values generated by our method are sufficiently accurate to allow rank ordering of a large number of compounds for "overall ADMET quality." In fact, *in vitro* methods also fail to predict *in vivo* ADMET properties under certain conditions. We have found that the *in silico* methods are comparable to *in vitro* methods for predictive capability.

We tested the usefulness of these *in silico* biopharmaceutical properties in predicting the rank order of human intestinal absorption (HIA%). The percentage absorbed for 266 drug molecules was collected from various literature sources [53, 54]. These drugs are known to be absorbed by a number of mechanisms including passive transcellular, passive paracellular, and active transport mechanism and some were actively effluxed. Starting from three-dimensional structures calculated by CORINA (http://www2.ccc.uni-erlangen.de/software/corina/), we used ADMET Predictor to generate molecular descriptors, to estimate the biopharmaceutical properties, and to calculate the ADMET Risk values as described above.

Results from ADMET Predictor using a rule-based method were first ranked by the value of ADMET Risk and, within an ADMET Risk category (0–5), were ranked by increasing permeability (*in silico* estimate of human effective permeability). Table 17.1 lists the experimental HIA% and compares their rank order with the rank order predicted for 266 drugs using the rule-based method. We found a significant Spearman rank correlation coefficient of 0.70 ($p < 0.001$) when the rule-based method of predicting the rank order of oral faction absorbed was applied to 209 passively absorbed compounds. Figure 17.1 shows a plot of the experimental rank order compared with the ADMET Risk-based rank order method for 209 compounds that are known to be absorbed through a passive transcellular or paracellular route. It can be seen that there is a good correlation for the passively absorbed compounds. By contrast, Figure 17.2 shows a similar plot for the 43 compounds that are known to be absorbed by an active route or are known to be actively effluxed. For these compounds the correlation is nonexistent.

In a comparison between Lipinski's rules and the ADMET Risk, we have found that the "rule-of-5" accurately identifies only a fraction of the compounds that have experimental absorption less than 50%. This high false-positive result allows many of the poor compounds to go undetected. By contrast, the ADMET Risk identifies a much higher fraction of the unfavorable compounds in addition to many of the well-absorbed compounds.

ADMET Predictor was used to generate *in silico* estimates of log *P*, aqueous solubility, the pH of a saturated solution in water, partial charges on H-bond donors

Table 17.1 Table of 266 common drugs used in the comparison of rule-based ranking with mechanistic simulation-based ranking.

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
AAFC	32	22	3	17	n/a	FC1N2C(=NC(=N)C=1)OC3C2OC(C3O)CO
Acarbose	2	9	4	1	n/a	O=CC(C(C(OC1OC(C(OC2OC(C(NC3C=C(C(O)C(O)C3O)CO)C(O)C2O)C)C(O)C1O)CO)C(O)CO)O)O
Acebutolol	80	50	0	79	n/a	C(=O)(c1c(OCC(O)CNC(C)C)ccc(NC(=O)CCC)c1)C
Acetaminophen	80	51	0	87	n/a	C(=O)(Nc1ccc(O)cc1)C
Acetylsalicylic acid	84	58	1	53	n/a	C(=O)(Oc1c(C(=O)O)cccc1)C
Acrivastine	88	39	0	46	n/a	C(c1nc(C=CC(=O)O)ccc1)(c2ccc(cc2)C)=CCN3CCCC3
Acyclovir	23	15	3	18	n/a	O=C1c2c(N=C(N1)N)n(cn2)COCCO
Adefovir	16	13	4	9	n/a	O=P(O)(O)COCCn1c2c(c(ncn2)N)nc1
Adriamycin	5	18	3	26	P-gp	C(=O)(C1(O)C(OC2OC(C(O)C(N)C2)C)Cc3c(c4C(=O)c5c(OC)cccc5C(=O)c4c(c3C1)O)O)CO
Allopurinol	90	151	0	143	Hypoxanthine	c12c(ncnc1[nH]nc2)O
Alpha-difluoromethylornithine	55	70	3	34	CAT1?	C(C(C(F)F)(CCCN)N)(=O)O
Alprazolam	90	49	0	71	n/a	Clc1cc2C(c3ccccc3)=NCc4nnc(n4-c2cc1)C
Alprenolol	93	81	0	121	n/a	O(c1c(cccc1)CC=C)CC(O)CNC(C)C
Amantadine	95	85	0	62	n/a	C12(N)CC3CC(C1)CC(C2)C3
Amiloride	50	65	1	91	n/a	C(=O)(c1c(nc(c(Cl)n1)N)N)NC(=N)N
Aminopyrine	100	123	0	113	n/a	N(C1C(=O)N(c2ccccc2)N(C=1C)C)(C)C
Amoxicillin	93	171	2	58	PepT1	C(=O)(C(c1ccc(O)cc1)N)NC2C(=O)N3C(C(=O)O)C(SC32)(C)C
Amphetamine	90	46	0	62	n/a	c1(ccccc1)CC(N)C
Amphotericin B	3	10	3	12	n/a	C(=O)(O)C1C2OC(O)(CC1O)CC(O)CC(O)C(O)CCC(O)CC(O)CC(=O)OC(C(C(O)C(C=CC=CC=CC=CC=CC=CC=CC(OC3OC(C(O)C(C3O)N)C)C2)C)C)C

(*Continued*)

Table 17.1 (*Continued*)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Ampicillin	62	84	2	60	PepT1	C(=O)(C(c1ccccc1)N)NC2C(=O)N3C(C(=O)O)C(SC32)(C)C
Amrinone	93	52	0	61	n/a	O=C1C(=CC(c2ccncc2)=CN1)N
Antipyrine	97	98	0	117	n/a	O=C1N(c2ccccc2)N(C(=C1)C)C
Ascorbic acid	35	23	4	11	n/a	O=C1C(=C(O)C(O1)C(O)CO)O
Atenolol	50	29	1	49	n/a	C(=O)(Cc1ccc(OCC(O)CNC(C)C)cc1)N
Atropine	98	58	0	37	n/a	C(=O)(OC1CC2N(C(C1)CC2)C)C(c3ccccc3)CO
Azithromycin	37	24	3	15	n/a	N(C1C(O)C(OC2C(O)(CC(CN(C(C(C(O)(C(OC(=O)C(C(OC3OC (C(C(OC)(C3)C)O)C)C2C)C)CC)C)O)C)C)C)C)OC(C1)C)(C)C
Azosemide	20	11	1	28	n/a	O=S(=O)(c1c(Cl)cc(c(-c2nnn[nH]2)c1)NCc3sccc3)N
Aztreonam	1	5	3	20	n/a	C(C(=O)NC1C(=O)N(S(=O)(=O)O)C1C)(=NOC(C(=O)O)(C)C) C2=CS=C(N2)N
Benazepril	37	14	0	52	n/a	C(=O)(OCC)C(NC1C(=O)N(c2c(cccc2)CC1)CC(=O)O) CCc3ccccc3
Benserazide	90	74	2	30	n/a	C(=O)(C(CO)N)NNCc1c(c(c(O)cc1)O)O
Benzylpenicillin	30	42	1	104	Intes. pept. carrier	C(=O)(NC1C(=O)N2C(C(=O)O)C(SC21)(C)C)Cc3ccccc3
Betaxolol	90	45	0	50	n/a	O(c1ccc(cc1)CCOCC2CC2)CC(O)CNC(C)C
Bornaprine	100	72	0	53	n/a	C(=O)(OCCCN(CC)CC)C1(c2ccccc2)C3CC(C1)CC3
Bretyliumtosylate	23	31	3	38	n/a	Brc1c(cccc1)C[N+](CC)(C)C
Bromazepam	84	59	0	100	n/a	O=C1Nc2c(C(c3ncccc3)=NC1)cc(Br)cc2
Bromocriptine	28	38	1	78	n/a	C(=O)(NC1(C(=O)N2C(O1)(O)C3N(C(=O)C2CC(C)C)CCC3)C (C)C)C4C=C5c6c7c(c(Br)[nH]c7ccc6)CC5N(C4)C
Bumetanide	96	189	1	100	Bile acid	C(=O)(c1cc(S(=O)(=O)N)c(Oc2ccccc2)c(NCCCC)c1)O
Bupropion	87	66	0	116	n/a	C(=O)(c1cc(Cl)ccc1)C(NC(C)(C)C)C
Caffeine	100	118	0	96	n/a	O=C1c2c(ncn2C)N(C(=O)N1C)C
Camazepam	100	134	0	130	n/a	C(N(C)C)(=O)OC1C(=O)N(c2c(C(c3ccccc3)=N1)cc(Cl)cc2)C

Capreomycin	50	63	4	8	n/a	C(=O)(NCC1C(=O)NC(C(=O)NC(C(=O)NCC(C(=O)NC(C(=O)N1)C)N)C2NC(=N)NCC2)=CNC(=O)N)CC(CCCN)N
Captopril	84	126	1	115	Intes. pept. carrier	C(=O)(N1C(C(=O)O)CCC1)C(CS)C
Carbamazepine	97	97	0	112	n/a	C(=O)(N1c2c(cccc2)CCc3c1cccc3)N
Carfecillin	99	207	0	147	Intes. pept. carrier	C(=O)(Oc1ccccc1)C(C(=O)NC2C(=O)N3C(C(=O)O)C(SC32)(C)C)c4ccccc4
Cefadroxil	100	220	2	61	Intes. pept. carrier	C(=O)(C1=C(CSC2N1C(=O)C2NC(=O)C(c3ccc(O)cc3)N)C)O
Cefatrizine	75	100	2	57	Intes. pept. carrier	C(=O)(C1=C(CSc2cnn[nH]2)CSC3N1C(=O)C3NC(=O)C(c4ccc(O)cc4)N)O
Cefetametpivoxil	47	26	1	48	n/a	C(C(=O)NC1C(=O)N2C(C(=O)OCOC(C(C)(C)C)=O)=C(CSC21)C)(=NOC)C3=CS=C(N3)N
Ceftizoxime	72	97	2	52	n/a	C(C(=O)NC1C(=O)N2C(C(=O)O)=CCSC21)(=NOC)C3=CSC(=N)N3
Ceftriaxone	1	6	2	29	n/a	C(C(=O)NC1C(=O)N2C(C(=O)O)=C(CSC3=NC(=O)C(=O)NN3C)CSC21)(=NOC)C4=CS=C(N4)N
Cefuroxime	1	3	4	5	n/a	C(C(=O)NC1C(=O)N2C(C(=O)O)=C(COC(=O)N)CSC21)(=NOC)c3occc3
Cefuroximeaxetil	44	56	2	41	n/a	C(C(=O)NC1C(=O)N2C(C(=O)OC(OC(=O)C)C)=C(COC(=O)N)CSC21)(=NOC)c3occc3
Cephalexin	100	224	2	62	PepT1	C(=O)(C1=C(CSC2N1C(=O)C2NC(=O)C(c3ccccc3)N)C)O
Chloramphenicol	90	78	1	59	n/a	C(=O)(C(Cl)Cl)NC(C(c1ccc([N+](=O)[O-])cc1)O)CO
Chlorothiazide	49	27	1	33	n/a	O=S(=O)(c1c(Cl)cc2c(S(=O)(=O)NC=N2)c1)N
Cicaprost	100	69	0	43	n/a	C(#CC1C(O)CC2C1CC(=CCOCC(=O)O)C2)C(C(CC#CCC)C)O
Cidofovir	3	4	5	1	n/a	O=P(O)(O)COC(CN1C(=O)NC(=CC1)N)CO
Cimetidine	64	22	1	19	n/a	C(#N)N=C(NCCSCC1=NC=NC1C)NC
Ciprofloxacin	69	24	2	14	n/a	C(=O)(C1C(=O)c2c(N(C=1)C3CC3)cc(c(F)c2)N4CCNCC4)O
Cisapride	100	116	0	91	n/a	C(=O)(c1c(OC)cc(c(Cl)c1)N)NC2C(OC)CN(CCCOc3ccc(F)cc3)CC2

(*Continued*)

Table 17.1 (Continued)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Clofibrate	97	99	0	131	n/a	C(C(Oc1ccc(Cl)cc1)(C)C)(=O)OCC
Clonazepam	98	104	0	111	n/a	O=C1Nc2c(C(c3c(Cl)cccc3)=NC1)cc([N+](=O)[O-])cc2
Clonidine	95	92	0	110	n/a	N(c1c(Cl)cccc1Cl)=C2NCCN2
Codeine	95	86	0	64	n/a	O(c1c2c3c(cc1)CC4N(CCC35C(O2)C(O)C=CC54)C)C
Corticosterone	100	106	1	42	n/a	C(=O)(C1C2(C(C3C(C4(C(=CC(=O)CC4)CC3)C)C(O)C2)CC1)C)CO
Cromolyn sodium	0.4	2	2	44	n/a	C(=O)(C1Oc2c(C(=O)C=1)c(OCC(O)COc3c4C(=O)C=C(C(=O)O)Oc4ccc3)ccc2)O
Cycloserine	73	99	1	89	hPAT1	O=C1C(N)CON1
Cymarin	47	18	1	16	n/a	O=C1OCC(=C1)C2C3(C(O)(C4C(C5(C(O)(CC(OC6OC(C(O)C(OC)C6)C)CC5)CC4)C=O)CC3)CC2)C
Cyproterone acetate	100	109	0	67	n/a	C(=O)(OC1(C(=O)C)C2(C(C3C=C(Cl)C4C(C5C(C(=O)C=4)C5)(C3CC2)C)CC1)C)C
Desipramine	100	73	0	73	n/a	c12c(cccc1)CCc3c(N2CCCNC)cccc3
Dexamethasone	80	30	1	22	n/a	C(=O)(C1(C2(C(C3C(F)(C4(C(=CC(=O)C=C4)CC3)C)C(O)C2)CC1C)C)O)CO
Diazepam	100	131	0	127	n/a	O=C1N(c2c(C(c3ccccc3)=NC1)cc(Cl)cc2)C
Diclofenac	100	264	0	259	n/a	C(=O)(O)Cc1c(Nc2c(Cl)cccc2Cl)cccc1
Diflunisal	100	135	0	132	n/a	C(=O)(c1c(O)ccc(-c2c(F)cc(F)cc2)c1)O
Digoxin	81	33	3	6	OATP/P-gp	O=C1OCC(=C1)C2C3(C(O)(C4C(C5(C(CC(OC6OC(C(OC7OC(C(OC8OC(C(O)C(O)C8)C)C(O)C7)C)C(O)C6)C)CC5)CC4)C)CC3O)CC2)C
Dihydrocodeine	89	40	0	35	n/a	O(c1c2c3c(cc1)CC4N(CCC35C(O2)C(O)CCC54)C)C
Diltiazem	100	115	0	89	n/a	C(=O)(OC1C(=O)N(c2c(SC1c3ccc(OC)cc3)cccc2)CCN(C)C)C
Distigmine	8	7	3	12	n/a	C(N(CCCCCCN(C(=O)Oc1c[n+](ccc1)C)C)C)(=O)Oc2c[n+](ccc2)C

Disulfuram	97	55	0	57	n/a	C(N(CC)CC)(=S)SSC(N(CC)CC)=S
Doxorubicin	12	8	3	7	n/a	C(=O)(C1(O)Cc2c(c3C(=O)c4c(C(=O)c3c(c2C(OC5OC(C(O)C(N)C5)C)C1)O)c(OC)ccc4)O)CO
Enalapril	66	90	0	185	Intes. pept. carrier	C(=O)(OCC)C(NC(C(=O)N1C(C(=O)O)CCC1)C)CCc2ccccc2
Enalaprilat	25	16	2	26	n/a	C(=O)(N1C(C(=O)O)CCC1)C(NC(C(=O)O)CCc2ccccc2)C
Erythromycin	35	13	3	9	n/a	N(C1C(O)C(OC2C(O)(CC(C(=O)C(C(C(O)(C(OC(=O)C(C(OC3OC(C(C(OC)(C3)C)O)C)C2C)C)CC)C)O)C)C)C)OC(C1)C)(C)C
Ethambutol	80	52	1	57	n/a	C(NCCNC(CO)CC)(CO)CC
Ethinylestradiol	100	74	0	54	n/a	C(C1(C2(C(C3C(c4c(cc(O)cc4)CC3)CC2)CC1)C)O)#C
Etoposide	50	64	3	24	P-gp	O=C1OCC2C(OC3OC4C(OC(OC4)C)C(O)C3O)c5c(cc6c(OCO6)c5)C(c7cc(OC)c(c(OC)c7)O)C12
Famciclovir	77	47	0	61	n/a	C(=O)(OCC(COC(=O)C)CCn1c2c(nc1)cnc(n2)N)C
Famotidine	38	25	3	13	n/a	C(=NS(=O)(=O)N)(CCSCc1nc(N=C(N)N)sc1)N
Felbamate	90	42	1	26	n/a	C(=O)(OCC(c1ccccc1)COC(=O)N)N
Felodipine	88	69	0	83	n/a	C(=O)(OCC)C1=C(NC(=C(C(=O)OC)C1c2c(Cl)c(Cl)ccc2)C)C
Fenclofenac	100	136	0	133	n/a	C(=O)(O)Cc1c(Oc2c(Cl)cc(Cl)cc2)cccc1
Fenoterol	60	38	1	54	n/a	c1(C(O)CNC(Cc2ccc(O)cc2)C)cc(O)cc(O)c1
Flecainide	81	34	0	69	n/a	C(=O)(c1c(OCC(F)(F)F)ccc(OCC(F)(F)F)c1)NCC2NCCCC2
Fluconazole	95	53	0	51	n/a	C(c1c(F)cc(F)cc1)(O)(Cn2ncnc2)Cn3ncnc3
Flumazenil	95	89	0	77	n/a	C(=O)(OCC)C1=NCN2c3c(C(=O)N(CC12)C)cc(F)cc3
Fluoxetine	80	54	0	135	n/a	C(F)(F)(F)c1ccc(OC(c2ccccc2)CCNC)cc1
Fluvastatin	100	226	1	111	n/a	C(=O)(O)CC(O)CC(C=Cc1c(c2c(n1C(C)C)cccc2)-c3ccc(F)cc3)O
Foscarnet	17	14	3	21	n/a	C(=O)(P(=O)(O)O)O
Fosfomycin	31	21	3	14	n/a	O=P(O)(O)C1OC1C
Fosinopril	36	50	0	144	n/a	C(=O)(OC(OP(=O)(CC(=O)N1C(C(=O)O)CC(C2CCCCC2)C1)CCCCc3ccccc3)C(C)C)CC
Fosmidomycin	30	19	4	10	n/a	N(C=O)(O)CCCP(=O)(O)O

(*Continued*)

Table 17.1 (*Continued*)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Furosemide	61	39	2	23	n/a	C(=O)(c1c(NCc2occc2)cc(Cl)c(S(=O)(=O)N)c1)O
Gabapentin	59	34	1	45	LAT1	C(=O)(O)CC1(CN)CCCCC1
Gallopamil	100	127	0	119	n/a	C(C(c1cc(OC)c(OC)c(OC)c1)(C(C)C)CCCN(CCc2cc(OC)c(OC)cc2)C)#N
Ganciclovir	3	5	3	8	n/a	O=C1c2c(N=C(N1)N)n(cn2)COC(CO)CO
Gentamicin	1	2	4	4	n/a	O(C1OC(C(NC)C)CCC1N)C2C(O)C(OC3OCC(O)(C(NC)C3O)C)C(N)CC2N
Gliclazide	65	42	1	35	n/a	C(=O)(NS(=O)(=O)c1ccc(cc1)C)NN2CC3C(C2)CCC3
Glipizide	100	105	1	34	n/a	C(=O)(c1ncc(nc1)C)NCCc2ccc(S(=O)(=O)NC(=O)NC3CCCCC3)cc2
Glyburide	100	71	0	34	n/a	C(=O)(c1c(OC)ccc(Cl)c1)NCCc2ccc(S(=O)(=O)NC(=O)NC3CCCCC3)cc2
Glycine	100	236	1	118	hPAT1	C(=O)(O)CN
Granisetron	100	113	0	82	n/a	C(=O)(c1c2c(n(n1)C)cccc2)NC3CC4N(C(C3)CCC4)C
Guanabenz	80	31	1	30	n/a	C(=N)(NN=Cc1c(Cl)cccc1Cl)N
Guanoxan	50	28	1	46	n/a	C(=N)(NCC1Oc2c(OC1)cccc2)N
Hydrochlorothiazide	65	23	2	13	n/a	O=S(=O)(c1c(Cl)cc2c(S(=O)(=O)NCN2)c1)N
Hydrocortisone	91	79	1	41	n/a	C(=O)(C1(C2(C(C3C(C4(C(=CC(=O)CC4)CC3)C)C(O)C2)CC1)C)O)CO
Ibuprofen	95	93	0	124	n/a	C(=O)(C(c1ccc(cc1)CC(C)C)C)O
Imipramine	100	133	0	129	n/a	N(CCCN1c2c(cccc2)CCc3c1cccc3)(C)C
Indapamide	97	96	0	95	n/a	C(=O)(c1cc(S(=O)(=O)N)c(Cl)cc1)NN2c3c(cccc3)CC2C
Indomethacin	100	124	0	114	n/a	C(=O)(c1ccc(Cl)cc1)n2c3c(c(c2C)CC(=O)O)cc(OC)cc3
Iothalamate sodium	1.9	8	1	58	n/a	C(=O)(c1c(I)c(C(=O)O)c(I)c(c1I)NC(=O)C)NC
Isoniazid	80	28	0	48	n/a	C(=O)(c1ccncc1)NN
Isoxicam	100	67	0	42	n/a	C(=O)(C1=C(c2c(S(=O)(=O)N1C)cccc2)O)NC3=CC(ON3)C

Isradipine	92	50	1	18	n/a	C(=O)(OC(C)C)C1=C(NC(=C(C(=O)OC)C1c2c3c(ccc2)NON3)C)C
Kanamycin	1	4	4	6	n/a	O(C1OC(C(O)C(C1O)N)CO)C2C(O)C(OC3OC(C(O)C(O)C3O)CN)C(N)CC2N
Ketoprofen	92	51	0	74	n/a	C(=O)(c1cc(C(C(=O)O)C)ccc1)c2ccccc2
Ketorolac	90	76	0	92	n/a	C(=O)(c1n2c(cc1)C(C(=O)O)CC2)c3ccccc3
K-Strophanthoside	16	12	4	4	n/a	O=C1OCC(=C1)C2C3(C(O)(C4C(C5(C(O)(CC(OC6OC(C(OC7OC(C(OC8OC(C(O)C(O)C8O)CO)C(O)C7O)CO)C(OC)C6)C)CC5)CC4)C=O)CC3)CC2)C
Labetalol	95	91	1	60	n/a	C(=O)(c1c(O)ccc(C(O)CNC(CCc2ccccc2)C)c1)N
Lactulose	0.6	2	4	8	n/a	O(C1OC(C(O)C(O)C1O)CO)C2C(C(OC2CO)(O)CO)O
Lamivudine	87	65	2	22	n/a	O=C1N(C2OC(SC2)CO)CC=C(N1)N
Lamotrigine	98	59	1	31	n/a	Clc1c(Cl)cccc1-c2c(nc(nn2)N)N
Lansoprazole	85	61	0	120	n/a	C(F)(F)(F)COc1c(c(ncc1)CS(=O)c2nc3c(cccc3)[nH]2)C
Leucine	100	247	1	127	B(0)AT2	C(=O)(C(CC(C)C)N)O
Levodopa	86	136	2	65	LAT1	C(=O)(C(Cc1cc(c(O)cc1)O)N)O
Levonorgestrel	100	75	0	45	n/a	C(C1(C2(C(C3C(C4C(=CC(=O)CC4)CC3)CC2)CC1)CC)O)#C
Lincomycin	28	17	2	24	n/a	C(=O)(NC(C(O)C)C1OC(SC)C(O)C(O)C1O)C2N(CC(CCC)C2)C
Lisinopril	28	37	2	46	Intes. pept. carrier	C(=O)(N1C(C(=O)O)CCC1)C(NC(C(=O)O)CCc2ccccc2)CCCCN
Loracarbef	100	232	2	64	Intes. pept. carrier	C(=O)(C1=C(Cl)CCC2N1C(=O)C2NC(=O)C(c3ccccc3)N)O
Lormetazepam	100	130	0	126	n/a	O=C1N(c2c(C(c3c(Cl)cccc3)=NC1O)cc(Cl)cc2)C
Lornoxicam	100	126	0	118	n/a	C(=O)(C1=C(C2C(S(=O)(=O)N1C)C=C(Cl)S=2)O)Nc3ncccc3
Lovastatin	10	21	0	150	OATP/P-gp	C(=O)(OC1C2C(C=CC(C2CCC3OC(=O)CC(O)C3)C)=CC(C1)C)C(CC)C
Mannitol	16	9	2	15	n/a	C(C(C(O)CO)O)(C(O)CO)O
Meloxicam	90	43	0	59	n/a	C(=O)(C1=C(c2c(S(=O)(=O)N1C)cccc2)O)Nc3ncc(s3)C
Mercaptoethanesulfonic acid	77	104	2	59	Influx?	O=S(=O)(O)CCS
Metaproterenol	44	17	1	25	n/a	c1(C(O)CNC(C)C)cc(O)cc(O)c1

(*Continued*)

Table 17.1 (*Continued*)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Metformin	53	31	3	19	hOCT1/PMAT	C(N(C)C)(=N)NC(=N)N
Methadone	80	32	0	75	n/a	C(C(c1ccccc1)(c2ccccc2)CC(N(C)C)C)(=O)CC
Methotrexate	70	93	3	35	RFCP	C(=O)(c1ccc(N(Cc2nc3c(nc(nc3N)N)nc2)C)cc1)NC(C(=O)O)CCC(=O)O
Methyldopa	41	15	3	11	n/a	C(C(Cc1cc(c(O)cc1)O)(N)C)(=O)O
Methylprednisolone	82	56	1	40	n/a	C(=O)(C1(C2(C(C3C(C4(C(=CC(=O)C=C4)C(C3)C)C)C(O)C2)CC1)C)O)CO
Metolazone	64	41	0	72	n/a	O=S(=O)(c1c(Cl)cc2c(C(=O)N(c3c(cccc3)C)C(N2)C)c1)N
Metoprolol	95	90	0	102	n/a	O(c1ccc(cc1)CCOC)CC(O)CNC(C)C
Mexiletine	100	132	0	128	n/a	O(c1c(cccc1C)C)CC(N)C
Mibefradil	69	25	0	49	n/a	C(=O)(OC1(C(c2c(cc(F)cc2)CC1)C(Cl)C)CCN(CCCc3nc4c(cccc4)[nH]3)C)COC
Miconazole	25	34	0	263	P-gp	O(C(c1c(Cl)cc(Cl)cc1)Cn2cncc2)Cc3c(Cl)cc(Cl)cc3
Mifobate	82	57	0	136	n/a	O=P(OC(P(=O)(OC)OC)c1ccc(Cl)cc1)(OC)OC
Minoxidil	98	57	1	21	n/a	C1(=NC(N(C(=C1)N)O)N)N2CCCCC2
Morphine	85	37	0	36	n/a	c12c3c(O)ccc1CC4N(CCC25C(O3)C(O)C=CC54)C
Moxonidine	88	70	0	85	n/a	O(c1c(c(Cl)nc(n1)C)NC2=NCCN2)C
Nadolol	57	72	1	99	P-gp	C(NCC(O)COc1c2c(ccc1)CC(O)C(O)C2)(C)(C)C
Naloxone	91	80	0	68	n/a	O=C1C2C34c5c(c(O)ccc5CC(C3(O)CC1)N(CC=C)CC4)O2
Naproxen	99	61	0	72	n/a	C(=O)(C(c1cc2c(cc(OC)cc2)cc1)C)O
Nefazodone	100	129	0	125	n/a	O=C1N(C(=NN1CCCN2CCN(c3cc(Cl)ccc3)CC2)CC)CCOc4ccccc4
Neomycin	1	1	4	3	n/a	O(C1OC(C(OC2OC(C(O)C(O)C2N)CN)C1O)CO)C3C(OC4OC(C(O)C(O)C4N)CN)C(N)CC(C3O)N
Netivudine	28	18	2	28	n/a	C(#CC)C1C(=O)NC(=O)N(C=1)C2OC(C(O)C2O)CO
Nicotine	100	117	0	93	n/a	c1(cnccc1)C2N(CCC2)C

Nicotinic acid	88	141	1	83	Intes. active transport	C(=O)(c1cnccc1)O
Nisoldipine	90	44	0	41	n/a	C(=O)(OCC(C)C)C1=C(NC(=C(C(=O)OC)C1c2c([N+](=O)[O-])cccc2)C)C
Nitrendipine	88	68	0	71	n/a	C(=O)(OCC)C1=C(NC(=C(C(=O)OC)C1c2cc([N+](=O)[O-])ccc2)C)C
Nitrofurantoin	100	110	0	69	n/a	N(N1C(=O)NC(=O)C1)=Cc2oc([N+](=O)[O-])cc2
Nizatidine	90	150	1	71	OCT	C(=C[N+](=O)[O-])(NCCSCC1=CS=C(N1)CN(C)C)NC
Nordiazepam	99	63	0	65	n/a	O=C1Nc2c(C(c3ccccc3)=NC1)cc(Cl)cc2
Norfloxacin	71	95	1	87	P-gp	C(=O)(C1C(=O)c2c(N(C=1)CC)cc(c(F)c2)N3CCNCC3)O
Ofloxacin	100	64	1	17	n/a	C(=O)(C1C(=O)c2c3c(c(c(F)c2)N4CCN(CC4)C)OCC(N3C=1)C)O
Olanzapine	75	45	0	94	n/a	c12C(=Nc3c(cccc3)Nc1sc(c2)C)N4CCN(CC4)C
Olsalazine	24	32	1	109	n/a	C(=O)(c1c(O)ccc(N=Nc2cc(C(=O)O)c(O)cc2)c1)O
Omeprazole	80	29	0	58	n/a	O=S(c1nc2c(cc(OC)cc2)[nH]1)Cc3c(c(OC)c(cn3)C)C
Ondansetron	100	121	0	99	n/a	O=C1c2c3c(n(c2CCC1Cn4c(ncc4)C)C)cccc3
Ouabain	1.4	7	4	7	n/a	O=C1OCC(=C1)C2C3(C(O)(C4C(C5(C(O)(CC(OC6OC(C(O)C(O)C6O)C)CC5O)CC4)CO)C(O)C3)CC2)C
Oxatomide	100	125	0	115	n/a	O=C1N(c2c(cccc2)N1)CCCN3CCN(C(c4ccccc4)c5ccccc5)CC3
Oxazepam	89	41	0	66	n/a	O=C1C(N=C(c2c(ccc(Cl)c2)N1)c3ccccc3)O
Oxprenolol	95	54	0	63	n/a	O(c1c(OCC(O)CNC(C)C)cccc1)CC=C
Oxyfedrine	85	62	0	123	n/a	C(=O)(c1cc(OC)ccc1)CCNC(C(c2ccccc2)O)C
Pafenolol	29	40	1	92	n/a	C(=O)(NC(C)C)NCCc1ccc(OCC(O)CNC(C)C)cc1
Paromomycin	3	11	4	2	n/a	O(C1OC(C(OC2OC(C(O)C(O)C2N)CN)C1O)CO)C3C(OC4OC(C(O)C(O)C4N)CO)C(N)CC(C3O)N
Pefloxacin	95	87	0	65	n/a	C(=O)(C1C(=O)c2c(N(C=1)CC)cc(c(F)c2)N3CCN(CC3)C)O
Penicillin V	30	20	1	51	n/a	C(=O)(NC1C(=O)N2C(C(=O)O)C(SC21)(C)C)COc3ccccc3
Phenglutarimide	100	119	0	97	n/a	N(CCC1(C(=O)NC(=O)CC1)c2ccccc2)(CC)CC
Phenoxymethylpenicillin	59	77	1	106	Intes. pept. carrier	C(=O)(NC1C(=O)N2C(C(=O)O)C(SC21)(C)C)COc3ccccc3

(*Continued*)

Table 17.1 (*Continued*)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Phenytoin	90	75	0	88	n/a	O=C1C(c2ccccc2)(c3ccccc3)NC(=O)N1
Pindolol	87	38	0	44	n/a	O(c1c2c(N=CC2)ccc1)CC(O)CNC(C)C
Pirbuterol	60	35	1	47	n/a	C(NCC(c1nc(c(O)cc1)CO)O)(C)(C)C
Pirenzepine	27	35	0	135	P-gp	C(=O)(N1c2c(cccn2)NC(=O)c3c1cccc3)CN4CCN(CC4)C
Piroxicam	100	68	0	56	n/a	C(=O)(C1=C(c2c(S(=O)(=O)N1C)cccc2)O)Nc3ncccc3
Piroximone	81	55	0	76	n/a	C(=O)(C1C(=NC(=O)N=1)CC)c2ccncc2
Practolol	95	88	0	75	n/a	C(=O)(Nc1ccc(OCC(O)CNC(C)C)cc1)C
Pravastatin	34	47	2	47	OATP	C(=O)(OC1C2C(C=CC(C2CCC(O)CC(O)CC(=O)O)C)=CC(O)C1)C(CC)C
Praziquantel	100	111	0	80	n/a	C(=O)(N1CC(=O)N2C(c3c(cccc3)CC2)C1)C4CCCCC4
Prazosin	86	64	0	74	n/a	C(=O)(c1occc1)N2CCN(c3nc4c(c(n3)N)cc(OC)c(OC)c4)CC2
Prednisolone	99	60	1	20	n/a	C(=O)(C1(C2(C(C3C(C4(C(=CC(=O)C=C4)CC3)C)C(O)C2)CC1)C)O)CO
Prefloxacin	100	108	0	66	n/a	C(=O)(C1C(=O)c2c(N(C=1)CC)cc(c(F)c2)N3CCN(CC3)C)O
Probenecid	100	128	0	122	n/a	C(=O)(c1ccc(S(N(CCC)CCC)(=O)=O)cc1)O
Progesterone	100	225	0	168	P-gp	C(=O)(C1C2(C(C3C(C4(C(=CC(=O)CC4)CC3)C)CC2)CC1)C)C
Propiverine	84	60	0	105	n/a	C(C(OCCC)(c1ccccc1)c2ccccc2)(=O)OC3CCN(CC3)C
Propranolol	99	62	0	67	n/a	O(c1c2c(ccc1)cccc2)CC(O)CNC(C)C
Propylthiouracil	76	46	0	78	n/a	c1(nc(O)cc(n1)CCC)S
Quinidine	81	120	0	184	P-gp	O(c1cc2c(nccc2C(O)C3N4CC(C=C)C(C3)CC4)cc1)C
Raffinose	0.3	1	4	5	n/a	O(C1(OC(C(O)C1O)CO)CO)C2OC(C(O)C(O)C2O)COC3OC(C(O)C(O)C3O)CO
Ranitidine	64	86	0	133	P-gp	C(=C[N+](=O)[O-])(NCCSCc1oc(cc1)CN(C)C)NC
Recainam	71	44	0	109	n/a	C(=O)(Nc1c(cccc1C)C)NCCCNC(C)C
Remikiren	10	22	1	113	n/a	C(=O)(C(NC(=O)C(CS(C(C)(C)C)(=O)=O)Cc1ccccc1)Cc2cnc[nH]2)NC(C(C(O)C3CC3)O)CC4CCCCC4

Reproterol	60	36	1	52	n/a	O=C1c2c(ncn2CCCNCC(c3cc(O)cc(O)c3)O)N(C(=O)N1C)C
Ribavirin	33	12	3	10	n/a	C(=O)(C1=NCN(C2OC(C(O)C2O)CO)N1)N
Rifabutine	53	30	3	16	n/a	C(=O)(OC1C(C(OC)C=COC2(C(=O)c3c4c(C(=O)C (=C5C4=NC6(N5)CCN(CC(C)C)CC6)NC(=O)C(=CC=CC(C(O) C(C(O)C1C)C)C)C)c(c(c3O2)C)O)C)C)C
Rimiterol	48	19	1	29	n/a	c1(c(O)ccc(C(O)C2NCCCC2)c1)O
Saccharin	88	67	2	25	n/a	O=C1c2c(S(=O)(=O)N1)cccc2
Salicylic acid	100	120	1	56	n/a	C(=O)(c1c(O)cccc1)O
Saquinavir	80	48	2	27	n/a	C(=O)(c1nc2c(cc1)cccc2)NC(C(=O)NC(C(O)CN3C(C(=O)NC(C) (C)C)CC4C(C3)CCCC4)Cc5ccccc5)CC(=O)N
Scopolamine	95	82	1	36	n/a	C(=O)(OC1CC2N(C(C3OC32)C1)C)C(c4ccccc4)CO
Sorivudine	82	123	2	56	Nucleoside?	O=C1C(C=CBr)=CN(C(=O)N1)C2OC(C(O)C2O)CO
Sotalol	95	84	1	38	n/a	O=S(=O)(Nc1ccc(C(O)CNC(C)C)cc1)C
Spironolactone	73	26	0	32	n/a	C(=O)(SC1C2C(C3(C(=CC(=O)CC3)C1)C)CCC4(C5(CC(=O) OC5)CCC42)C)C
Stavudine	100	66	0	40	n/a	O=C1C(=CN(C(=O)N1)C2OC(C=C2)CO)C
Streptomycin	1	3	4	1	n/a	C(=N)(NC1C(OC2OC(C(C=O)(O)C2OC3OC(C(O)C(O)C3NC) CO)C)C(O)C(O)C(NC(=N)N)C1O)N
Sudoxicam	100	122	0	104	n/a	C(=O)(C1=C(c2c(S(=O)(=O)N1C)cccc2)O)Nc3nccs3
Sulfamethoxazole	100	107	0	63	n/a	O=S(=O)(c1ccc(cc1)N)Nc2noc(c2)C
Sulfasalazine	59	33	1	43	n/a	C(=O)(c1c(O)ccc(N=Nc2ccc(S(=O)(=O)Nc3ncccc3)cc2)c1)O
Sulindac	90	47	0	64	n/a	C(=O)(O)CC1c2c(C(C=1C)=Cc3ccc(S(=O)C)cc3)ccc(F)c2
Sulpiride	44	16	0	38	n/a	C(=O)(c1c(OC)ccc(S(=O)(=O)N)c1)NCC2N(CC)CCC2
Sultopride	89	72	0	98	n/a	C(=O)(c1c(OC)ccc(S(=O)(=O)CC)c1)NCC2N(CC)CCC2
Sumatriptan	57	21	0	39	n/a	N(CCC1c2c(N=C1)ccc(c2)CS(=O)(=O)NC)(C)C
Telmisartan	90	48	0	70	n/a	C(=O)(c1c(-c2ccc(cc2)Cn3c4c(c(cc(-c5nc6c(n5C)cccc6)c4)C) nc3CCC)cccc1)O
Tenidap	89	71	0	73	n/a	C(=O)(N1C(=O)C(=C(c2sccc2)O)c3c1ccc(Cl)c3)N

(*Continued*)

Table 17.1 (*Continued*)

Name	HIA%	F_a rank	ADMET Risk	AR rank	Active transport	Smiles
Tenoxicam	100	65	0	60	n/a	C(=O)(C1=C(C2C(S(=O)(=O)N1C)C=CS=2)O)Nc3ncccc3
Terazosin	90	73	0	70	n/a	C(=O)(N1CCN(c2nc3c(c(n2)N)cc(OC)c(OC)c3)CC1)C4OCCC4
Terbinafine	80	53	0	134	n/a	C(#CC=CCN(Cc1c2c(ccc1)cccc2)C)C(C)(C)C
Terbutaline	62	40	1	44	n/a	C(NCC(c1cc(O)cc(O)c1)O)(C)(C)C
Testosterone	100	112	0	81	n/a	O=C1C=C2C(C3C(C4C(C(O)CC4)(CC3)C)CC2)(CC1)C
Theophylline	100	238	0	197	Active uptake?	O=C1c2c(nc[nH]2)N(C(=O)N1C)C
Thiacetazone	20	10	1	27	n/a	C(=O)(Nc1ccc(C=NNC(=S)N)cc1)C
Tiagabine	90	155	0	170	n/a	C(c1c(ccs1)C)(c2c(ccs2)C)=CCCN3CC(C(=O)O)CCC3
Timolol	95	83	1	37	n/a	C(NCC(O)COc1c(nsn1)N2CCOCC2)(C)(C)C
Tinidazole	100	114	0	86	n/a	O=S(=O)(CCn1c(ncc1[N+](=O)[O-])C)CC
Tolbutamide	85	36	0	33	n/a	C(=O)(NS(=O)(=O)c1ccc(cc1)C)NCCCC
Tolmesoxide	98	102	0	107	n/a	O=S(c1c(cc(OC)c(OC)c1)C)C
Topiramate	86	63	1	32	n/a	O=S(=O)(OCC12OC(OC1C3OC(OC3CO2)(C)C)(C)C)N
Torasemide	96	94	1	31	n/a	C(=O)(NS(=O)(=O)c1c(Nc2cc(ccc2)C)ccnc1)NC(C)C
Toremifene	100	137	0	137	n/a	C(=C(c1ccccc1)CCCl)(c2ccc(OCCN(C)C)cc2)c3ccccc3
Tramadol	90	77	0	106	n/a	N(CC1C(c2cc(OC)ccc2)(O)CCCC1)(C)C
Tranexamic acid	55	20	1	24	n/a	C(=O)(O)C1CCC(CN)CC1
Trapidil	96	95	0	101	n/a	N(C1n2c(N=C(C=1)C)ncn2)(CC)CC
Trimethoprim	97	191	1	112	RFCP?	O(c1c(OC)cc(cc1OC)Cc2c(nc(nc2)N)N)C
Trovofloxicin	37	52	0	132	MDR1 in bacteria	C(=O)(C1C(=O)c2c(nc(c(F)c2)N3CC4C(C4C3)N)N(c5c(F)cc(F)cc5)C=1)O
Urapidil	78	27	0	47	n/a	O=C1N(C(=O)C=C(N1C)NCCCN2CCN(c3c(OC)cccc3)CC2)C
Valproic acid	100	70	0	68	n/a	C(=O)(C(CCC)CCC)O
Venlafaxine	97	56	0	55	n/a	N(CC(c1ccc(OC)cc1)C2(O)CCCCC2)(C)C
Verapamil	100	258	0	250	P-gp	C(C(c1cc(OC)c(OC)cc1)(C(C)C)CCCN(CCc2cc(OC)c(OC)cc2)C)#N

Vigabatrin	58	32	1	50	n/a	C(=O)(O)CCC(C=C)N
Viloxazine	98	100	0	90	n/a	O(c1c(OCC)cccc1)CC2OCCNC2
Viomycin	85	35	4	2	n/a	C(=O)(NC1C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC1)C2NC(=N)NC(O)C2)=CNC(=O)N)CO)CO)CC(CCCN)N
Warfarin	98	101	0	103	n/a	C(=O)(CC(C1C(=O)Oc2c(C=1O)cccc2)c3ccccc3)C
Xamoterol	5	6	1	23	n/a	C(=O)(N1CCOCC1)NCCNCC(O)COc2ccc(O)cc2
Ximoprofen	98	103	0	108	n/a	C(=O)(C(c1ccc(cc1)C2CC(=NO)CCC2)C)O
Xipamide	70	43	1	55	n/a	C(=O)(c1c(O)cc(Cl)c(S(=O)(=O)N)c1)Nc2c(cccc2C)C
Zidovudine	100	214	0	131	Nucleoside	N(=[N+]=[N-])C1C(OC(N2C(=O)NC(=O)C(=C2)C)C1)CO
Ziprasidone	60	37	0	84	n/a	O=C1Nc2c(cc(c(Cl)c2)CCN3CCN(c4c5c(sn4)cccc5)CC3)C1
Zopiclone	80	49	1	39	n/a	C(=O)(OC1c2c(C(=O)N1c3ncc(Cl)cc3)nccn2)N4CCN(CC4)C

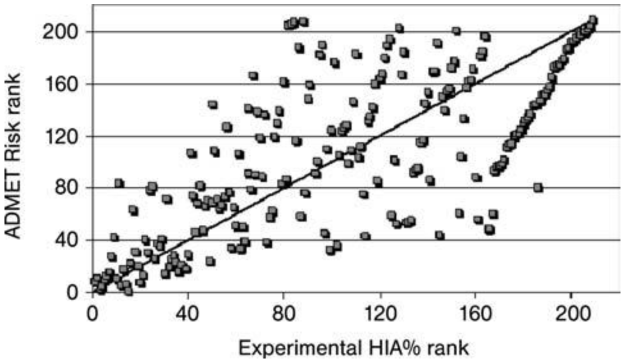

Figure 17.1 Correlation of rank order for ADMET Risk and human intestinal absorption (HIA%). The ADMET Risk score ranged from 0 to 5, with 5 being compounds with the greatest risk of having poor ADMET properties. Within a single ADMET Risk number, the compounds were ranked according to ascending estimated human jejunal permeability (ADMET Predictor, Simulations Plus, Inc.). Spearman rank correlation coefficient was 0.7 ($p < 0.001$).

and acceptors, and human jejunal permeability from 3D molecular structures. The predictive performance of Lipinski's original rule-of-5 [47] was compared with that of the ADMET Risk. A positive was defined as HIA% greater than or equal to 50%, and a negative was defined as less than or equal to 50%.

Both sets of rules correctly identified over 99% of the true positives. However, because of the liberal criteria found in Lipinski's rule-of-5, only 20% true negatives were predicted correctly whereas 80% were predicted as false positives. The default ADMET Risk rules predicted 64% of the true negatives and lowered the false positives to only 36%. Unlike the original rule-of-5, however, ADMET Risk marked one

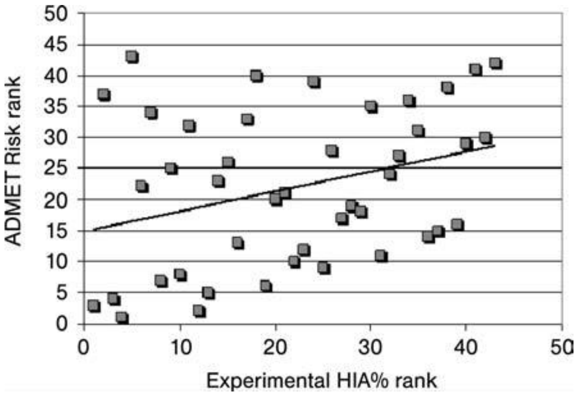

Figure 17.2 Correlation of rank order for 43 compounds that are known substrates for influx or efflux transporters. Spearman rank correlation coefficient was 0.3.

compound as a false negative – a result of more conservative rules. Rifabutine has $F_a = 53\%$ and was assigned an ADMET Risk score of 3, so it was right on the border of having poor ADMET properties. Thus, application of ultrahigh-throughput *in silico* estimation of biopharmaceutical properties to the generation of computational alerts has the potential to improve compound selection to those drug candidates that are likely to have less trouble in development.

17.3.2
Complex Rules That Include Toxicity

Computational alerts can be extended from having rules that predict absorption properties only to customized rules that include distribution and toxicity. Here, we will consider *in silico* models for four types of toxicity. When the "hits" in the ADME rules are combined with additional rules for estimated toxicity, we generate a more general computational "ADMET Risk."

Maximal recommended therapeutic dose (MRTD as defined by the US FDA) has been correlated with potential toxicity. Matthews *et al.* have reported *in silico* multicase models for MRTD [55]. According to their analysis, more toxic molecules will have MRTD values less than 2.7 mg/kg/day and less toxic molecules will have MRTD more than 4.99 mg/kg/day. If one builds an *in silico* model for MRTD, then a new rule can be added to the computational alert ADME rules described above.

The US Environmental Protection Agency (EPA) has released a number of toxicological databases on its Distributed Structure-Searchable Toxicity (DSSTox) database network. These databases make a tremendous resource for building *in silico* models of toxicity. The first toxicity we will consider from the EPA is for molecules that bind to the estrogen receptor and have the potential to produce endocrine disruption. The primary quantitative end point for this toxicity is a ratio of the estimated estrogen receptor-binding affinities for 17b-estradiol divided by the affinity estimated for the unknown molecule (RBA). A high value for the ratio ($RBA > 1$) would imply that the new molecule has a binding affinity for the estrogen receptor greater than estradiol. The EPA database contains 232 molecules with measured RBA values, and the median ratio is 0.02. Thus, half of the molecules would have ratios greater than 0.02 (more toxic), and half would have ratios less than 0.02 (less toxic). When a virtual screen is conducted on a set of new molecules, this RBA estimation would be a valuable addition to the overall ADMET Risk score.

When considering environmental chemicals, another important measure of toxicity is based on the EPA acute fathead minnow toxicity database. In this assay, 28–36-day-old fathead minnows are exposed to varying concentrations of a test molecule in a flow-through apparatus for 96 h [56]. The concentration of the organic chemical that produced 50% lethality (LD_{50}) was reported in the EPA database. The median LD_{50} from 586 molecules was 21.5 mg/l. This value could comprise another cutoff in a computational alert.

Rat carcinogenicity is another important measure of toxicity reported in one of the DSSTox databases. TD_{50} is the daily dose that will induce tumors in half of the test animals that would have remained tumor free at zero dose [57]. The median TD_{50}

from 265 compounds tested was 1.15 mg/kg/day. In screening for ADMET Risk, thus, one might be able to estimate if the new molecules were more or less carcinogenic than the most carcinogenic half of the molecules tested by the EPA.

The DSSTox web site also provides a database for *Salmonella* mutagenicity with 506 molecules approximately half of which are designated as positive for mutagenicity [57]. In addition, Simulations Plus, Inc., has a database with approximately 5000 molecules that covers 10 strains of bacterial mutagenesis.

Finally, we can consider the inhibition of human ether-a-go-go related gene (hERG) product, which encodes a voltage-gated potassium channel in the cardiac myocyte. Inhibition of the channel predisposes patients to long QT syndrome, the characteristic "Torsades de Pointes" arrhythmia, and sudden cardiac death. Several publications have reported the results of electrophysiological measurement of the hERG IC_{50} for a variety of drugs [58, 59]. In building a model for hERG IC_{50}, it is important to screen the data for a common target, preferably the expression of hERG on mammalian cells (HEK, CHO, COS, cardiac myocytes, and neuroblastoma cells). For 93 drugs, the median half-maximal inhibition of potassium channel current tested in patch clamp electrophysiological apparatus on mammalian cells expressing human ERG gene was 1.58 μM. This represents a reasonable cutoff for hERG toxicity.

Thus, an extension of the ADMET Risk rules to include toxicity would include

- MRTD values less than 2.7 mg/kg/day;
- RBA ratios more than 0.02;
- fathead minnow LD_{50} less than 21.5 mg/l;
- Positive prediction for *Salmonela* mutagenicity;
- hERG IC_{50} less than 1.58 μM.

17.4 Mechanistic Simulation (ACAT Models) in Early Discovery

We have developed a two-step procedure for the *in silico* screening of compound libraries based on biopharmaceutical property estimation linked to a mechanistic simulation of GI absorption. The first step involves biopharmaceutical property estimation by application of machine learning procedures to empirical data modeled with a set of molecular descriptors derived from 2D and 3D molecular structures. *In silico* methods were used to estimate such biopharmaceutical properties as effective human jejunal permeability, cell culture permeability, aqueous solubility, and molecular diffusivity. In the second step, differential equations for the advanced compartmental absorption and transit model were numerically integrated to determine the rate, extent, and approximate GI location of drug liberation (for controlled release), dissolution, and absorption. Figure 17.3 shows the schematic diagram of the ACAT model in which each one of the arrows represents an ordinary differential equation (ODE).

The form of the ACAT model implemented in GastroPlus describes the release, dissolution, luminal degradation (if any), metabolism, and absorption/exsorption of

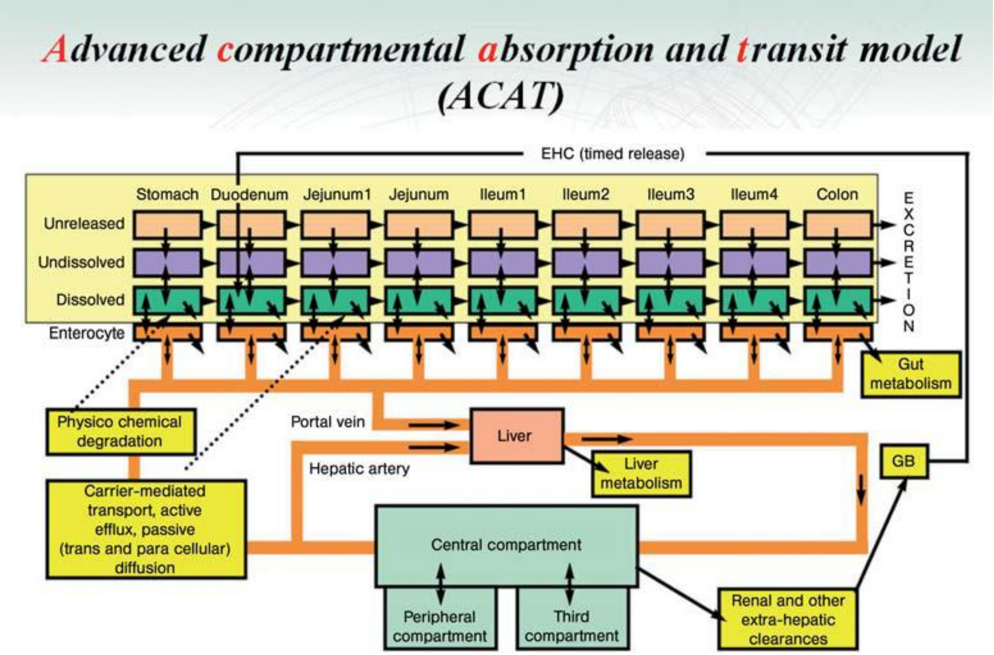

Figure 17.3 Schematic diagram of the advanced compartmental absorption and transit model as implemented in GastroPlus.

a drug as it transits through successive compartments. The kinetics associated with these processes is modeled by a system of coupled linear and nonlinear rate equations. The equations include the consideration of 6 states (unreleased, undissolved, dissolved, degraded, metabolized, and absorbed), 18 compartments (9 GI – 1 stomach, 6 small intestine, and 2 colon– and 9 enterocyte), 3 states of excreted material (unreleased, undissolved, and dissolved), and the amount of drug in up to 3 pharmacokinetic compartments (when pharmacokinetic parameters are available) or in a whole body physiologically-based pharmacokinetic model. The total amount of absorbed material is summed over the integrated amounts being absorbed/exsorbed from each absorption/transit compartment.

For example, the rate of change of dissolved drug concentration in a luminal GI compartment depends on six different processes: (1) transit of drug into a compartment, (2) transit of drug out of a compartment, (3) release of drug from the formulation in the compartment, (4) dissolution/precipitation of drug particles, (5) luminal degradation of the drug, and (6) absorption/exsorption of the drug . The timescale associated with luminal transit through a compartment is determined by a transfer rate constant, k_t, that is calculated as one divided by the mean transit time within the compartment. Transit times within each compartment are determined as the product of the physical volume of fluid in the compartment (milliliter) divided by the average fluid flow rate (ml/h). The timescale of the dissolution process is set by a rate constant, k_d, that is computed from a drug's solubility (as a function of pH), its effective particle size, particle density, lumen concentration, diffusion coefficient, and the diffusion layer thickness (Equation 17.1). The timescale associated with the absorption process is set by a rate coefficient, k_a, that depends on the effective permeability of the drug (P_{eff}, units of cm/s) multiplied by an absorption scale factor (ASF with units of cm^{-1}) for each compartment (Equation 17.2). The nominal value of the ASF is the surface-to-volume ratio of the compartment, which reduces to 2/radius of the SI compartment. ASFs are adjusted from these nominal values to correct for the changes in permeability due to changing physiology along the GI tract; for example, absorption surface area, pH, tight junction gap width, and transport protein (influx or efflux) densities. The rates of absorption and exsorption depend on the concentration gradients across the apical and basolateral enterocyte membranes (Equations 17.3 and 17.4). The timescale for luminal degradation is set by a rate constant k_{Degrad} that is determined by interpolation from an input table of degradation rate (or half-life) versus pH and the pH in the compartment.

The system of differential equations is integrated using CVODE numerical integration package. CVODE is a solver for stiff and nonstiff ordinary differential equation systems [60]. The fraction of dose absorbed is calculated as the sum of all drug amounts crossing the apical membrane as a function of time, divided by the dose, or by the sum of all doses if multiple dosing is used.

$$k_{(i)\text{d}} = 3\gamma \frac{C_{\text{S}(i)} - C_{(i)\text{L}}}{\rho r_0 T}, \tag{17.1}$$

$$k'_{(i)\text{a}} = \alpha_{(i)} P_{\text{eff}\,(i)}, \tag{17.2}$$

$$\frac{\text{Absorption}}{\text{Exsorption}}_{(i)} = k'_{(i)\text{a}} V_{(i)} (C_{(i)\text{L}} - C_{(i)\text{E}}), \tag{17.3}$$

$$\text{Basolateral transfer}_{(i)} = k'_{(i)\text{b}} V_{(i)} (C_{(i)\text{E}} - C_{\text{p}}), \tag{17.4}$$

where $k_{(i)\text{d}}$ is dissolution rate constant for the ith compartment; $k'_{(i)\text{a}}$ is absorption rate coefficient for the ith compartment; $k'_{(i)\text{b}}$ is absorption rate coefficient specific for the basolateral membrane of the ith compartment; C_{S} is aqueous solubility at local pH; $C_{(i)\text{L}}$ is lumen concentration for the ith compartment; $C_{(i)\text{E}}$ is intracellular enterocyte concentration for the ith compartment; C_{p} is plasma central compartment concentration; $V_{(i)}$ is lumen volume of ith compartment; γ is molecular diffusion coefficient; ρ is drug particle density; r_0 is effective initial drug particle radius; T is diffusion layer thickness; $\alpha_{(i)}$ is compartmental absorption scale factor for ith compartment; and $P_{\text{eff}(i)}$ is human effective permeability for ith compartment.

As one part of our software validation, we tested the accuracy of GastroPlus simulation of fraction absorbed. Starting from two-dimensional structures, ADMET Predictor was used to generate the molecular descriptors and estimates of log P, solubility, permeability, and diffusivity that were used in GI simulations. The extent of GI absorption for each drug was determined *in silico* using the ACAT model after making the following simplifying assumptions: (default dose (100 mg), particle radius (was adjusted to achieve 100% dissolution in 3.3 h), and human fasted physiology). The simulation results from GastroPlus were compared with literature values. The simplest assumption for the regional dependence of the rate of absorption (Equation 17.2) is that the compartmental absorption scale factor is equal to 2 divided by the radius of the small intestine and that this value of ASF is applied to all compounds equally.

17.4.1 Automatic Scaling of k'_a as a Function of P_{eff}, pH, log D, and GI Surface Area

The size and shape of a drug molecule, its acid and base dissociation constants, and the pH of the GI tract all influence the absorption rate constant for specific regions of the GI. Pade and coworkers measured the Caco-2 cellular permeability for a diverse set of acidic and basic drug molecules at two pH values [61]. They concluded that the permeability coefficient of the acidic drugs was greater at pH 5.4, whereas that of the basic drugs was greater at pH 7.2 and that the transcellular pathway was the favored pathway for most drugs, probably due to its larger accessible surface area. The paracellular permeability of the drugs depended on size and charge. The permeability of the drugs through the tight junctions decreased with increasing molecular size. Further, the pathway also appeared to be cation selective, with the positively charged cations of weak bases permeating the aqueous pores of the paracellular pathway at a faster rate than the negatively charged anions of weak acids. Thus, the extent to which the paracellular and transcellular routes are utilized in drug transport is influenced by the fraction of ionized and unionized species (which in turn depends on the pK_a of the drug and the pH of the solution), the intrinsic partition coefficient of the drug, and molecule size and charge.

Figure 17.4 is a representation of regional permeability coefficients of 19 drugs with different physicochemical properties determined by Ungell *et al.* by using excised segments from three regions of rat intestine: jejunum, ileum, and colon [62].

They observed a significant decrease in the permeability to hydrophilic drugs and a significant increase in the permeability for hydrophobic drugs aborally to the small intestine ($p < 0.0001$). Figure 17.4 illustrates that for hydrophilic drugs (low permeability and low log *D*), the ratio of colon: jejunal permeability was less than 1, whereas for hydrophobic drugs (higher permeability and higher log *D*), the ratio of colon: jejunal permeability was observed to be greater than 1. At certain pH values, the permeability of small hydrophilic drugs may have a large paracellular component [63], and it is well known that the transepithelial electrical resistance (TEER) of colon is much higher than that of the small intestine. TEER increases as the width of tight junctions decreases, and the tight junction width has been determined to be 0.75–0.8 nm in jejunum, 0.3–0.35 nm in ileum, and 0.2–0.25 nm in colon [64–67]. The narrower tight junctions in colon suggest that the paracellular transport will be much less significant in the colon, which helps to explain the lower ratio of colon:

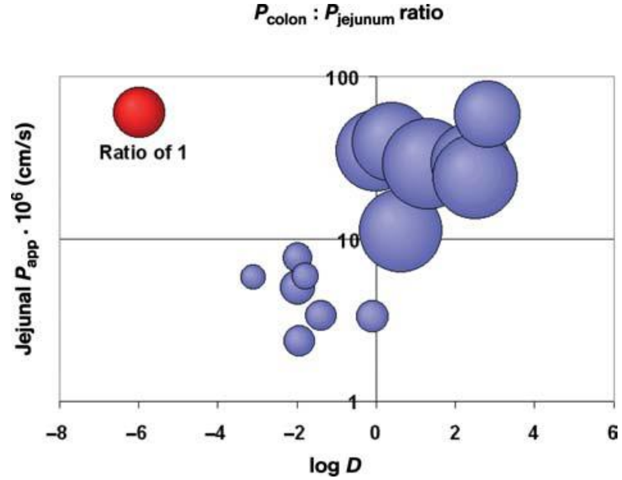

Figure 17.4 Relationship between distribution coefficient at pH = 7.4 and the intestinal permeability of jejunum, ileum, and colon. Data were collected from [62].

jejunal permeability for hydrophilic drugs. To our knowledge, a conclusive explanation for the increased colon permeability of drugs with high small intestine permeability is not yet available. We have used the ACAT model with experimental biopharmaceutical properties for a series of hydrophilic and hydrophobic drug molecules to calibrate a "log D model" that explains the observed rate and extent of absorption.

The mechanistic simulation ACAT model was modified to automatically account for the change in small intestinal and colon k'_a as a function of the local (pH-dependent) log D of the drug molecule. The rank order of HIA% from GastroPlus was directly compared with the rank order experimental HIA% with this correction for the log D of each molecule in each of the pH environments of the small intestine. The mechanistic simulation produced 82% of HIA% predictions within 25% of the experimental values.

17.4.2
Mechanistic Corrections for Active Transport and Efflux

Table 17.2 lists the 43 molecules used in this study that are known to be substrates for active transport or active efflux. The mechanistic ACAT model was modified to accommodate saturable uptake and efflux by using standard Michaelis–Menten equations. It was assumed that transporters responsible for active uptake of drug molecules from the lumen and active efflux from the enterocytes to the lumen were homogeneously dispersed within each luminal compartment and each corresponding enterocyte compartment, respectively. Equation 17.5 represents the

Table 17.2 Forty-three drugs with some evidence of active uptake or efflux.

Name	Influx/efflux	Reference	Name	Influx/efflux	References
Adriamycin	P-gp	[111]	Lovastatin	OATP/P-gp	[112–114]
Allopurinol	Hypoxanthine	[115]	Mercaptoethanesulfonic acid	Influx?	[116]
Alpha-difluoromethylornithine	CAT1?	[117]	Metformin	hOCT1/PMAT	[118]
Amoxicillin	PepT1 conc. dep.	[119]	Methotrexate	RFCP	[120]
Ampicillin	PepT1/2	[121]	Miconazole	P-gp	[122, 123]
Benzylpenicillin	Intes. pept. carrier	[119]	Nadolol	P-gp	[124]
Bumetanide	Bile acid	[125]	Nicotinic acid	Intes. active trans.	[126]
Captopril	Intes. pept. carrier	[127]	Nizatidine	OCT	[128]
Carfecillin	Intes. pept. carrier	[119]	Norfloxacin	P-gp	[129]
Cefadroxil	Intes. pept. carrier	[119]	Phenoxymethylpenicillin	Intes. pept. carrier	[130]
Cefatrizine	Intes. pept. carrier	[119]	Pirenzepine	P-gp	[131, 132]
Cephalexin	PepT1	[133]	Pravastatin	OATP	[134]
Cycloserine	hPAT1	[135, 136]	Progesterone	P-gp	[137]
Digoxin	OATP/P-gp	[138, 139]	Quinidine	P-gp	[140]
Enalapril	Intes. pept. carrier	[141]	Ranitidine	P-gp	[142]
Etoposide	P-gp	[140]	Sorivudine	Nucleoside?	Structural analogy to zidovudine
Gabapentin	LAT1	[143]	Theophylline	Active uptake?	[144]
Glycine	hPAT1	[145, 146]	Trimethoprim	RFCP?	[147, 148]
Leucine	B(0)AT2	[149]	Trovofloxicin	MDR1 in bacteria	[150]
Levodopa	LAT1	[151]	Verapamil	P-gp	[152]
Lisinopril	Intes. pept. carrier	[130]	Zidovudine	Nucleoside	[153]
Loracarbef	Intes. pept. carrier	[154, 155]			

overall mass balance for drug in the enterocyte compartment lining the intestinal wall:

$$\frac{\mathrm{d}M_{\mathrm{ent}(i)}}{\mathrm{d}t} = \mathrm{ADR}_{(i)} + \mathrm{ATR}_{(i)} - \mathrm{BDR}_{(i)} - \mathrm{GMR}_{(i)}, \tag{17.5}$$

$$\mathrm{ATR} = \mathrm{DF}_{\mathrm{influx}(i)} \frac{V_{\mathrm{max(influx)}} C_{\mathrm{i}}}{K_{\mathrm{m(influx)}} + C_{\mathrm{i}}} - \mathrm{DF}_{\mathrm{efflux}(i)} \frac{V_{\mathrm{max(efflux)}} C_{\mathrm{ent}(i)}}{K_{\mathrm{m(efflux)}} + C_{\mathrm{ent}(i)}}, \tag{17.6}$$

where $M_{\mathrm{ent}(i)}$ is the mass of drug in the enterocyte compartment i; $\mathrm{ADR}_{(i)}$ is the apical diffusion rate for ith compartment; $\mathrm{ATR}_{(i)}$ is the apical transport rate for ith compartment; $\mathrm{BDR}_{(i)}$ is the basolateral diffusion rate for ith compartment; $\mathrm{GMR}_{(i)}$ is the gut metabolism rate for ith compartment; $\mathrm{DF}_{\mathrm{influx}(i)}$ is distribution factor for influx transporter in compartment i; $\mathrm{DF}_{\mathrm{efflux}(i)}$ is distribution factor for efflux transporter in compartment i; $V_{\mathrm{max(influx\ or\ efflux)}}$ is the maximal velocity of the saturable transporter; $K_{\mathrm{m(influx}\ \mathit{or}\ \mathrm{efflux)}}$ is the Michaelis constant for the saturable transporter; C_{i} is the concentration of drug inside the lumen of the intestine; $C_{\mathrm{ent}(i)}$ is the concentration of drug inside the enterocyte in compartment i.

Because the amounts and density of these transporters vary along the GI tract, it is necessary to introduce a correction factor for the varying transport rates in the different luminal and enterocyte compartments. Owing to the lack of experimental data for the regional distribution and Michaelis–Menten constants for each drug in Table 17.2, we fitted an intrinsic (concentration-independent) transport rate for each drug to closely approximate the experimental HIA%. Figure 17.5 shows a correlation

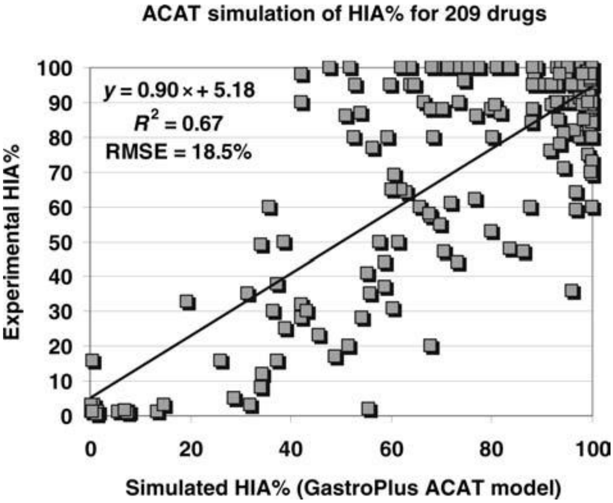

Figure 17.5 Correlation of experimental and simulated percentage absorbed. Percentage absorbed is defined as the percentage of the dose that crosses the apical membrane of the intestine. Percentage absorbed was simulated using the ACAT model as described in the text.

between the simulated HIA% and the experimental HIA% for 209 passively absorbed compounds.

17.4.3 PBPK and *In Silico* Estimation of Distribution

Prior to 2002, most studies published on physiologically-based pharmacokinetic models focused on the distribution and elimination of environmental toxins such as dioxin, styrene, and organic solvents [68–70]. PBPK models for drug molecules generally relied on tissue/plasma partition coefficients (K_{ps}) measured in rat [71–73]. The earliest models for calculation of tissue/plasma partition coefficients from structure were typical QSAR models developed from a database of empirical K_p values [74]. However, several years back, one group had already started a revolution in PBPK modeling with the introduction of a mechanistic model for tissue/plasma partition coefficients based on the tissue composition of neutral and phospholipids and volume fraction of water [75]. Years after the introduction of the tissue composition method, a comparison of methods for calculating K_p values observed that the mechanistic tissue composition methods worked well to estimate K_p values for volatile organic molecules [76]. However, the widespread use of these models for calculating drug distribution achieved great momentum from the work of Poulin and Theil [77–80]. Recent reviews and validation studies have confirmed the utility of PBPK modeling in early drug discovery [8, 10, 11, 81].

To address the inaccuracy of estimates for volume at steady state (V_{ss}) for strongly basic drug molecules, Rodgers and Rowland have proposed an extension to the tissue composition method that takes into account the volume fraction of acidic phospholipids. Presumably, the higher values of V_{ss} observed for these basic molecules is due to the interaction of the cationic state (at physiological pH) of the base with the anionic state of the acidic phospholipids [82–85]. Several commercial software programs are now extensively used in the pharmaceutical industry for PBPK modeling [86, 41, 87, 11].

17.5 Mechanistic Simulation of Bioavailability (Drug Development)

In addition to the mechanistic simulation of absorptive and secretive saturable carrier-mediated transport, we have developed a model of saturable metabolism for the gut and liver that simulates nonlinear responses in drug bioavailability and pharmacokinetics [44]. Hepatic extraction is modeled using a modified venous equilibrium model that is applicable under transient and nonlinear conditions. For drugs undergoing gut metabolism by the same enzymes responsible for liver metabolism (e.g., CYPs 3A4 and 2D6), gut metabolism kinetic parameters are scaled from liver metabolism parameters by scaling V_{max} by the ratios of the amounts of metabolizing enzymes in each of the intestinal enterocyte compartments relative to the liver. Significant work in identifying the distribution of CYP3A4 and CYP2D6

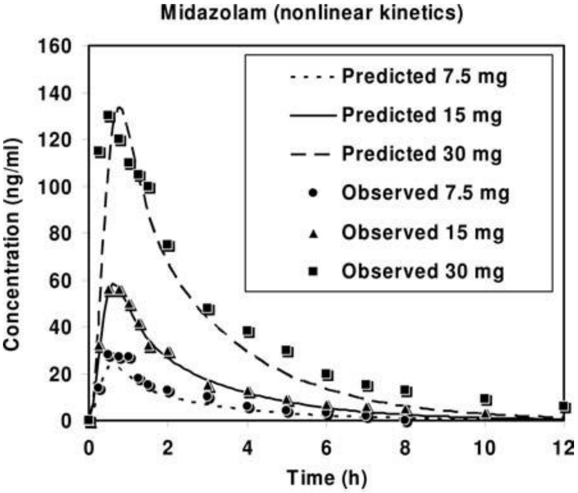

Figure 17.6 Experimental and simulated plasma concentration versus time profiles for three solution doses of midazolam. Data were collected from the literature [90].

isozymes in the gut has been done by Paine *et al.* and Madani *et al.*, respectively [88, 89], and their data were used in our simulations. We have validated the model against experimental data for drugs that undergo liver metabolism alone (propranolol and metoprolol), gut metabolism, and liver metabolism (midazolam), and efflux, gut metabolism, and liver metabolism (saquinavir).

We used *in vitro* kinetic constants obtained from homogenate or whole-cell experiments under controlled conditions and scaled the constants to the *in vivo* scenario by using appropriate physiological scale factors. Figure 17.6 shows our simulated results for absorption and metabolism of midazolam at three solution doses [90]. Midazolam is metabolized in the gut and liver by cytochrome 3A4, and Figure 17.6 shows the accurate simulation of the nonlinear dose dependence due to saturation of CYP3A4. Saquinavir is also metabolized in the gut and liver by 3A4, and it is also a substrate for efflux by P-glycoprotein. Figure 17.7 shows our simulated results for absorption and metabolism of saquinavir when dosed with and without grapefruit juice [91]. It can be seen that the simulation correctly predicts the increase in oral AUC and bioavailability when the drugs are dosed after the patient ingested grapefruit juice. It is well known that grapefruit juice is able to inhibit CYP3A4 metabolism in the gut by approximately 60% but not in the liver. Our results show that *in vitro* kinetic constants can be used to predict drug behavior *in vivo*, provided adequate data on enzyme distribution and activity are available, and that the *in vitro* method adequately measures the metabolic processes for the compound. The use of *in vitro* data from human liver microsomes, as was done for midazolam and saquinavir above, is adequate when the metabolism of the compound is well described by only phase-I processes that take place in microsomes. For compounds with significant phase-II metabolism, such as propranolol, microsomal measure-

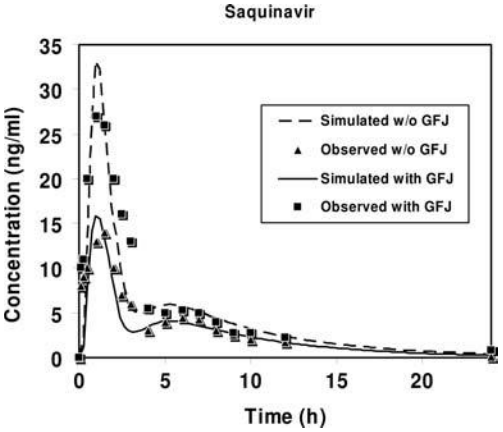

Figure 17.7 Experimental and simulated plasma concentration versus time profiles for a single dose of saquinavir administered with and without grape fruit juice [91].

ments will not reflect the entire metabolism, and clearance will be underpredicted. Data from hepatocytes can provide both phase-I and phase-II metabolism, and so the use of hepatocytes would be preferred when phase-II metabolism is involved. Even with the best of experimental data, factors such as interindividual variability in enzyme content and activity strongly limit the extension of predictions across different demographics.

More experimental information is needed regarding distribution and densities of metabolizing enzymes and efflux proteins in the GI tract. This information is crucial since dissolution and absorption are site dependent all along the GI tract. Knowledge of the variation in enzyme and efflux transporter amounts in the intestine and colon can also be used to design formulations with increased bioavailabilities by avoiding sites of high intestinal first pass and efflux. For example, the bioavailability of oxybutynin, a CYP3A4 substrate, is increased by modifying the formulation to release most of the drug in the distal GI [92]. Similarly, the bioavailability of a P-gp substrate might be increased by using a gastric-retentive formulation to release the drug in proximal GI where the P-gp density is relatively low. The influence of inhibitors and inducers of enzymes can be modeled by using appropriate scale factors to mimic changes in enzyme amounts, activity, and competitive inhibition. Similarly, drug–drug interactions can be modeled using the same techniques.

In spite of its limitations, the ACAT model combined with modeling of saturable processes has become a powerful tool in the study of oral absorption and pharmacokinetics. To our knowledge, it is the only tool that can translate *in vitro* data from early drug discovery experiments all the way to plasma concentration profiles and nonlinear dose-relationship predictions. As more experimental data become available, we believe that the model will become more comprehensive, and its predictive capabilities will be further enhanced.

17.5.1 Approaches to *In Silico* Estimation of Metabolism

In silico estimation of metabolism is still an area of intense study and development. Accurate prediction of intrinsic clearance is still not possible with the currently available methods [15]. Most of the progress in this area has been focused on the mixed function oxidase cytochrome P450 enzyme family. Advances in this area have been focused on three areas: (1) prediction of the cytochrome P450 (CYP) enzyme isotype that is responsible for the major metabolism, (2) prediction of the chemical site of a molecule that is most likely to undergo biotransformation by oxidative metabolism, and (3) structure-based docking studies of CYP enzyme substrate complexes .

Unsupervised machine learning based on the application of Kohonen self-organizing maps to groups of isotype-specific molecules has been applied to predict the CYP enzyme isotype involved in the major metabolism [93]. The same group used similar computational methods to estimate the catalytic K_m values for P450 substrates [94]. The most successful methods for predicting the P450 metabolism site utilize a method for calculation of the activation energy for homolytic cleavage of a C—H bond in the substrate [95–99]. Homolytic H atom abstraction is the rate-limiting step in P450 oxidative metabolism, and the C—H bonds with lowest activation energy are generally the sites of major metabolism for CYP enzymes such as 3A4 that has a large binding pocket that can easily accommodate the substrate in a variety of orientations. This method is less predictive for other CYP enzymes such as 2D6 or 2C9 that have a definite pharmacophore that helps orient the substrate so that oxidation can occur at carbons that have a higher activation energy for homolytic cleavage. A newer, empirical method for estimating the H-atom abstraction energy was shown to be more accurate than the classical methods based on semiempirical AM1 calculations [100, 101].

The availability of X-ray crystallographic structures and homology models of the CYP450 enzymes allows the application of structure-based methods to predict P450 metabolism [102–105]. Newer approaches that have promise in this area include hybrid methods that use an energy calculation with some knowledge of the steric interaction of a given CYP enzyme. Metasite, a software program, combines the calculation of H-atom abstraction energetics with a method based on a comparison between alignment-independent descriptors derived from GRID molecular interaction fields for the active site and a distance-based representation of the substrate [106, 107].

17.6 Regulatory Aspects of Modeling and Simulation (FDA Critical Path Initiative)

Pharmaceutical productivity has been falling and costs have been rising. In 2004, the US FDA introduced the Critical Path Initiative to modernize drug development by introducing advancements in genomics, modeling and simulation, and advanced imaging [22, 108, 109, 20, 110]. One important use of such data will be to construct quantitative models of disease processes, incorporating what is known about biomarkers, clinical outcomes, and the effects of various interventions. These models can then be used for trial simulations to better design appropriate trials and clinical outcome measures. These methods have been dubbed "Model-Based Drug Development and have the potential to improve the success rate in regulatory approval [20]."

17.7 Conclusions

The application of ultrahigh-throughput *in silico* estimation of biopharmaceutical properties to generate rule-based computational alerts has the potential to improve compound selection for those drug candidates that are likely to have less trouble in development. Extension of purely *in silico* methods to the realm of mechanistic simulation further enhances our ability to predict the impact of physiological and biochemical processes on drug absorption and bioavailability. Quantitative prediction of metabolic rates is still a future goal, but great progress has been achieved in calculating substrate specificity, sites of metabolism, and relative binding interactions with metabolic enzymes. It remains to be seen if all of these innovations combined with clinical trial simulations and model-based drug development will lead to a faster and less expensive drug development.

References

1 Modi, S. (2003) Computational approaches to the understanding of ADMET properties and problems. *Drug Discovery Today,* **8** (14), 621–623.

2 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modelling: towards prediction paradise? *Nature Reviews. Drug Discovery,* **2** (3), 192–204.

3 Stoner, C.L., Gifford, E., Stankovic, C., Lepsy, C.S., Brodfuehrer, J., Prasad, J.V. and Surendran, N. (2004) Implementation of an ADME enabling selection and visualization tool for drug discovery. *Journal of Pharmaceutical Sciences,* **93** (5), 1131–1141.

4 Yamashita, F. and Hashida, M. (2004) *In silico* approaches for predicting ADME properties of drugs. *Drug Metabolism and Pharmacokinetics,* **19** (5), 327–338.

5 Balakin, K.V., Ivanenkov, Y.A., Savchuk, N.P., Ivashchenko, A.A. and Ekins, S. (2005) Comprehensive computational assessment of ADME properties using mapping techniques. *Current Drug Discovery Technologies,* **2** (2), 99–113.

6 Lupfert, C. and Reichel, A. (2005) Development and application of physiologically based pharmacokinetic-modeling tools to support drug discovery. *Chemistry & Biodiversity,* **2** (11), 1462–1486.

7 Cai, H., Stoner, C., Reddy, A., Freiwald, S., Smith, D., Winters, R., Stankovic, C. and Surendran, N. (2006) Evaluation of an integrated *in vitro–in silico* PBPK (physiologically based pharmacokinetic) model to provide estimates of human bioavailability. *International Journal of Pharmaceutics*, **308** (1–2), 133–139.

8 Jones, H.M., Parrott, N., Jorga, K. and Lave, T. (2006) A novel strategy for physiologically based predictions of human pharmacokinetics. *Clinical Pharmacokinetics*, **45** (5), 511–542.

9 Jones, H.M., Parrott, N., Ohlenbusch, G. and Lave, T. (2006) Predicting pharmacokinetic food effects using biorelevant solubility media and physiologically based modelling. *Clinical Pharmacokinetics*, **45** (12), 1213–1226.

10 De Buck, S.S., Sinha, V.K., Fenu, L.A., Gilissen, R.A., Mackie, C.E. and Nijsen, M.J. (2007) The prediction of drug metabolism, tissue distribution, and bioavailability of 50 structurally diverse compounds in rat using mechanism-based absorption, distribution, and metabolism prediction tools. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **35** (4), 649–659.

11 De Buck, S.S., Sinha, V.K., Fenu, L.A., Nijsen, M.J., Mackie, C.E. and Gilissen, R.A. (2007) Prediction of human pharmacokinetics using physiologically based modeling: a retrospective analysis of 26 clinically tested drugs. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **35** (10), 1766–1780.

12 Rowland, M., Balant, L. and Peck, C. (2004) Physiologically based pharmacokinetics in drug development and regulatory science: a workshop report (Georgetown University, Washington, DC, May 29–30, 2002). *AAPS Journal*, **6** (1), E6.

13 Bolger, M.B., Fraczkiewicz, R. and Steere, B. (2006) *In silico* surrogates for vivo properties: profiling for ADME and toxicological behaviour, in *Exploiting Chemical Diversity for Drug Discovery* (eds P.A. Bartlett and M. Entzeroth), Royal Society of Chemistry, London, pp. 364–381.

14 Bolger, M.B., Fraczkiewicz, R., Entzeroth, M. and Steere, B. (2006) Concepts for *in vitro* profiling: Drug activity, selectivity, and liability, in *Exploiting Chemical Diversity for Drug Discovery* (eds P.A. Bartlett and M. Entzeroth), Royal Society of Chemistry, London, pp. 336–362.

15 Jolivette, L.J. and Ekins, S. (2007) Methods for predicting human drug metabolism. *Advances in Clinical Chemistry*, **43**, 131–176.

16 Haworth, I.S. (2006) Computational drug delivery. *Advanced Drug Delivery Reviews*, **58** (12–13), 1271–1273.

17 Ekins, S., Nikolsky, Y. and Nikolskaya, T. (2005) Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in Pharmacological Sciences*, **26** (4), 202–209.

18 Ekins, S. (2006) Systems-ADME/Tox: resources and network approaches. *Journal of Pharmacological and Toxicological Methods*, **53** (1), 38–66.

19 Powell, J.R. and Gobburu, J.V. (2007) Pharmacometrics at FDA: evolution and impact on decisions. *Clinical Pharmacology and Therapeutics*, **82** (1), 97–102.

20 Lalonde, R.L., Kowalski, K.G., Hutmacher, M.M., Ewy, W., Nichols, D.J., Milligan, P.A., Corrigan, B.W., Lockwood, P.A., Marshall, S.A., Benincosa, L.J., Tensfeldt, T.G., Parivar, K., Amantea, M., Glue, P., Koide, H. and Miller, R. (2007) Model-based drug development. *Clinical Pharmacology and Therapeutics*, **82** (1), 21–32.

21 Zhang, L., Sinha, V., Forgue, S.T., Callies, S., Ni, L., Peck, R. and Allerheiligen, S.R. (2006) Model-based drug development: the road to quantitative pharmacology. *Journal of Pharmacokinetics and Pharmacodynamics*, **33** (3), 369–393.

22 Miller, R., Ewy, W., Corrigan, B.W., Ouellet, D., Hermann, D., Kowalski, K.G., Lockwood, P., Koup, J.R., Donevan, S., El-Kattan, A., Li, C.S., Werth, J.L., Feltner, D.E. and Lalonde, R.L. (2005) How modeling and simulation have enhanced decision making in new drug

development. *Journal of Pharmacokinetics and Pharmacodynamics*, **32** (2), 185–197.

23 Hall, S.D., Thummel, K.E., Watkins, P.B., Lown, K.S., Benet, L.Z., Paine, M.F., Mayo, R.R., Turgeon, D.K., Bailey, D.G., Fontana, R.J. and Wrighton, S.A. (1999) Molecular and physical mechanisms of first-pass extraction. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **27** (2), 161–166.

24 Yu, L.X., Lipka, E., Crison, J.R. and Amidon, G.L. (1996) Transport approaches to the biopharmaceutical design of oral drug delivery system: prediction of intestinal absorption. *Advanced Drug Delivery Reviews*, **19**, 359–376.

25 Jacobs, M.H. (1940) Some aspects of cell permeability to weak electrolytes. *Cold Spring Harbor Symposia on Quantitative Biology*, **8**, 30–39.

26 Hogben, C.A.M., Tocco, D.J., Brodie, B.B. and Schanker, L.S. (1959) On the mechanism of intestinal absorption of drugs. *The Journal of Pharmacology and Experimental Therapeutics*, **125**, 275–282.

27 Palm, K., Luthman, K., Ros, J., Grasjo, J. and Artursson, P. (1999) Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs. *The Journal of Pharmacology and Experimental Therapeutics*, **291** (2), 435–443.

28 Suzuki, A., Higuchi, W.I. and Ho, N.F. (1970) Theoretical model studies of drug absorption and transport in the gastrointestinal tract. 2. *Journal of Pharmaceutical Sciences*, **59** (5), 651–659.

29 Suzuki, A., Higuchi, W.I. and Ho, N.F. (1970) Theoretical model studies of drug absorption and transport in the gastrointestinal tract.1. *Journal of Pharmaceutical Sciences*, **59** (5), 644–651.

30 Ho, N.F., Higuchi, W.I. and Turi, J. (1972) Theoretical model studies of drug absorption and transport in the GI tract. 3. *Journal of Pharmaceutical Sciences*, **61** (2), 192–197.

31 Ho, N.F. and Higuchi, W.I. (1971) Quantitative interpretation of *in vivo* buccal absorption of n-alkanoic acids by the physical model approach. *Journal of Pharmaceutical Sciences*, **60** (4), 537–541.

32 Dressman, J.B., Fleisher, D. and Amidon, G.L. (1984) Physicochemical model for dose-dependent drug absorption. *Journal of Pharmaceutical Sciences*, **73** (9), 1274–1279.

33 Suttle, A.B., Pollack, G.M. and Brouwer, K.L. (1992) Use of a pharmacokinetic model incorporating discontinuous gastrointestinal absorption to examine the occurrence of double peaks in oral concentration–time profiles. *Pharmaceutical Research*, **9** (3), 350–356.

34 Wright, J.D., Ma, T., Chu, C.K. and Boudinot, F.D. (1996) Discontinuous oral absorption pharmacokinetic model and bioavailability of 1-(2-fluoro-5-methyl-beta-L-arabinofuranosyl)uracil (L-FMAU) in rats. *Biopharmaceutics & Drug Disposition*, **17** (3), 197–207.

35 Grass, G.M. (1997) Simulation models to predict oral drug absorption from *in vitro* data. *Advanced Drug Delivery Reviews*, **23**, 199–219.

36 Norris, D.A., Leesman, G.D., Sinko, P.J. and Grass, G.M. (2000) Development of predictive pharmacokinetic simulation models for drug discovery. *Journal of Controlled Release*, **65** (1–2), 55–62.

37 Cong, D., Doherty, M. and Pang, K.S. (2000) A new physiologically based, segregated-f model to explain route-dependent intestinal metabolism. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **28** (2), 224–235.

38 Kalampokis, A., Argyrakis, P. and Macheras, P. (1999) A heterogeneous tube model of intestinal drug absorption based on probabilistic concepts. *Pharmaceutical Research*, **16** (11), 1764–1769.

39 Kalampokis, A., Argyrakis, P. and Macheras, P. (1999) Heterogeneous tube model for the study of small intestinal transit flow. *Pharmaceutical Research*, **16** (1), 87–91.

40 Ito, K., Kusuhara, H. and Sugiyama, Y. (1999) Effects of intestinal CYP3A4 and P-glycoprotein on oral drug absorption–theoretical approach. *Pharmaceutical Research*, **16** (2), 225–231.

41 Rostami-Hodjegan, A. and Tucker, G.T. (2007) Simulation and prediction of *in vivo* drug metabolism in human populations from *in vitro* data. *Nature Reviews. Drug Discovery*, **6** (2), 140–148.

42 Willmann, S., Schmitt, W., Keldenich, J. and Dressman, J.B. (2003) A physiologic model for simulating gastrointestinal flow and drug absorption in rats. *Pharmaceutical Research*, **20** (11), 1766–1771.

43 Willmann, S., Schmitt, W., Keldenich, J., Lippert, J. and Dressman, J.B. (2004) A physiological model for the estimation of the fraction dose absorbed in humans. *Journal of Medicinal Chemistry*, **47** (16), 4022–4031.

44 Agoram, B., Woltosz, W.S. and Bolger, M.B. (2001) Predicting the impact of physiological and biochemical processes on oral drug bioavailability. *Advanced Drug Delivery Reviews*, **50** (Suppl. 1) S41–S67.

45 Aarons, L., Karlsson, M.O., Mentre, F., Rombout, F., Steimer, J.L. and van Peer, A. (2001) Role of modelling and simulation in phase I drug development. *European Journal of Pharmaceutical Sciences*, **13** (2), 115–122.

46 Yu, L.X. and Amidon, G.L. (1999) A compartmental absorption and transit model for estimating oral drug absorption. *International Journal of Pharmaceutics*, **186** (2), 119–125.

47 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23** (1–3), 3–25.

48 Andrews, C.W., Bennett, L. and Yu, L.X. (2000) Predicting human oral bioavailability of a compound: development of a novel quantitative structure–bioavailability relationship *Pharmaceutical Research*, **17** (6), 639–644.

49 Oprea, T.I. and Gottfries, J. (1999) Toward minimalistic modeling of oral drug absorption. *Journal of Molecular Graphics & Modelling*, **17** (5–6), 261–274, 329.

50 Stenberg, P., Luthman, K., Ellens, H., Lee, C.P., Smith, P.L., Lago, A., Elliott, J.D. and Artursson, P. (1999) Prediction of the intestinal absorption of endothelin receptor antagonists using three theoretical methods of increasing complexity. *Pharmaceutical Research*, **16** (10), 1520–1526.

51 Matter, H., Baringhaus, K.H., Naumann, T., Klabunde, T. and Pirard, B. (2001) Computational approaches towards the rational design of drug-like compound libraries. *Combinatorial Chemistry & High Throughput Screening*, **4** (6), 453–475.

52 Osterberg, T. and Norinder, U. (2000) Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics. *Journal of Chemical Information and Computer Sciences*, **40** (6), 1408–1411.

53 Chiou, W.L. and Barve, A. (1998) Linear correlation of the fraction of oral dose absorbed of 64 drugs between humans and rats. *Pharmaceutical Research*, **15** (11), 1792–1795.

54 Zhao, Y.H., Le, J., Abraham, M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Butina, D., Beck, G., Sherborne, B., Cooper, I., Platts, J.A. and Boutina, D. (2001) Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences*, **90** (6), 749–784.

55 Matthews, E.J., Kruhlak, N.L., Benz, R.D. and Contrera, J.F. (2004) Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Current Drug Discovery Technologies*, **1** (1), 61–76.

56 Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E. and Drummond, R.A. (1997) Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry/ SETAC*, **16** (5), 948–967.

57 Gold, L.S., Manley, N.B., Slone, T.H. and Rohrbach, L. (1999) Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environmental Health Perspectives*, **107** (Suppl. 4) 527–600.

58 Bains, W., Basman, A. and White, C. (2004) HERG binding specificity and binding site structure: evidence from a fragment-based evolutionary computing SAR study. *Progress in Biophysics and Molecular Biology*, **86** (2), 205–233.

59 Keseru, G.M. (2003) Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorganic & Medicinal Chemistry Letters*, **13** (16), 2773–2775.

60 Cohen, S.D. and Hindmarsh, A.C. (1996) CVODE, a stiff/nonstiff ODE solver in C. *Computers in Physics*, **10** (2), 138–143.

61 Pade, V. and Stavchansky, S. (1997) Estimation of the relative contribution of the transcellular and paracellular pathway to the transport of passively absorbed drugs in the Caco-2 cell culture model. *Pharmaceutical Research*, **14** (9), 1210–1215.

62 Ungell, A.L., Nylander, S., Bergstrand, S., Sjoberg, A. and Lennernäs, H. (1998) Membrane transport of drugs in different regions of the intestinal tract of the rat. *Journal of Pharmaceutical Sciences*, **87** (3), 360–366.

63 Adson, A., Burton, P.S., Raub, T.J., Barsuhn, C.L., Audus, K.L. and Ho, N.F. (1995) Passive diffusion of weak organic electrolytes across Caco-2 cell monolayers: uncoupling the contributions of hydrodynamic, transcellular, and paracellular barriers. *Journal of Pharmaceutical Sciences*, **84** (10), 1197–1204.

64 Fordtran, J.S., Rector, F.C., Jr. Ewton, M.F., Soter, N. and Kinney, J. (1965) Permeability characteristics of the human small intestine. *The Journal of Clinical Investigation*, **44** (12), 1935–1944.

65 Soergel, K.H., Whalen, G.E. and Harris, J.A. (1968) Passive movement of water and sodium across the human small intestinal mucosa. *Journal of Applied Physiology*, **24** (1), 40–48.

66 Artursson, P. and Karlsson, J. (1991) Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochemical and Biophysical Research Communications*, **175** (3), 880–885.

67 Billich, C.O. and Levitan, R. (1969) Effects of sodium concentration and osmolality on water and electrolyte absorption form the intact human colon. *The Journal of Clinical Investigation*, **48** (7), 1336–1347.

68 Maruyama, W., Yoshida, K., Tanaka, T. and Nakanishi, J. (2002) Determination of tissue–blood partition coefficients for a physiological model for humans, and estimation of dioxin concentration in tissues. *Chemosphere*, **46** (7), 975–985.

69 Sarangapani, R., Teeguarden, J.G., Cruzan, G., Clewell, H.J. and Andersen, M.E. (2002) Physiologically based pharmacokinetic modeling of styrene and styrene oxide respiratory-tract dosimetry in rodents and humans. *Inhalation Toxicology*, **14** (8), 789–834.

70 Thrall, K.D., Soelberg, J.J., Weitz, K.K. and Woodstock, A.D. (2002) Development of a physiologically based pharmacokinetic model for methyl ethyl ketone in F344 rats. *Journal of Toxicology and*

Environmental Health. Part A, **65** (13), 881–896.

71 Nestorov, I.A., Aarons, L.J. and Rowland, M. (1997) Physiologically based pharmacokinetic modeling of a homologous series of barbiturates in the rat: a sensitivity analysis. *Journal of Pharmacokinetics and Biopharmaceutics*, **25** (4), 413–447.

72 Kawai, R., Lemaire, M., Steimer, J.L., Bruelisauer, A., Niederberger, W. and Rowland, M. (1994) Physiologically based pharmacokinetic study on a cyclosporin derivative, SDZ IMM 125. *Journal of Pharmacokinetics and Biopharmaceuticst*, **22** (5), 327–365.

73 Bjorkman, S., Wada, D.R., Berling, B.M. and Benoni, G. (2001) Prediction of the disposition of midazolam in surgical patients by a physiologically based pharmacokinetic model. *Journal of Pharmaceutical Sciences*, **90** (9), 1226–1241.

74 Fouchecourt, M.O., Beliveau, M. and Krishnan, K. (2001) Quantitative structure–pharmacokinetic relationship modelling. *The Science of the Total Environment*, **274** (1–3), 125–135.

75 Pelekis, M., Poulin, P. and Krishnan, K. (1995) An approach for incorporating tissue composition data into physiologically based pharmacokinetic models. *Toxicology and Industrial Health*, **11** (5), 511–522.

76 Payne, M.P. and Kenny, L.C. (2002) Comparison of models for the estimation of biological partition coefficients. *Journal of Toxicology and Environmental Health. Part A*, **65** (13), 897–931.

77 Poulin, P. and Theil, F.P. (2000) *A priori* prediction of tissue:plasma partition coefficients of drugs to facilitate the use of physiologically-based pharmacokinetic models in drug discovery. *Journal of Pharmaceutical Sciences*, **89** (1), 16–35.

78 Luttringer, O., Theil, F.P., Poulin, P., Schmitt-Hoffmann, A.H., Guentert, T.W. and Lave, T. (2003) Physiologically based pharmacokinetic (PBPK) modeling of disposition of epiroprim in humans. *Journal of Pharmaceutical Sciences*, **92** (10), 1990–2007.

79 Theil, F.P., Guentert, T.W., Haddad, S. and Poulin, P. (2003) Utility of physiologically based pharmacokinetic models to drug development and rational drug discovery candidate selection. *Toxicology Letters*, **138** (1–2), 29–49.

80 Poulin, P. and Theil, F.P. (2002) Prediction of pharmacokinetics prior to *in vivo* studies. II. Generic physiologically based pharmacokinetic models of drug disposition. *Journal of Pharmaceutical Sciences*, **91** (5), 1358–1370.

81 De Buck, S.S. and Mackie, C.E. (2007) Physiologically based approaches towards the prediction of pharmacokinetics: *in vitro–in vivo* extrapolation. *Expert Opinion on Drug Metabolism and Toxicology*, **3** (6), 865–878.

82 Rodgers, T., Leahy, D. and Rowland, M. (2005) Tissue distribution of basic drugs: accounting for enantiomeric, compound and regional differences amongst beta-blocking drugs in rat. *Journal of Pharmaceutical Sciences*, **94** (6), 1237–1248.

83 Rodgers, T., Leahy, D. and Rowland, M. (2005) Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. *Journal of Pharmaceutical Sciences*, **94** (6), 1259–1276.

84 Rodgers, T. and Rowland, M. (2006) Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. *Journal of Pharmaceutical Sciences*, **95** (6), 1238–1257.

85 Rodgers, T. and Rowland, M. (2007) Mechanistic approaches to volume of distribution predictions: understanding the processes. *Pharmaceutical Research*, **24** (5), 918–933.

86 Brightman, F.A., Leahy, D.E., Searle, G.E. and Thomas, S. (2006) Application of a generic physiologically based pharmacokinetic model to the estimation of xenobiotic levels in rat plasma. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **34** (1), 84–93.

87 Willmann, S., Hohn, K., Edginton, A., Sevestre, M., Solodenko, J., Weiss, W., Lippert, J. and Schmitt, W. (2007) Development of a physiology-based whole-body population model for assessing the influence of individual variability on the pharmacokinetics of drugs. *Journal of Pharmacokinetics and Pharmacodynamics*, **34** (3), 401–431.

88 Paine, M.F., Khalighi, M., Fisher, J.M., Shen, D.D., Kunze, K.L., Marsh, C.L., Perkins, J.D. and Thummel, K.E. (1997) Characterization of interintestinal and intraintestinal variations in human CYP3A-dependent metabolism. *The Journal of Pharmacology and Experimental Therapeutics*, **283** (3), 1552–1562.

89 Madani, S., Paine, M.F., Lewis, L., Thummel, K.E. and Shen, D.D. (1999) Comparison of CYP2D6 content and metoprolol oxidation between microsomes isolated from human livers and small intestines. *Pharmaceutical Research*, **16** (8), 1199–1205.

90 Bornemann, L.D., Min, B.H., Crews, T., Rees, M.M., Blumenthal, H.P., Colburn, W.A. and Patel, I.H. (1985) Dose dependent pharmacokinetics of midazolam. *European Journal of Clinical Pharmacology*, **29** (1), 91–95.

91 Kupferschmidt, H.H., Fattinger, K.E., Ha, H.R., Follath, F. and Krahenbuhl, S. (1998) Grapefruit juice enhances the bioavailability of the HIV protease inhibitor saquinavir in man. *British Journal of Clinical Pharmacology*, **45** (4), 355–359.

92 Gupta, S.K. and Sathyan, G. (1999) Pharmacokinetics of an oral once-a-day controlled-release oxybutynin formulation compared with immediate-release oxybutynin. *Journal of Clinical Pharmacology*, **39** (3), 289–296.

93 Korolev, D., Balakin, K.V., Nikolsky, Y., Kirillov, E., Ivanenkov, Y.A., Savchuk, N.P., Ivashchenko, A.A. and Nikolskaya, T. (2003) Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *Journal of Medicinal Chemistry*, **46** (17), 3631–3643.

94 Balakin, K.V., Ekins, S., Bugrim, A., Ivanenkov, Y.A., Korolev, D., Nikolsky, Y.V., Skorenko, A.V., Ivashchenko, A.A., Savchuk, N.P. and Nikolskaya, T. (2004) Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **32** (10), 1183–1189.

95 Korzekwa, K.R., Jones, J.P. and Gillette, J.R. (1990) Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *Journal of the American Chemical Society*, **112**, 7042–7046.

96 Korzekwa, K.R., Trager, W.F., Mancewicz, J. and Osawa, Y. (1993) Studies on the mechanism of aromatase and other cytochrome P450 mediated deformylation reactions. *The Journal of Steroid Biochemistry and Molecular Biology*, **44** (4–6), 367–373.

97 Korzekwa, K.R. and Jones, J.P. (1993) Predicting the cytochrome P450 mediated metabolism of xenobiotics. *Pharmacogenetics*, **3** (1), 1–18.

98 Jones, J.P. and Korzekwa, K.R. (1996) Predicting the rates and regioselectivity of reactions mediated by the P450 superfamily. *Methods in Enzymology*, **272**, 326–335.

99 Jones, J.P., Mysinger, M. and Korzekwa, K.R. (2002) Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **30** (1), 7–12.

100 Singh, S.B., Shen, L.Q., Walker, M.J. and Sheridan, R.P. (2003) A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *Journal of Medicinal Chemistry*, **46** (8), 1330–1306.

101 Sheridan, R.P., Korzekwa, K.R., Torres, R.A. and Walker, M.J. (2007) Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *Journal of Medicinal Chemistry*, **50** (14), 3173–3184.

102 de Groot, M.J., Vermeulen, N.P., Kramer, J.D., van Acker, F.A. and Donne-Op den Kelder, G.M. (1996) A three-dimensional protein model for human cytochrome P450 2D6 based on the crystal structures of P450 101, P450 102, and P450 108. *Chemical Research in Toxicology*, **9** (7), 1079–1091.

103 de Groot, M.J., Alex, A.A. and Jones, B.C. (2002) Development of a combined protein and pharmacophore model for cytochrome P450 2C9. *Journal of Medicinal Chemistry*, **45** (10), 1983–1993.

104 de Groot, M.J., Kirton, S.B. and Sutcliffe, M.J. (2004) *In silico* methods for predicting ligand binding determinants of cytochromes P450. *Current Topics in Medicinal Chemistry*, **4** (16), 1803–1824.

105 de Groot, M.J. (2006) Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discovery Today*, **11** (13–14), 601–606.

106 Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T. and Vianello, R. (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of Medicinal Chemistry*, **48** (22), 6970–6979.

107 Zamora, I., Afzelius, L. and Cruciani, G. (2003) Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9. *Journal of Medicinal Chemistry*, **46** (12), 2313–2324.

108 O'Neill, R.T. (2006) FDA's critical path initiative: a perspective on contributions of biostatistics. *Biometrical Journal*, **48** (4), 559–564.

109 Woosley, R.L. and Cossman, J. (2007) Drug development and the FDA's critical path initiative. *Clinical Pharmacology and Therapeutics*, **81** (1), 129–133.

110 Woodcock, J. and Woosley, R. (2008) The FDA critical path initiative and its influence on new drug development. *Annual Review of Medicine*, **59**, 1–12.

111 Leonce, S., Pierre, A., Anstett, M., Perez, V., Genton, A., Bizzari, J.P. and Atassi, G. (1992) Effects of a new triazinoaminopiperidine derivative on adriamycin accumulation and retention in cells displaying P-glycoprotein-mediated multidrug resistance. *Biochemical Pharmacology*, **44** (9), 1707–1715.

112 Hsiang, B., Zhu, Y., Wang, Z., Wu, Y., Sasseville, V., Yang, W.P. and Kirchgessner, T.G. (1999) A novel human hepatic organic anion transporting polypeptide (OATP2). Identification of a liver-specific human organic anion transporting polypeptide and identification of rat and human hydroxymethylglutaryl-CoA reductase inhibitor transporters. *The Journal of Biological Chemistry*, **274** (52), 37161–37168.

113 Chen, C., Lin, J., Smolarek, T. and Tremaine, L. (2007) P-Glycoprotein has differential effects on the disposition of statin acid and lactone forms in mdr1a/b knockout and wild-type mice. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **35** (10), 1725–1729.

114 Holtzman, C.W., Wiggins, B.S. and Spinler, S.A. (2006) Role of P-glycoprotein in statin drug interactions. *Pharmacotherapy*, **26** (11), 1601–1607.

115 de Koning, H.P. and Jarvis, S.M. (1997) Hypoxanthine uptake through a purine-selective nucleobase transporter in Trypanosoma brucei brucei procyclic cells is driven by protonmotive force. *European Journal of Biochemistry*, **247** (3), 1102–1110.

116 Balch, W.E. and Wolfe, R.S. (1979) Transport of coenzyme M (2-mercaptoethanesulfonic acid) in methanobacterium ruminantium. *Journal of Bacteriology*, **137** (1), 264–273.

117 Shima, Y., Maeda, T., Aizawa, S., Tsuboi, I., Kobayashi, D., Kato, R. and Tamai, I. (2006) L-arginine import via cationic amino acid transporter CAT1 is essential for both differentiation and proliferation of erythrocytes. *Blood*, **107** (4), 1352–1356.

118 Zhou, M., Xia, L. and Wang, J. (2007) Metformin transport by a newly cloned proton-stimulated organic cation transporter (plasma membrane monoamine transporter) expressed in human intestine. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **35** (10), 1956–1962.

119 Swaan, P.W. and Tukker, J.J. (1997) Molecular determinants of recognition for the intestinal peptide carrier. *Journal of Pharmaceutical Sciences*, **86** (5), 596–602.

120 Nagakubo, J., Tomimatsu, T., Kitajima, M., Takayama, H., Aimi, N. and Horie, T. (2001) Characteristics of transport of fluoresceinated methotrexate in rat small intestine. *Life Sciences*, **69** (7), 739–747.

121 Luckner, P. and Brandsch, M. (2005) Interaction of 31 beta-lactam antibiotics with the H + /peptide symporter PEPT2: analysis of affinity constants and comparison with PEPT1. *European Journal of Pharmaceutics and Biopharmaceutics*, **59** (1), 17–24.

122 Katiyar, S.K. and Edlind, T.D. (2001) Identification and expression of multidrug resistance-related ABC transporter genes in *Candida krusei*. *Medical Mycology*, **39** (1), 109–116.

123 Chang, C., Bahadduri, P.M., Polli, J.E., Swaan, P.W. and Ekins, S. (2006) Rapid identification of P-glycoprotein substrates and inhibitors. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, **34** (12), 1976–1984.

124 Terao, T., Hisanaga, E., Sai, Y., Tamai, I. and Tsuji, A. (1996) Active secretion of drugs from the small intestinal epithelium in rats by P-glycoprotein functioning as an absorption barrier. *The Journal of Pharmacy and Pharmacology*, **48** (10), 1083–1089.

125 Honscha, W., Schulz, K., Muller, D. and Petzinger, E. (1993) Two different mRNAs from rat liver code for the transport of bumetanide and taurocholate in *Xenopus laevis* oocytes. *European Journal of Pharmacology*, **246** (3), 227–232.

126 Sadoogh-Abasian, F. and Evered, D.F. (1980) Absorption of nicotinic acid and nicotinamide from rat small intestine *in vitro*. *Biochimica et Biophysica Acta*, **598** (2), 385–391.

127 Hu, M. and Amidon, G.L. (1988) Passive and carrier-mediated intestinal absorption components of captopril. *Journal of Pharmaceutical Sciences*, **77** (12), 1007–1011.

128 Nakamura, H., Sano, H., Yamazaki, M. and Sugiyama, Y. (1994) Carrier-mediated active transport of histamine H2 receptor antagonists, cimetidine and nizatidine, into isolated rat hepatocytes: contribution of type I system. *The Journal of Pharmacology and Experimental Therapeutics*, **269** (3), 1220–1227.

129 de Lange, E.C., Marchand, S., van den Berg, D., van der Sandt, I.C., de Boer, A.G., Delon, A., Bouquet, S. and Couet, W. (2000) *In vitro* and *in vivo* investigations on fluoroquinolones; effects of the P-glycoprotein efflux transporter on brain distribution of sparfloxacin. *European Journal of Pharmaceutical Sciences*, **12** (2), 85–93.

130 Swaan, P.W., Stehouwer, M.C. and Tukker, J.J. (1995) Molecular mechanism for the relative binding affinity to the intestinal peptide carrier. Comparison of three ACE-inhibitors: enalapril, enalaprilat, and lisinopril. *Biochimica et Biophysica Acta*, **1236** (1), 31–38.

131 Lauterbach, F. (1987) Intestinal permeation of nonquaternary amines: a study with telenzepine and pirenzepine in the isolated mucosa of guinea pig jejunum and colon. *The Journal of Pharmacology and Experimental Therapeutics*, **243** (3), 1121–1130.

132 Boulton, D.W., DeVane, C.L., Liston, H.L. and Markowitz, J.S. (2002) *In vitro* P-glycoprotein affinity for atypical and conventional antipsychotics. *Life Sciences*, **71** (2), 163–169.

133 Covitz, K.M., Amidon, G.L. and Sadee, W. (1996) Human dipeptide transporter, hPEPT1, stably transfected into Chinese hamster ovary cells. *Pharmaceutical Research*, **13** (11), 1631–1634.

134 Hatanaka, T. (2000) Clinical pharmacokinetics of pravastatin: mechanisms of pharmacokinetic events. *Clinical Pharmacokinetics*, **39** (6), 397–412.

135 Metzner, L., Neubert, K. and Brandsch, M. (2006) Substrate specificity of the amino acid transporter PAT1. *Amino Acids*, **31** (2), 111–117.

136 Anderson, C.M. and Thwaites, D.T. (2005) Indirect regulation of the intestinal H + -coupled amino acid transporter hPAT1 (SLC36A1). *Journal of Cellular Physiology*, **204** (2), 604–613.

137 van Kalken, C.K., Broxterman, H.J., Pinedo, H.M., Feller, N., Dekker, H., Lankelma, J. and Giaccone, G. (1993) Cortisol is transported by the multidrug resistance gene product P-glycoprotein. *British Journal of Cancer*, **67** (2), 284–289.

138 Yao, H.M. and Chiou, W.L. (2006) The complexity of intestinal absorption and exsorption of digoxin in rats. *International Journal of Pharmaceutics*, **322** (1–2), 79–86.

139 Lau, Y.Y., Wu, C.Y., Okochi, H. and Benet, L.Z. (2004) *Ex situ* inhibition of hepatic uptake and efflux significantly changes metabolism: hepatic enzyme–transporter interplay. *The Journal of Pharmacology and Experimental Therapeutics*, **308** (3), 1040–1045.

140 Makhey, V.D., Guo, A., Norris, D.A., Hu, P., Yan, J. and Sinko, P.J. (1998) Characterization of the regional intestinal kinetics of drug efflux in rat and human intestine and in Caco-2 cells. *Pharmaceutical Research*, **15** (8), 1160–1167.

141 Swaan, P.W. and Tukker, J.J. (1995) Carrier-mediated transport mechanism of foscarnet (trisodium phosphonoformate hexahydrate) in rat intestinal tissue. *The Journal of Pharmacology and Experimental Therapeutics*, **272** (1), 242–247.

142 Collett, A., Higgs, N.B., Sims, E., Rowland, M. and Warhurst, G. (1999) Modulation of the permeability of H2 receptor antagonists cimetidine and ranitidine by P-glycoprotein in rat intestine and the human colonic cell line Caco-2. *The Journal of Pharmacology and Experimental Therapeutics*, **288** (1), 171–178.

143 Uchino, H., Kanai, Y., Kim, D.K., Wempe, M.F., Chairoungdua, A., Morimoto, E., Anders, M.W. and Endou, H. (2002) Transport of amino acid-related compounds mediated by L-type amino acid transporter 1 (LAT1): insights into the mechanisms of substrate recognition. *Molecular Pharmacology*, **61** (4), 729–737.

144 Johansson, O., Lindberg, T., Melander, A. and Wahlin-Boll, E. (1985) Different effects of different nutrients on theophylline absorption in man. *Drug-Nutrient Interactions*, **3** (4), 205–211.

145 Metzner, L., Kottra, G., Neubert, K., Daniel, H. and Brandsch, M. (2005) Serotonin, L-tryptophan, and tryptamine are effective inhibitors of the amino acid transport system PAT1. *FASEB Journal*, **19** (11), 1468–1473.

146 Metzner, L. and Brandsch, M. (2006) Influence of a proton gradient on the transport kinetics of the H + /amino acid cotransporter PAT1 in Caco-2 cells. *European Journal of Pharmaceutics and Biopharmaceutics*, **63** (3), 360–364.

147 Lambie, D.G. and Johnson, R.H. (1985) Drugs and folate metabolism. *Drugs*, **30** (2), 145–155.

148 Zimmerman, J., Selhub, J. and Rosenberg, I.H. (1987) Competitive inhibition of folate absorption by dihydrofolate reductase inhibitors, trimethoprim and pyrimethamine. *The American Journal of Clinical Nutrition*, **46** (3), 518–522.

149 Broer, S. (2006) The SLC6 orphans are forming a family of amino acid transporters. *Neurochemistry International*, **48** (6–7), 559–567.

150 Zhang, L., Li, X.Z. and Poole, K. (2001) Fluoroquinolone susceptibilities of efflux-mediated multidrug-resistant *Pseudomonas aeruginosa*, *Stenotrophomonas maltophilia* and *Burkholderia cepacia*. *The Journal of Antimicrobial Chemotherapy*, **48** (4), 549–552.

151 Uchino, H., Kanai, Y., Kim do, K., Wempe, M.F., Chairoungdua, A., Morimoto, E., Anders, M.W. and Endou, H. (2002) Transport of amino acid-related compounds mediated by L-type amino acid transporter 1 (LAT1): insights into the mechanisms of substrate recognition. *Molecular Pharmacology*, **61** (4), 729–737.

152 Saitoh, H. and Aungst, B.J. (1995) Possible involvement of multiple P-glycoprotein-mediated efflux systems in the transport of verapamil and other organic cations across rat intestine. *Pharmaceutical Research*, **12** (9), 1304–1310.

153 Ngo, L.Y., Patil, S.D. and Unadkat, J.D. (2001) Ontogenic and longitudinal activity of Na(+)-nucleoside transporters in the human intestine. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, **280** (3), G475–G481.

154 Wenzel, U., Thwaites, D.T. and Daniel, H. (1995) Stereoselective uptake of beta-lactam antibiotics by the intestinal peptide transporter. *British Journal of Pharmacology*, **116** (7), 3021–3027.

155 Hu, M., Chen, J., Zhu, Y., Dantzig, A.H., Stratford, R.E., Jr and Kuhfeld, M.T. (1994) Mechanism and kinetics of transcellular transport of a new beta-lactam antibiotic loracarbef across an intestinal epithelial membrane model system (Caco-2). *Pharmaceutical Research*, **11** (10), 1405–1413.

18
Toward Understanding P-Glycoprotein Structure–Activity Relationships

Anna Seelig

Abbreviations

ABC	ATP-binding cassette (transport protein)
ATP	Adenosine triphosphate
BBB	Blood–brain barrier
GRIND	Grid independent descriptors
IUPAC	International Union of Pure and Applied Chemistry
LDA	Linear discriminant analysis
MDR	Multidrug resistance
NBD	Nucleotide-binding domain
PCA	Principle component analysis
PLS-DA	Partial least square discriminant analysis
P-gp	P-Glycoprotein (MDR1, ABCB1)
(Q)SAR	(Quantitative) structure–activity relationship
Sav1866	ABC transporter from *Staphylococcus aureus*
SVM	Support vector machine
TMD	Transmembrane domain

Symbols

C_{Saq}	Substrate concentration in aqueous solution
K_1	Substrate concentration at half-maximum P-gp activation
K_2	Substrate concentration of at half-minimum P-gp activation
V_0	Basal P-gp activity in the absence of substrates
V_1	Maximum transporter activity
V_2	Minimum transporter activity
V_{Saq}	Transporter activity at a given substrate concentration in aqueous solution
k	Rate constant
$K_{\mathrm{tw(1)}}$	Binding constant of a drug from water to the activating binding region of the transporter

$K_{\mathrm{tl}(1)}$	Binding constant of a drug from the lipid phase to the inhibitory binding region of the transporter
K_{lw}	Lipid–water partition coefficient
$\Delta G^0_{\mathrm{tw}(1)}$	Free energy of binding of a substrate from water to the activating binding region of the transporter
$\Delta G^0_{\mathrm{tl}(1)}$	Free energy of binding of a substrate from the lipid phase to the activating binding region of the transporter
ΔG^0_{lw}	Free energy of partitioning of a substrate from water to the lipid membrane
J	Net flux
Φ	Passive flux
IC_{50}	Half-maximum (50%) inhibitory concentration

18.1 Introduction

P-Glycoprotein (P-gp/MDR1/ABCB1) is an efflux transporter of broad substrate specificity that is encoded by the multidrug resistance (MDR) 1 gene (*MDR1*) [1]. P-gp was first observed in multidrug resistant cancer cells [2]. It is also highly expressed in different plasma membrane barriers with protective functions, such as the intestinal barrier (IB) [3, 4], the blood–brain barrier (BBB) [3, 5], the placental barrier [6], and the blood–testis barrier [7], where it reduces or even prevents the absorption of a broad range of drugs and toxins (for review see Ref. [3, 8]). Recently, P-gp was detected in the nuclear membrane [9] where it contributes to an additional protection shell around the nucleus. P-gp not only prevents absorption but also plays a role in the excretion of drugs, toxins, and their metabolites, for example, in proximal tubules of the kidney and biliary ducts of the liver [3].

Cells can be induced to overexpress P-gp after exposure to a single agent (e.g., anticancer drugs, certain antibiotics, or food components) [10] or even after exposure to physical stress, such as X-ray [11], UV light irradiation [12], or heat shock [13]. Overexpression of P-gp leads to multidrug resistance, that is, to a resistance toward all drugs that are substrates for P-gp. The expression level of P-gp not only depends on the exposure of cells to various stimuli but also on genetic factors [14].

The same type of stimuli that induce MDR due to P-gp in human can also induce MDR in bacteria, parasites, and fungi by promoting the expression of related ABC transporters. MDR is detrimental not only for the treatment of cancers (for review see Ref. [15]), but also for the treatment of bacterial [16], parasitic [17], and fungal [18] diseases and can be considered as a general problem for pharmacotherapy.

18.1.1 Similarity Between P-gp and Other ABC Transporters

*A*TP-*b*inding *c*assette transport proteins (ABCs) are phylogenetically highly conserved and transport a large variety of compounds across cell membranes. The 48 human ABC transporters are grouped into seven subfamilies (A–G) according to

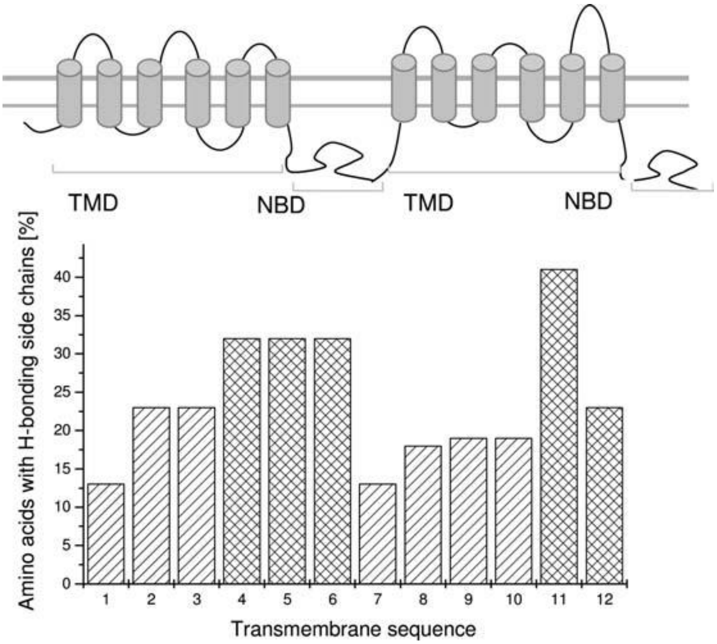

Figure 18.1 The putative transmembrane domains of P-gp derived from hydropathy plots. Hydropathy analyses search for all clusters of about 20–22 amino acids in a protein, which are hydrophobic enough to form a transmembrane sequence. Upper panel: P-gp in a 2D model derived from hydropathy plots comprises two halves each consisting of six putative α-helices (gray tubes) (TMDs) followed by a nucleotide-binding domain. Lower panel: the percentage of hydrogen-bond donor side chains in the putative transmembrane sequences of P-gp. The crosshatched putative α-helices are known to be especially important for binding and transport of substrates (updated version of Figure 18.3 in Ref. [26]).

similarities in their amino acid sequences [19]. On the basis of hydropathy plots, most human ABC transporters are predicted to consist of two homologous parts, each consisting of a transmembrane domain (TMD) and a cytosolic nucleotide-binding domain (NBD) coupled by a cytosolic linker region. P-gp (MDR1/ABCB1) (Figure 18.1) is the best studied example.

Some transporters (e.g., BCRP/ABCG2) are half-transporters (with one TMD comprising six transmembrane α-helices and one cytosolic NBD) that only function as homodimers, like the prokaryotic ABC transporter (e.g., Sav1866). Other members (e.g., ABCC1, ABCC2, ABCC3, ABCC6, and ABCC10) exhibit an additional amino-terminal TMD [20]. Despite these variations, overlapping substrate specificity has been observed, for example, between P-gp/ABCB1and MRP1/ABCC1 [21] as well as between P-gp/ABCB1 and BCRP/ABCG2 [22].

Like many membrane proteins, P-gp (170 kDa) has been recalcitrant to crystallization. So far, only a low-resolution (~8 Å) structure from two-dimensional crystals of P-glycoprotein trapped in the nucleotide-bound state has been obtained by electron microscopy [23]. A high-resolution crystal structure is available for a homologous bacterial ABC transporter, Sav1866 (64.9 kDa). It was also crystallized in the nucleotide-bound state [24, 25] and was found to be a homodimer, formed from two TMD units, each consisting of six transmembrane α-helices.

18.1.2 Why P-gp Is Special

P-gp differs from many well-characterized membrane transporters such as sugar or amino acid transporters. First, it transports not one specific class of compounds but an intriguing number of chemically unrelated drugs, toxins, and metabolites (see, e.g., Ref. [26]). Second, it seems to exhibit not one single well-defined binding site but several binding sites [1, 27, 28]. The different binding sites may not even be well-defined, lock–key-type binding sites but may constitute a binding region that is occupied only transiently [21]. Third, P-gp recognizes its substrates not when they are dissolved in aqueous phase but when they are dissolved in the lipid membrane [29]; more precisely, when the substrates are dissolved in the membrane leaflet facing the cytosol [30, 31]. This implies that binding occurs in two consecutive binding steps, partitioning from water into lipid followed by partitioning from lipid into the P-gp-binding region. The membrane concentration of the substrate thus determines the P-gp activity [32].

In silico methods that are able to predict quantitative aspects of the interaction of a substrate with P-gp would be of great value. So far, modeling was applied mainly to lock–key-type reactions taking place in aqueous solution. The structural diversity and lipid solubility of P-gp substrates and the fact that their encounter with the transporter takes place in the lipid membrane and not in aqueous solution are new challenges for *in silico* predictions. Since all *in silico* models are based on experimental data, we first provide a short introduction to various P-gp assays and discuss their underlying principles (18.2). Secondly, we summarize the different *in silico* approaches (18.3), and, lastly, we discuss the parameters that are most relevant for the different *in silico* models (18.4).

18.2 Measurement of P-gp Function

Different assays are used to monitor the function of P-gp such as (i) ATPase assays; (ii) drug transport assays across confluent, polarized cell monolayers; and (iii) competition assays with reference substrates. The different assays address different functional aspects of P-gp.

18.2.1 P-gp ATPase Activity Assay

P-gp ATPase activity is measured using either inside-out cellular vesicles of MDR1-transfected cells or reconstituted proteoliposomes. In both types of systems, NBDs are oriented at least partially toward the extravesicular side, and ATP hydrolysis can therefore be monitored with a colorimetric [33–35] or a coupled enzyme assay [36].

For cells *in vitro*, glycolysis is the main metabolic pathway and yields one molecule of lactic acid per molecule of ATP synthesized; the lactic acid leaves the cell as a waste product. At steady state, the rate of ATP synthesis corresponds to the rate of ATP hydrolysis and can therefore be monitored in living MDR1-transfected cells by measuring the rate of lactic acid extrusion by the cell. Lactic acid extrusion can be measured either by a spectroscopic approach [37] or by recording the extracellular acidification rate (ECAR) in NIH-MDR1-G185 cells [32, 38] with a micro-pH meter based on silicon chip technology (Cytosensor microphysiometer) [39]. A graphical representation of these assays is shown in Figure 18.2.

P-gp shows a basal ATPase activity in the absence of exogenous compounds; on addition of drugs, the ATPase activity can increase or decrease. Drug-induced inorganic phosphate release [33–35] or ECAR [32] shows a bell-shaped dependence on the drug concentration (log scale), first increasing to a maximum and then decreasing at high concentrations (Figure 18.3). Both equimolar (e.g., Ref. [40]) and equitoxic (e.g., Refs [41, 42]) concentrations have been used to classify compounds as substrates, modulators, or inhibitors. The fact that the same drugs can either activate or inhibit P-gp depending on the assay concentration (Figure 18.3) may explain the numerous inconsistencies in the classification of drugs with respect to their effects on P-gp.

Different models have been used to analyze P-gp activity profiles [32–34]. Here, we describe the modified Michaelis–Menten equation proposed by Litman *et al.* [33]. It

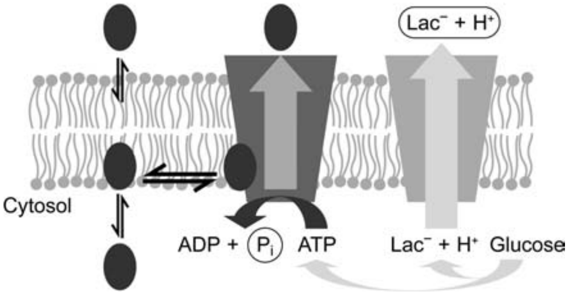

Figure 18.2 ATPase assays: In living cells, the drug first partitions into the extracellular leaflet of the plasma membrane and then crosses the membrane by passive diffusion. Once the drug reaches the cytosolic membrane leaflet, it either escapes to the cytosol or is captured by P-gp (indicated in dark gray). The diffusion process, that is, passive influx, can vary by several orders of magnitude. If the drug is bound by P-gp (which is more likely if diffusion through the intracellular leaflet is slow), it can be exported out of the cell at the expense of ATP hydrolysis. ATP in cultured cells is produced via glycolysis; whereby an equimolar amount of lactic acid is formed, which leaves the cell as a waste product, and dissociates extracellularly to lactate and a proton. This can be monitored with a Cytosensor as an extracellular acidification rate. Cytosensor assays are performed under steady-state conditions. In contrast, in inside-out plasma membrane vesicles, the drug first partitions into the cytosolic leaflet of the plasma membrane. P-gp activation can be measured by monitoring inorganic phosphate released by ATPase activity using a colorimetric assay.

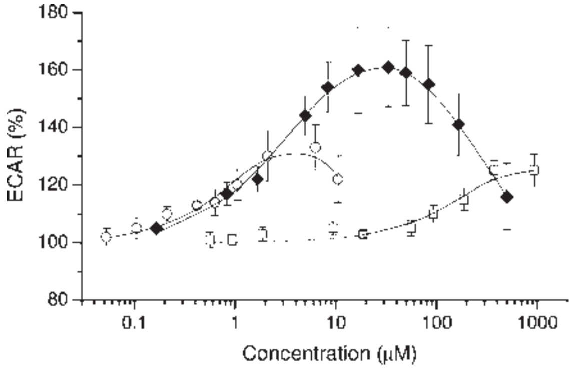

Figure 18.3 P-gp activity profiles measured with living MDR1-transfected cells as a function of drug concentration. The extracellular acidification rate is expressed as a percentage of the basal rate (100%): verapamil is represented by lozenges, lidocaine by open squares, and trifluopromazine by open circles. (Data are taken from Ref. [32].)

assumes activation with one substrate molecule, S, bound, and inhibition with two substrate molecules bound to P-gp as described by Scheme 18.1:

$$\begin{array}{l} \mathrm{T} + \mathrm{ATP} + \mathrm{S} \underset{\mathrm{k}_{-1}}{\overset{\mathrm{k}_1}{\rightleftharpoons}} \mathrm{T(ATP)S} + \mathrm{S} \underset{\mathrm{k}_{-2}}{\overset{\mathrm{k}_2}{\rightleftharpoons}} \mathrm{T(ATP)S_2} \xrightarrow{\mathrm{k}''} \mathrm{T(ADP)} + 2\mathrm{S}_{\mathrm{rel}} + \mathrm{P_i} + \mathrm{H}^+ \\ \qquad\qquad\qquad\quad \downarrow \mathrm{k}' \\ \qquad\qquad \mathrm{T(ADP)} + \mathrm{P_i} + \mathrm{H}^+ + \mathrm{S}_{\mathrm{rel}} \end{array}$$

Scheme 18.1

T(ATP)S and T(ATP)S_2 are transporter–ATP complexes with one and two substrate molecules bound, respectively; T(ADP) is the transporter–ADP complex; P_i inorganic phosphate; S_{rel} is the substrate molecule flipped to the outer leaflet or released extracellularly; k_1, k_{-1}, and k_2, k_{-2} are the rate constants of the first and the second substrate binding steps, respectively; and k' and k'' the rate constants of the catalytic steps. For this model, the rate of ATP hydrolysis is a function of the P-gp-stimulating drug concentration:

$$V_{\mathrm{Saq}} = \frac{K_1 K_2 V_0 + K_2 V_1 C_{\mathrm{Saq}} + V_2 C_{\mathrm{Saq}}^2}{K_1 K_2 + K_2 C_{\mathrm{Saq}} + C_{\mathrm{Saq}}^2}, \qquad (18.1)$$

where V_{Saq} is the rate of P_i release as a function of the substrate concentration in solution, C_{Saq}; V_0 is the basal activity in the absence of substrate; V_1 is the maximal ATPase rate that is achieved only when the inhibitory second step is negligible; V_2 is the minimal rate at infinite substrate concentration and lower than V_1; K_1 is the drug concentration at half-maximum activation, that is, $V_1/2$; K_2, the drug concentration at half-minimum activation, that is, $V_2/2$. At low drug concentrations, Equation 18.1 simplifies to the Michaelis–Menten equation. The catalytic rate constant (k') corresponds to $V_1/[T_0]$, where $[T_0]$ is the transporter concentration.

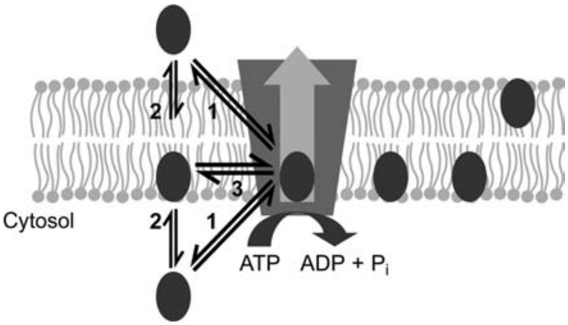

Figure 18.4 Binding from water to the transporter (1) is divided into two steps: membrane partitioning (2) and transporter binding (3). These processes are fast and can be considered (to a first approximation) as equilibrium processes in inside-out vesicles as well as in cells under steady-state conditions in a Cytosensor: this is indicated by the double arrows. The free energy of binding of the drug from water to the transporter ($\Delta G^0_{\mathrm{tw}(1)}$) was derived from ATPase assays, and the free energy of binding of the drug from water to the lipid membrane (ΔG^0_{lw}) was derived from surface activity measurements or from isothermal titration calorimetry. The free energy of binding of the drug from the lipid membrane to the transporter (ΔG^0_{tl}) is not directly accessible by experiment but can be estimated as the difference between the two free energies $\Delta G^0_{\mathrm{tw}(1)}$ and ΔG^0_{lw} (see Equation 18.3).

Evidence for a direct correlation between the turnover number for vinblastine-stimulated ATP hydrolysis and vinblastine transport rate was provided by Ambudkar and Stein [43]. Since compounds that interact with P-gp often exhibit a high lipid–water partition coefficient and can cross the lipid bilayer by passive diffusion, the stoichiometry between ATP hydrolysis and drug transport is difficult to assess. Using a permanently charged spin-labeled analogue of verapamil that cannot cross the membrane by passive diffusion [44], a direct correlation between ATP hydrolysis and drug transport was demonstrated [45].

18.2.1.1 **Quantification of Substrate–Transporter Interactions**

Substrate binding from water to the transporter can be described as a two-step binding process [32] as illustrated in Figure 18.4.

ATPase activation experiments are performed under steady-state conditions, and the catalytic rate (rate constant, $k_1 \approx 1$–$5\,\mathrm{s}^{-1}$) of P-gp is much slower than the rates of drug and ATP binding. Hence, the concentration at half-maximum activation (K_1) can be considered as the dissociation constant and $1/K_1$ as the binding constant of a drug to the activating binding region of the transporter ($K_{\mathrm{tw}(1)}$) to a first approximation. The binding constant of the drug to the transporter ($K_{\mathrm{tw}(1)}$) can then be expressed as the product of the lipid–water partition coefficient (K_{lw}) and the binding constant of the substrate from the lipid membrane to the activating binding site of the transporter ($K_{\mathrm{tl}(1)}$),

$$\frac{1}{K_1} \cong K_{\mathrm{tw}(1)} \cong K_{\mathrm{tl}(1)} \cdot K_{\mathrm{lw}}. \tag{18.2}$$

This leads to the free energy relationship,

$$\Delta G^0_{\mathrm{tw}(1)} \cong \Delta G^0_{\mathrm{tl}(1)} + \Delta G^0_{\mathrm{lw}}, \tag{18.3}$$

where the superscript zero refers to a biological standard state (pH 7.4 and 37 °C). The free energy of substrates binding from water to the transporter, $\Delta G^0_{\mathrm{tw}(1)}$, and the free energy of partitioning into the lipid membrane, ΔG^0_{lw}, are defined as

$$\Delta G^0_{\mathrm{tw}(1)} \cong -RT\ln(C_{\mathrm{w}} K_{\mathrm{tw}(1)}) \tag{18.4}$$

and

$$\Delta G^0_{\mathrm{lw}} = -RT\ln(C_{\mathrm{w}} K_{\mathrm{lw}}), \tag{18.5}$$

respectively, where C_{w} (55.3 mol/l) corresponds to the molar concentration of water at 37 °C. Analogous equations can be formulated for the binding constant, $K_{\mathrm{tw}(2)}$, and the free energy of binding, $\Delta G^0_{\mathrm{tw}(2)}$, to the second binding region as outlined previously [32]. The more negative the free energy of binding is, the higher is the binding affinity to the transporter. For 15 drugs [32], the free energy of drug partitioning from water to the lipid membrane, ΔG^0_{lw}, was somewhat more negative than the free energy of drug binding from the lipid phase to the transporter, $\Delta G^0_{\mathrm{tl}(1)}$. However, the variation in $\Delta G^0_{\mathrm{tl}(1)}$ was more pronounced (~fourfold) than that in ΔG^0_{lw} (~1.5 fold) as shown in Figure 18.5.

18.2.1.2 Relationship between Substrate–Transporter Affinity and Rate of Transport

As seen in Figure 18.6, the maximal extent of P-gp ATPase stimulation, which correlates with the rate of intrinsic transport, $\ln k_1$, decreases as the affinity of drugs to the transporter increases or as the free energy of binding, $\Delta G^0_{\mathrm{tw}(1)}$, decreases. Molecules with low affinity are thus transported more rapidly and tend to be smaller (Figure 18.7).

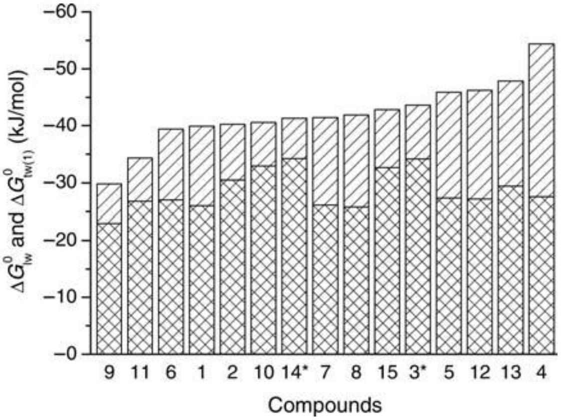

Figure 18.5 The free energy of drug binding from water to the activating binding region of P-gp ($\Delta G^0_{\mathrm{tw}(1)}$) (hatched and cross-hatched bars) in comparison to the free energy of drug partitioning from water to the lipid membrane (ΔG^0_{lw}) (cross-hatched bars). The difference between $\Delta G^0_{\mathrm{tw}(1)}$ and ΔG^0_{lw} represents the free energy of drug binding from lipid membrane to the transporter ($\Delta G^0_{\mathrm{tl}(1)}$) (hatched bar). Amitriptyline (1), chlorpromazine (2), *cis*-flupenthixol (3), cyclosporin A (4), daunorubicin (5), dibucaine (6), diltiazem (7), glivec (8), lidocaine (9) progesterone (10), promazine (11), verapamil (12), reserpine (13), trifluoperazine (14), and trifluopromazine (15) measured at pH 7.4 and 37 °C. (Data are taken from Ref. [32].)

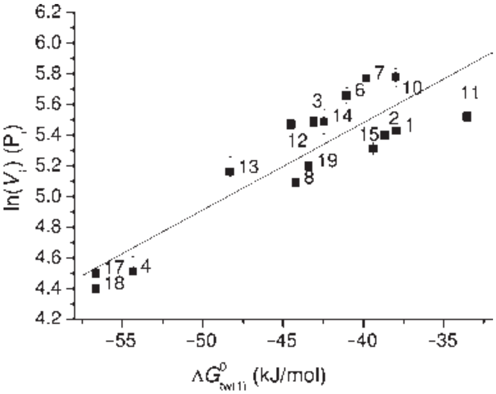

Figure 18.6 Correlation between the logarithm of the maximum P-gp activity (ln V_1) obtained from phosphate release measurements at pH 7.0 (37 °C) and the free energy of drug binding from water to transporter ($\Delta G^0_{\mathrm{tw}(1)}$). The maximum activity of P-gp (V_1) is expressed as a percentage of the basal rates taken as 100%. Data are presented as average of two to 15 measurements. The solid line is a linear regression to the data with a slope 0.06 ± 0.01 and an intercept 7.76 ± 0.34 ($R^2 = 0.79$). Compounds are as in Figure 18.5 (1–15); daunorubicin (5), and lidocaine (9) were excluded from the fit due to experimental problems; extra data for OC144-093 (17), PSC-833 (18), and vinblastine (19) (data taken from Ref. [35]).

In summary, ATP hydrolysis by P-gp correlates well with the intrinsic rate of substrate transport. A complete characterization of the interaction of a compound with P-gp is obtained by measuring the ATPase activity as a function of concentration. The rate of intrinsic substrate transport first increases with increasing concentration, reaches a maximum, and decreases again at high concentrations. The rate of intrinsic transport by P-gp depends not only on the substrate concentration but also on its affinity to the transporter; substrates with high affinities for P-gp are transported more slowly than those with low affinities.

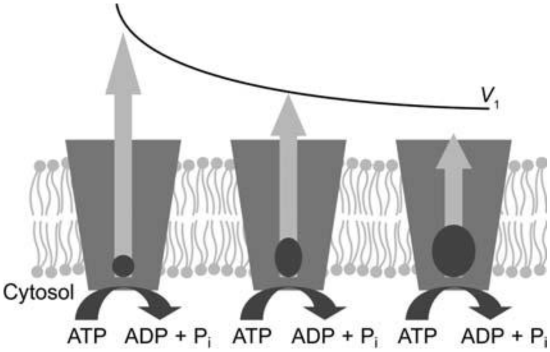

Figure 18.7 Intrinsic transport by P-gp. P-gp transports drugs at different rates. Small drugs often have a lower affinity to the transporter and are transported more rapidly (at least at low concentrations when only one drug is bound) than larger drugs with higher affinity.

18.2.2 Transport Assays

P-gp-specific transport is often assayed with confluent monolayers of polarized epithelial cells transfected with the MDR1 gene (e.g., kidney cells) using radioactively labeled compounds [46]. Instead of MDR1-transfected cells, Caco-2 cell lines have been used [47]. Caco-2 cells can express a number of different transporters, including P-gp, and thus show activities typical of all transporters. To assess transport, basolateral-to-apical flux ($J_{B\rightarrow A}$) is generally compared with apical-to-basolateral flux ($J_{A\rightarrow B}$) of a compound across the confluent cell monolayer using identical initial compound concentrations in the donor compartments (Figure 18.8) (see e.g., Ref. [48]). If P-gp or other efflux transporters are present in the basolateral membrane, the basolateral-to-apical flux ($J_{B\rightarrow A}$) is enhanced and the apical-to-basolateral flux is reduced resulting in a flux ratio

$$\frac{J_{B\rightarrow A}}{J_{A\rightarrow B}}>1. \tag{18.6}$$

As illustrated in Figure 18.8, the net flux (J) across a membrane is the sum of passive and active transport processes. To estimate the net flux across a membrane, the following simplifying assumptions are made. The net flux ($J_{B\rightarrow A}$) from the basolateral to the apical side of the membrane is assumed to be the sum of the passive flux ($\Phi_{B\rightarrow A}$) plus the active transport rate ($+V$), and the net flux from the apical to the basolateral side ($J_{A\rightarrow B}$) is the sum of the passive flux ($\Phi_{A\rightarrow B}$) less the active transport rate ($-V$):

$$J_{B\rightarrow A} = \Phi_{B\rightarrow A} + V, \tag{18.7}$$

$$J_{A\rightarrow B} = \Phi_{A\rightarrow B} - V. \tag{18.8}$$

To illustrate the role of passive influx in transport assays, we plotted the flux ratio $J_{B\rightarrow A}/J_{A\rightarrow B}$ as a function of the passive flux (Φ). Passive flux varies enormously

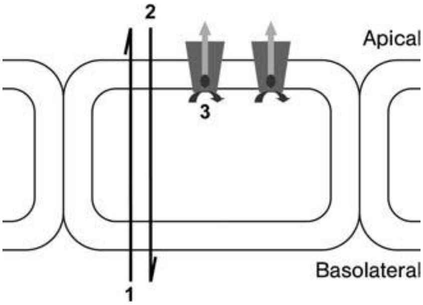

Figure 18.8 Transport processes across a confluent cell monolayer of P-gp-expressing cells. The flux from the basolateral to the apical side of the membrane ($J_{B\rightarrow A}$) is the sum of the passive flux ($\Phi_{B\rightarrow A}$) (black arrow 1) plus an active efflux component (V) (light gray arrow 3). The flux from the apical to the basolateral side of the membrane ($J_{A\rightarrow B}$) is the sum of the passive flux ($\Phi_{A\rightarrow B}$) (black arrow 2) and an active efflux component ($-V$) (light gray arrow 3).

(from about 10^{12} to less than 10^6 molecules/s/cell). It decreases exponentially with increasing cross-sectional area (A_D) and the charge (pK_a) of the molecule. Further factors that influence the passive flux are the lateral membrane packing density (π_M), which depends on the lipid composition, and the pH of the solution [49]. In contrast, active efflux varies by less than one order of magnitude for a given cell line (see Figure 18.6) [50]. Major factors influencing active efflux are the drug concentration (Equation 18.1) (see also Ref. [51]) and the expression level of P-gp.

Passive flux $|\Phi|$ is orders of magnitude higher than active efflux $|V|$ for small drugs (intrinsic substrates) and is therefore "masked" in assays (Figure 18.9). As the passive flux tends to decrease strongly with molecular size and the intrinsic transport tends to decrease only slightly [49, 50], membrane-specific limiting cross-sectional areas (A_D) can be defined for drug permeation and have been reported for the blood–brain barrier [52] and the intestinal barrier [53].

In summary, assays with confluent cell monolayers reveal the net flux (J) that is the result of passive and active transport processes. Substrate transport is observable only if the magnitude of passive flux $|\Phi|$ is similar to that of active efflux $|V|$ but is "masked" if the passive flux is significantly higher than active efflux. Since the passive flux decreases exponentially with the cross-sectional area (A_D) and the ionization status (pK_a), these two parameters dominate the flux, or the apparent transport, across a cell

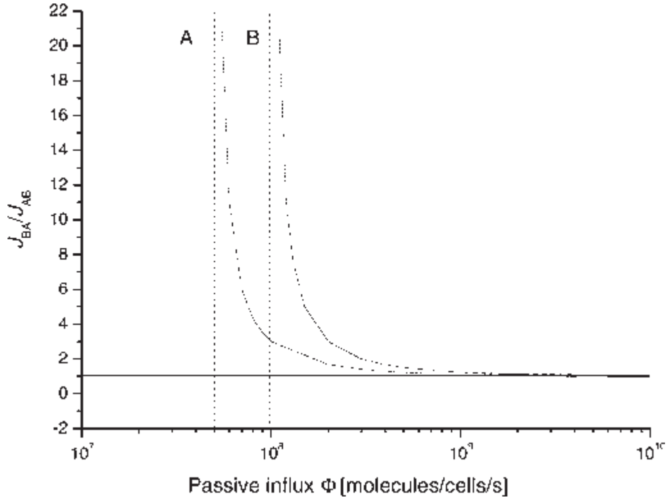

Figure 18.9 Transport across confluent cell monolayers depends on many parameters (see Section 20.2.2). The quotient $J_{B\rightarrow A}/J_{A\rightarrow B}$ is plotted as a function of the passive influx (Φ). It was assumed that $\Phi_{B\rightarrow A}$ and $\Phi_{A\rightarrow B}$ are identical. Active export was taken as $V = 5 \times 10^7$ molecules/cell/s (A) or $V = 1 \times 10^8$ molecules/cell/s (B). If the passive flux $|\Phi|$ is high, active transport by P-gp is masked, leading to flux ratio $J_{B\rightarrow A}/J_{A\rightarrow B}$ close to 1; if the flux $|\Phi|$ is similar to the rate of active transport $|V|$, then $J_{B\rightarrow A}/J_{A\rightarrow B} > 1$; if the passive flux $|\Phi|$ is very low (lower limits indicated by dotted lines), the quotient cannot be determined experimentally because all substrate molecules that then permeate the plasma membrane are exported again and none are able to permeate into the cytosol. (Data are taken from Ref. [49].)

layer. Compounds that are determined as substrates in transport assays are called apparent substrates to distinguish them from intrinsic substrates determined by ATPase assays.

18.2.3 Competition Assays

The most frequently used competition assays are the calcein–AM [54–56] and the cytotoxicity assays (e.g., performed with doxorubicin) [41, 42], carried out with living cells expressing P-gp. Calcein–AM as well as doxorubicin is used as reference substrates for P-gp as their concentrations in the cytosol are reduced by its action. If a second compound that competes for the transporter P-gp is added to the cell, the reference substrates accumulate above control levels in the cytosol in a concentration-dependent manner up to a maximal value. The more effective the competing agent, the larger is the increase in cytosolic concentration of the reference substrates. In the case of the calcein–AM that is hydrolyzed as soon as it reaches the cytosol, the fluorescent hydrolysis product, calcein, is assayed by fluorescence spectroscopy. The inhibitory potencies of compounds are measured as half-inhibitory constants, K_I.

In the case of the cytotoxin doxorubicin, its concentration in the cytosol is generally estimated via the incumbent cytotoxicity, which reduces cell growth rates. Defined concentrations of doxorubicin inhibit growth of the MDR-expressing cells by 50 or 20% (IC_{50} or IC_{20} values). When P-gp is inhibited by a competing drug, lower concentrations of doxorubicin are required to cause the same level of toxicity; hence, drug potencies are expressed as effect–concentration ratio or MDR ratio.

$$\mathrm{MDR\,ratio} = \frac{\mathrm{IC}_{50}(\text{cytotoxic drug alone})}{\mathrm{IC}_{50}(\text{cytotoxic drug} + \text{modulator})}. \qquad (18.9)$$

In summary, competition assays yield information on the affinity of a drug to the transporter relative to the affinity of a reference substrate, for example, calcein–AM or doxorubicin. The higher the affinity of a drug for P-gp (or the more negative the free energy of binding to P-gp), the greater is the ability to suppress efflux of a reference substrate.

18.3 Predictive *In Silico* Models

Different prediction models have been reported including (i) pharmacophore models that take into account structural features, (ii) linear discriminant models that do not consider structural features, (iii) a modular-binding approach, and (iv) rule-based approaches. The focus of the following discussion is to identify the most important descriptors in the different approaches and relate them to the physicochemical parameters determined in the different P-gp assays.

18.3.1 Introduction to Structure–Activity Relationship

Structure–activity relationship (SAR) studies are based on the assumption that similar molecules elicit similar activities in a lock/key-type manner. Quantitative structure–activity relationships (QSARs) correlate the extent of a change in a biological response (e.g., activity) elicited by a specific compound with its physicochemical and/or its structural properties,

$$\text{activity} = f(\text{physical} - \text{ chemical parameters and/or structural properties}). \tag{18.10}$$

Individual physicochemical parameters are also called molecular descriptors, and the net structural properties describe a pharmacophore. According to IUPAC, a pharmacophore is defined as the ensemble of steric and electronic features that is necessary to ensure interactions with a specific biological target and that induces or blocks its biological response.

QSAR modeling has been successfully applied for elucidating the stereochemical features relevant for the function of small ligands binding to an acceptor in a lock/key-type mechanism in aqueous solution. For such a process, the following assumptions are appropriate: (i) the modeled conformation is the bioactive one (i.e., the pharmacophore); (ii) the binding site and/or mode is the same for all modeled compounds; (iii) interactions between the drug and the binding site are mainly due to enthalpic processes such as van der Waals interactions; and (iv) solvent or membrane effects are negligible [57].

Extending QSAR models to P-gp is nontrivial since no high-resolution structure of P-gp is available yet [23]. In addition, the binding site or binding region is not well defined, but is most likely large and flexible [58]; as it is located in the interior membrane, electrostatic and hydrogen-bond interactions are more specific and stronger than van der Waals interactions, due to the low dielectric constant of the environment [32]. Substrates have extremely diverse and flexible structure [59]. A further difficulty arises from the complexity of the biological data used as the basis for QSAR or SAR. Generally, data from transport or competition assays are used, although the underlying principles are more complex than in ATPase assays.

18.3.2 3D-QSAR Pharmacophore Models

Examples of pharmacophore models are discussed in this chapter. The earlier models are primarily based on competition assays whereas the newer models are rather based on transport assays.

On the basis of competition assays, Pajeva and Wiese [60] proposed pharmacophores with different interaction points with the transporter using the program GASP (Tripos software). GASP elucidates pharmacophore models while allowing ligand flexibility, without requiring prior knowledge of pharmacophore elements or constraints. The pharmacophore model consists of two hydrophobic points, three hydrogen-bond acceptor points, and one hydrogen-bond donor point. In this model, the affinity of the substrate to the transporter depends on the number of pharmacophore points per substrate and thus allows for variable binding to the transporter.

Ekins *et al*. built QSAR models using Catalyst software to rank and predict inhibitors for P-gp substrate transport. In their first attempt, four different pharmacophores were derived from the analysis of inhibitors of digoxin transport, vinblastine binding, or intracellular accumulation of vinblastine and calcein [61]. These data were then combined with experiments using verapamil as inhibitor and led to the construction of a unique pharmacophore consisting of one hydrogen-bond acceptor, one aromatic ring, and two hydrophobic centers [62].

Langer *et al*. [63] used a training set of propafenone-type MDR modulators tested with a daunorubicin competition assay and developed a pharmacophore for P-gp inhibition using Catalyst software. The pharmacophore features identified by this model were one hydrogen-bond acceptor, one hydrophobic area, two aromatic centers, and (iv) one positively ionizable group.

On the basis of transport data, Penzotti *et al*. [64] constructed and validated a model for recognizing P-gp transport substrates. The model consists of an ensemble of 100 two-, three-, and four-point pharmacophores. The “point pharmacophores” were selected from the following descriptors: hydrogen-bond acceptors, hydrogen-bond donors, hydrophobic centers, negative and positive charges, aromatic groups, and the associated six interfeature distances. Together, these were assumed to describe the various chemotypes that interact with P-gp.

Cianchetta *et al*. [65] selected compounds from Caco-2 cell transport assays and investigated them for their ability to inhibit calcein–AM efflux. Using GRIND (*gr*id *in*dependent *d*escriptors), they then proposed a unique pharmacophore containing two hydrophobic groups separated by 16.5 Å and two hydrogen-bond acceptor groups separated by 11.5 Å. Moreover, they observed that the dimensions of the molecule play a significant role for substrate transport.

Applying supervised machine learning techniques, Li *et al*. [66] proposed a model that differentiates substrates from nonsubstrates of P-gp based on a simple tree using nine distinct pharmacophores. Four-point 3D pharmacophores were employed to increase the amount of shape information and resolution and possessed the ability to distinguish chirality. Relevant features were hydrogen-bond acceptors, hydrophobicity indices, and a cationic charge.

18.3.3
Linear Discriminant Models

Linear discriminant analysis (LDA) is used in statistics and machine learning methods to find the best linear combination of descriptors that distinguish two or more classes of objects or events, and, in the present case, to distinguish between substrates and nonsubstrates of P-gp. A linear classifier achieves this by making a classification decision based on the value of the linear combination of descriptors.

The linear discriminant models applied to P-gp [67, 68] are essentially based on data from transport assays. Several methods such as the support vector machine approach (SVM) [69], principle component analysis (PCA) [69], partial least square discriminant analysis (PLS-DA) [70, 71], or the machine learning approach (neural network) [70, 72] are derived from related principles. Svetnik *et al.* [73] used boosting tree or bagging tree techniques for P-gp substrate classification, each of which consists of a sequence of about 100 tree classifiers based on 1522 binarized atom pair descriptors. The investigation was performed with the transport data set of Penzotti *et al.* [64] and it revealed that when chlorine and fluorine substitutions enhanced permeability [74] they also lowered the tendency of a compound to be effluxed by P-gp. All these procedures start with a very large number of general descriptors that are then reduced to lower numbers of essential ones; size- and charge-related parameters dominate again.

18.3.4 Modular Binding Approach

To get the broadest possible information on the nature of P-gp/substrate interactions, we chose data from all types of assays and analyzed 3D structures by visual inspection. Chemically very diverse compounds known to interact (or not to interact) with P-gp were analyzed, with their molecular weight ranging from approximately 250 to 1250. The only recognition elements found in all compounds interacting with P-gp were hydrogen-bond acceptor groups. Patterns with two hydrogen-bond acceptors with a spatial separation of 2.5 ± 0.3 Å (type I units) were observed in all P-gp substrates. In addition, patterns with three hydrogen-bond acceptor groups with a spatial separation of the outer two acceptor groups of 4.6 ± 0.6 Å or two hydrogen-bond acceptor groups with a spatial separation of 4.6 ± 0.6 Å (type II units) were observed in many substrates and all inducers of P-gp [21, 26]. Hydrogen-bond acceptor patterns were therefore suggested to serve as binding modules interacting with the hydrogen-bond donor-rich transmembrane domains of P-gp.

In lipid environments, exhibiting a low dielectric constant, the hydrogen-bonding interactions are stronger and more specific than van der Waals interactions. It was therefore suggested that the measured total free energy of binding of a drug from the lipid membrane to the transporter $\Delta G^0_{\mathrm{tl}(1)}$ is the sum of the free energies, $\Delta G^0_{\mathrm{H}i}$, of the individual hydrogen bonds formed between the substrate and the transporter [26, 75]

$$\Delta G^0_{\mathrm{tl}(1)} \approx \sum_{i=1}^{n} \Delta G^0_{\mathrm{H}i}. \tag{18.11}$$

To test this hypothesis, the experimentally determined free energy of binding to P-gp, $\Delta G^0_{\mathrm{tl}(1)}$, for a given drug was divided by the number of possible hydrogen bonds formed thus yielding the free energy per hydrogen bond of $\Delta G^0_{\mathrm{H}i} \approx -2.5\ \mathrm{kJ}/mol$ as a lower limit. This value is in good agreement with expectations [32]. Combining Equations 18.3 and 18.11, the free energy of binding of a substrate from water to the transporter can then be estimated as

$$\Delta G^0_{\mathrm{tw}(1)} \approx \sum_{i=1}^{n} \Delta G^0_{\mathrm{H}i} + \Delta G^0_{\mathrm{lw}}. \tag{18.12}$$

The requirement to bind to P-gp is thus the ability to partition into the inner lipid leaflet and to carry hydrogen-bond acceptor groups (arranged in type I or type II units).

18.3.5 Rule-Based Approaches

One of the first and most cited rule-based approaches is the "rule-5" by Lipinski [76], which predicts whether a compound will be absorbed from the intestinal tract, that is, cross the intestinal barrier. Although transporters are not explicitly mentioned, they play a role in intestinal absorption. "Lipinski's rule-of-5" states that, in general, a well-absorbed drug violates no more than one of the following criteria: no more than five hydrogen-bond donors (i.e., nitrogen or oxygen atoms with one or more hydrogen atoms); not more than 10 hydrogen-bond acceptors (nitrogen or oxygen atoms); molecular weight below 500 Da; and an octanol–water partition coefficient (log *P*) below 5.

An approach that is related to the "rule-of-5" was proposed by Didziapetris *et al.* [77] to predict whether a drug is a substrate for P-glycoprotein. On the basis of transcellular transport experiments, they suggested that compounds with more than eight oxygen and nitrogen atoms, molecular weights above 400 Da, and acidic $\mathrm{p}K_a$ more than 4 are likely to be P-glycoprotein substrates; compounds with less than four oxygen and nitrogen atoms, molecular weights below 400 Da, and $\mathrm{p}K_a$ less than 8 are likely to be nonsubstrates.

The cross-sectional area, A_D, of a compound oriented in an amphiphilic gradient such as the air–water or lipid–water interface has been shown to be even more reliable for permeability predictions than the molecular weight [52]. For BBB permeation, the limiting cross-sectional area, A_D, was determined as $A_D \approx 73\,\text{Å}^2$, and the limiting ionization constants, $\mathrm{p}K_a$s, for bases and acids were determined as 9 and 4, respectively [52]. For intestinal barrier permeation, the limiting cross-sectional area was assessed as $A_D \approx 100\,\text{Å}^2$, and the limiting ionization constants ($\mathrm{p}K_a$s) for bases as 9 and for acids as 2 [53]. In this approach, the role of P-gp is again implicit [49].

To predict membrane barrier permeation *in silico*, we developed an algorithm that determines the molecular axis of amphiphilicity and the cross-sectional area, A_{Dcalc}, perpendicular to this axis. Starting with the conformational ensemble of each molecule, the three-dimensional membrane-binding conformation was determined as the one with the smallest cross-sectional area, A_{DcalcM}, and the strongest amphiphilicity. The calculated cross-sectional areas, A_{DcalcM}, were then correlated with the calculated octanol–water distribution coefficients, log $D_{7.4}$, of the 55 compounds with known abilities to permeate the blood–brain barrier, to predict the probability of blood–brain barrier permeation. The limiting cross-sectional area was $A_{\text{DcalcM}} = 70\,\text{Å}^2$, and the optimal range of octanol–water distribution coefficients was $-1.4 \leq \log D_{7.4} < 5.0$. The correlation was validated with an independent set of 43 compounds with known abilities to permeate the blood–brain barrier and yielded a prediction accuracy of 86% [59].

18.4 Discussion

We now compare the main descriptors obtained from the different *in silico* models (Section 18.3) with the physicochemical parameters that allow a quantitative description of the different aspects of P-gp transport (Section 18.2).

18.4.1 Prediction of Substrate-P-gp Interactions

Most pharmacophore models for P-gp substrate prediction have similar predictive accuracies. All three-dimensional pharmacophores use similar features such as size-related parameters, hydrogen-bond acceptors, hydrophobicity parameters, aromatic rings, and ionizable groups. Thereby size- and charge-related parameters and hydrogen-bond acceptor groups were generally considered as the most relevant descriptors. Despite these agreements, the resulting pharmacophores showed no similarity at all, raising the question whether three-dimensional lock/key-type pharmacophores are appropriate for P-gp. Moreover, it is difficult to envisage how a single pharmacophore could account for the different affinities of drugs to the transporter. Variable binding affinities can, however, only be explained with modular binding approaches.

18.4.2 Prediction of ATPase Activity or Intrinsic Transport

The binding affinity of a substrate to P-gp seems to directly influence the rate of transport as seen in Figure 18.6. ATPase activity (or the intrinsic transport rate by P-gp) decreases with increasing binding affinities of substrates to the transporter. Rather accurate predictions are possible assuming modular binding [26, 32, 78] using Equation 18.12 [79]. The modular binding principles also allow the prediction of the extent of competition between different substrates for binding P-gp (see Ref. [79]).

18.4.3 Prediction of Transport (i.e., Apparent Transport)

Pharmacophore models, linear discriminate models, and rule-based models agree with respect to the relevance of size and charge for transport by P-gp and also agree with experimental investigations [80]. This is also consistent with the analysis of net influx (Section 18.2.2), which shows that increasing size (or cross-sectional area) and/or charge of the molecule diminishes the rate of diffusion that in turn unmasks active transport rates [49]. It can thus be concluded that compounds that are large, have hydrogen-bond acceptors, and are cationic are likely to be apparent substrates for P-gp. Large molecules with hydrogen-bond acceptor groups are also at risk when being effluxed by transporters with overlapping substrate specificities.

18.4.4 Prediction of Competition

Hydrogen-bond acceptor groups and size-related descriptors are also dominant in QSAR models predicting P-gp inhibitors on the basis of competition assays. However, substrate–transporter interactions are most likely not determined by the size of the substrate as such, but rather by a concomitant increase in residues that can undergo specific interactions with the transmembrane sequences of P-gp (for details see Ref. [32]). A good estimate of the competitive potential of compounds is possible on the basis of the total hydrogen-bond acceptor strength alone, if the lipid-binding properties of the compounds are comparable [35, 81].

18.4.5 Conclusions

Considerable effort has been put into the development of *in silico* methods that predict apparent transport by P-gp and competition for P-gp-binding sites. The models cited include pharmacophore, linear discriminant and rule-based models, respectively, and a modular binding approach. The first three *in silico* models are strongly based on transport assays that are very complex and determine apparent rather than intrinsic transport rates by P-gp. Although the different models discussed are very diverse, they nevertheless agree with respect to the relevance of size and charge of the molecule for apparent transport; they also agree on the relevance of hydrogen-bond acceptor groups as recognition elements for P-gp. However, they disagree greatly with respect to the proposed structural arrangement of recognition elements. At present, only the modular binding approach that considers small hydrogen-bond acceptor patterns as binding modules allows modeling the different binding affinities of the enormously diverse intrinsic substrates for P-gp. Translation of the results obtained from these investigations into the synthesis of new ligands or into an optimization of known ligands could lead to a reduction of MDR.

References

1 Gottesman, M.M. and Ambudkar, S.V. (2001) Overview: ABC transporters and human disease. *Journal of Bioenergetics and Biomembranes*, **33**, 453–458.

2 Juranka, P.F., Zastawny, R.L. and Ling, V. (1989) P-Glycoprotein: multidrug-resistance and a superfamily of membrane-associated transport proteins. *FASEB Journal*, **3**, 2583–2592.

3 Cordon-Cardo, C., O'Brien, J.P., Boccia, J., Casals, D., Bertino, J.R. and Melamed, M.R. (1990) Expression of the multidrug resistance gene product (P-glycoprotein) in human normal and tumor tissues. *The Journal of Histochemistry and Cytochemistry*, **38**, 1277–1287.

4 Mukhopadhyay, T., Batsakis, J.G. and Kuo, M.T. (1988) Expression of the mdr (P-glycoprotein) gene in Chinese hamster digestive tracts. *Journal of the National Cancer Institute*, **80**, 269–275.

5 Thiebaut, F., Tsuruo, T., Hamada, H., Gottesman, M.M., Pastan, I. and Willingham, M.C. (1989)

Immunohistochemical localization in normal tissues of different epitopes in the multidrug transport protein P170: evidence for localization in brain capillaries and crossreactivity of one antibody with a muscle protein. *The Journal of Histochemistry and Cytochemistry*, **37**, 159–164.

6 Smit, J.W., Huisman, M.T., van Tellingen, O., Wiltshire, H.R. and Schinkel, A.H. (1999) Absence or pharmacological blocking of placental P-glycoprotein profoundly increases fetal drug exposure. *The Journal of Clinical Investigation*, **104**, 1441–1447.

7 Holash, J.A., Harik, S.I., Perry, G. and Stewart, P.A. (1993) Barrier properties of testis microvessels. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 11069–11073.

8 Sarkadi, B., Muller, M. and Hollo, Z. (1996) The multidrug transporters – proteins of an ancient immune system. *Immunology Letters*, **54**, 215–219.

9 Calcabrini, A., Meschini, S., Stringaro, A., Cianfriglia, M., Arancia, G. and Molinari, A. (2000) Detection of P-glycoprotein in the nuclear envelope of multidrug resistant cells. *The Histochemical Journal*, **32**, 599–606.

10 Endicott, J.A. and Ling, V. (1989) The biochemistry of P-glycoprotein-mediated multidrug resistance. *Annual Review of Biochemistry*, **58**, 137–171.

11 McClean, S., Hosking, L.K. and Hill, B.T. (1993) Dominant expression of multiple drug resistance after *in vitro* X-irradiation exposure in intraspecific Chinese hamster ovary hybrid cells. *Journal of the National Cancer Institute*, **85**, 48–53.

12 Uchiumi, T., Kohno, K., Tanimura, H., Matsuo, K., Sato, S., Uchida, Y. and Kuwano, M. (1993) Enhanced expression of the human multidrug resistance 1 gene in response to UV light irradiation. *Cell Growth & Differentiation: The Molecular Biology Journal of the American Association for Cancer Research*, **4**, 147–157.

13 Chin, K.V., Tanaka, S., Darlington, G., Pastan, I. and Gottesman, M.M. (1990) Heat shock and arsenite increase expression of the multidrug resistance (MDR1) gene in human renal carcinoma cells. *The Journal of Biological Chemistry*, **265**, 221–226.

14 Sauna, Z.E., Kim, I.W. and Ambudkar, S.V. (2007) Genomics and the mechanism of P-glycoprotein (ABCB1). *Journal of Bioenergetics and Biomembranes*, **39**, 481–487.

15 Gottesman, M.M., Fojo, T. and Bates, S.E. (2002) Multidrug resistance in cancer: role of ATP-dependent transporters. *Nature Reviews Cancer*, **2**, 48–58.

16 Van Bambeke, F.B.E. and Tulkens, P.M. (2000) Antibiotic efflux pumps. *Biochemical Pharmacology*, **60**, 457–470.

17 Borst, P. and Ouellette, M. (1995) New mechanisms of drug resistance in parasitic protozoa. *Annual Review of Microbiology*, **49**, 427–460.

18 Wolfger, H., Mamnun, Y.M. and Kuchler, K. (2001) Fungal ABC proteins: pleiotropic drug resistance, stress response and cellular detoxification. *Research in Microbiology*, **152**, 375–389.

19 Dean, M. and Annilo, T. (2005) Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annual Review of Genomics and Human Genetics*, **6**, 123–142.

20 Haimeur, A., Conseil, G., Deeley, R.G. and Cole, S.P. (2004) The MRP-related and BCRP/ABCG2 multidrug resistance proteins: biology, substrate specificity and regulation. *Current Drug Metabolism*, **5**, 21–53.

21 Seelig, A., Blatter, X.L. and Wohnsland, F. (2000) Substrate recognition by P-glycoprotein and the multidrug resistance-associated protein MRP1: a comparison. *International Journal of Clinical Pharmacology and Therapeutics*, **38**, 111–121.

22 Matsson, P., Englund, G., Ahlin, G., Bergstrom, C.A., Norinder, U. and Artursson, P. (2007) A global drug

inhibition pattern for the human ATP-binding cassette transporter breast cancer resistance protein (ABCG2). *The Journal of Pharmacology and Experimental Therapeutics*, **323**, 19–30.

23 Rosenberg, M.F., Callaghan, R., Modok, S., Higgins, C.F. and Ford, R.C. (2005) Three-dimensional structure of P-glycoprotein: the transmembrane regions adopt an asymmetric configuration in the nucleotide-bound state. *The Journal of Biological Chemistry*, **280**, 2857–2862.

24 Dawson, R.J. and Locher, K.P. (2006) Structure of a bacterial multidrug ABC transporter. *Nature*, **443**, 180–185.

25 Dawson, R.J. and Locher, K.P. (2007) Structure of the multidrug ABC transporter Sav1866 from *Staphylococcus aureus* in complex with AMP-PNP. *FEBS Letters*, **581**, 935–938.

26 Seelig, A. (1998) A general pattern for substrate recognition by P-glycoprotein. *European Journal of Biochemistry*, **251**, 252–261.

27 Dey, S., Ramachandra, M., Pastan, I., Gottesman, M.M. and Ambudkar, S.V. (1997) Evidence for two nonidentical drug-interaction sites in the human P-glycoprotein. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 10594–10599.

28 Martin, C., Berridge, G., Higgins, C.F., Mistry, P., Charlton, P. and Callaghan, R. (2000) Communication between multiple drug binding sites on P-glycoprotein. *Molecular Pharmacology*, **58**, 624–632.

29 Raviv, Y., Pollard, H.B., Bruggemann, E.P., Pastan, I. and Gottesman, M.M. (1990) Photosensitized labeling of a functional multidrug transporter in living drug-resistant tumor cells. *The Journal of Biological Chemistry*, **265**, 3975–3980.

30 Chen, Y., Pant, A.C. and Simon, S.M. (2001) P-glycoprotein does not reduce substrate concentration from the extracellular leaflet of the plasma membrane in living cells. *Cancer Research*, **61**, 7763–7769.

31 Shapiro, A.B. and Ling, V. (1997) Extraction of Hoechst 33342 from the cytoplasmic leaflet of the plasma membrane by P-glycoprotein. *European Journal of Biochemistry*, **250**, 122–129.

32 Gatlik-Landwojtowicz, E., Aanismaa, P. and Seelig, A. (2006) Quantification and characterization of P-glycoprotein–substrate interactions. *Biochemistry*, **45**, 3020–3032.

33 Litman, T., Zeuthen, T., Skovsgaard, T. and Stein, W.D. (1997) Structure–activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. *Biochimica et Biophysica Acta*, **1361**, 159–168.

34 Al-Shawi, M.K., Polar, M.K., Omote, H. and Figler, R.A. (2003) Transition state analysis of the coupling of drug transport to ATP hydrolysis by P-glycoprotein. *The Journal of Biological Chemistry*, **278**, 52629–52640.

35 Aanismaa, P. and Seelig, A. (2007) P-Glycoprotein kinetics measured in plasma membrane vesicles and living cells. *Biochemistry*, **46**, 3394–3404.

36 Garrigues, A., Nugier, J., Orlowski, S. and Ezan, E. (2002) A high-throughput screening microplate test for the interaction of drugs with P-glycoprotein. *Analytical Biochemistry*, **305**, 106–114.

37 Broxterman, H.J., Pinedo, H.M., Kuiper, C.M., Schuurhuis, G.J. and Lankelma, J. (1989) Glycolysis in P-glycoprotein-overexpressing human tumor cell lines. Effects of resistance-modifying agents. *FEBS Letters*, **247**, 405–410.

38 Gatlik-Landwojtowicz, E., Aanismaa, P. and Seelig, A. (2004) The rate of P-glycoprotein activation depends on the metabolic state of the cell. *Biochemistry*, **43**, 14840–14851.

39 McConnell, H.M., Owicki, J.C., Parce, J.W., Miller, D.L., Baxter, G.T., Wada, H.G. and Pitchford, S. (1992) The cytosensor microphysiometer: biological applications of silicon technology. *Science*, **257**, 1906–1912.

40 Polli, J.W., Wring, S.A., Humphreys, J.E., Huang, L., Morgan, J.B., Webster, L.O. and Serabjit-Singh, C.S. (2001) Rational use of *in vitro* P-glycoprotein assays in drug discovery. *The Journal of Pharmacology and Experimental Therapeutics*, **299**, 620–628.

41 Ford, J.M., Prozialeck, W.C. and Hait, W.N. (1989) Structural features determining activity of phenothiazines and related drugs for inhibition of cell growth and reversal of multidrug resistance. *Molecular Pharmacology*, **35**, 105–115.

42 Toffoli, G., Simone, F., Corona, G., Raschack, M., Cappelletto, B., Gigante, M. and Boiocchi, M. (1995) Structure–activity relationship of verapamil analogs and reversal of multidrug resistance. *Biochemical Pharmacology*, **50**, 1245–1255.

43 Ambudkar, S.V., Cardarelli, C.O., Pashinsky, I. and Stein, W.D. (1997) Relation between the turnover number for vinblastine transport and for vinblastine-stimulated ATP hydrolysis by human P-glycoprotein. *The Journal of Biological Chemistry*, **272**, 21160–21166.

44 Saparov, S.M., Antonenko, Y.N. and Pohl, P. (2006) A new model of weak acid permeation through membranes revisited: does Overton still rule? *Biophysical Journal*, **90** (11), L86–L88.

45 Omote, H. and Al-Shawi, M.K. (2002) A novel electron paramagnetic resonance approach to determine the mechanism of drug transport by P-glycoprotein. *The Journal of Biological Chemistry*, **277**, 45688–45694.

46 Schinkel, A.H., Wagenaar, E., van Deemter, L., Mol, C.A. and Borst, P. (1995) Absence of the mdr1a P-glycoprotein in mice affects tissue distribution and pharmacokinetics of dexamethasone, digoxin, and cyclosporin A. *The Journal of Clinical Investigation*, **96**, 1698–1705.

47 Anderle, P., Niederer, E., Rubas, W., Hilgendorf, C., Spahn-Langguth, H., Wunderli-Allenspach, H., Merkle, H.P. and Langguth, P. (1998) P-Glycoprotein (P-gp) mediated efflux in Caco-2 cell monolayers: the influence of culturing conditions and drug exposure on P-gp expression levels. *Journal of Pharmaceutical Sciences*, **87**, 757–762.

48 Schwab, D., Fischer, H., Tabatabaei, A., Poli, S. and Huwyler, J. (2003) Comparison of *in vitro* P-glycoprotein screening assays: recommendations for their use in drug discovery. *Journal of Medicinal Chemistry*, **46**, 1716–1725.

49 Seelig, A. (2007) The role of size and charge for blood–brain barrier permeation of drugs and fatty acids. *Journal of Molecular Neuroscience*, **33**, 32–41.

50 Seelig, A. and Gatlik-Landwojtowicz, E. (2005) Inhibitors of multidrug efflux transporters: their membrane and protein interactions. *Mini Reviews in Medicinal Chemistry*, **5**, 135–151.

51 Hochman, J.H., Yamazaki, M., Ohe, T. and Lin, J.H. (2002) Evaluation of drug interactions with P-glycoprotein in drug discovery: *in vitro* assessment of the potential for drug–drug interactions with P-glycoprotein. *Current Drug Metabolism*, **3**, 257–273.

52 Fischer, H., Gottschlich, R. and Seelig, A. (1998) Blood–brain barrier permeation: molecular parameters governing passive diffusion. *The Journal of Membrane Biology*, **165**, 201–211.

53 Fischer, H., Seelig, A., Chou, R.C. and van de Waterbeemd, J. (1997) The difference between the diffusion through the blood–brain barrier and the gastro-intestinal membrane. 4th International Conference on Drug Absorption, Edinborough.

54 Hollo, Z., Homolya, L., Davis, C.W. and Sarkadi, B. (1994) Calcein accumulation as a fluorometric functional assay of the multidrug transporter. *Biochimica et Biophysica Acta*, **1191**, 384–388.

55 Essodaigui, M., Broxterman, H.J. and Garnier-Suillerot, A. (1998) Kinetic analysis of calcein and calcein–acetoxymethylester efflux mediated by the multidrug resistance protein and P-glycoprotein. *Biochemistry*, **37**, 2243–2250.

56 Fricker, G. (2002) Drug transport across the blood–brain barrier, in *Pharmacokinetic Challenges in Drug Discovery*, Vol. 37 (eds O. Pelkonen, A. Baumann and A. Reichel), Springer, pp. 139–154.

57 Ekins, S., Waller, C.L., Swaan, P.W., Cruciani, G., Wrighton, S.A. and Wikel, J.H. (2000) Progress in predicting human ADME parameters *in silico*. *Journal of Pharmacological and Toxicological Methods*, **44**, 251–272.

58 Ekins, S., Ecker, G.F., Chiba, P. and Swaan, P.W. (2007) Future directions for drug transporter modelling. *Xenobiotica; The Fate of Foreign Compounds in Biological Systems*, **37**, 1152–1170.

59 Gerebtzoff, G. and Seelig, A. (2006) *In silico* prediction of blood–brain barrier permeation using the calculated molecular cross-sectional area as main parameter. *Journal of Chemical Information and Modeling*, **46**, 2638–2650.

60 Pajeva, I.K. and Wiese, M. (2002) Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis). *Journal of Medicinal Chemistry*, **45**, 5671–5686.

61 Ekins, S., Kim, R.B., Leake, B.F., Dantzig, A.H., Schuetz, E.G., Lan, L.B., Yasuda, K., Shepard, R.L., Winter, M.A., Schuetz, J.D., Wikel, J.H. and Wrighton, S.A. (2002) Three-dimensional quantitative structure–activity relationships of inhibitors of P-glycoprotein. *Molecular Pharmacology*, **61**, 964–973.

62 Ekins, S., Kim, R.B., Leake, B.F., Dantzig, A.H., Schuetz, E.G., Lan, L.B., Yasuda, K., Shepard, R.L., Winter, M.A., Schuetz, J.D., Wikel, J.H. and Wrighton, S.A. (2002) Application of three-dimensional quantitative structure–activity relationships of P-glycoprotein inhibitors and substrates. *Molecular Pharmacology*, **61**, 974–981.

63 Langer, T., Eder, M., Hoffmann, R.D., Chiba, P. and Ecker, G.F. (2004) Lead identification for modulators of multidrug resistance based on *in silico* screening with a pharmacophoric feature model. *Archiv der Pharmazie*, **337**, 317–327.

64 Penzotti, J.E., Lamb, M.L., Evensen, E. and Grootenhuis, P.D. (2002) A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *Journal of Medicinal Chemistry*, **45**, 1737–1740.

65 Cianchetta, G., Singleton, R.W., Zhang, M., Wildgoose, M., Giesing, D., Fravolini, A., Cruciani, G. and Vaz, R.J. (2005) A pharmacophore hypothesis for P-glycoprotein substrate recognition using GRIND-based 3D-QSAR. *Journal of Medicinal Chemistry*, **48**, 2927–2935.

66 Li, W.X., Li, L., Eksterowicz, J., Ling, X.B. and Cardozo, M. (2007) Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. *Journal of Chemical Information and Modeling*, **47**, 2429–2438.

67 Gombar, V.K., Polli, J.W., Humphreys, J.E., Wring, S.A. and Serabjit-Singh, C.S. (2004) Predicting P-glycoprotein substrates by a quantitative structure–activity relationship model. *Journal of Pharmaceutical Sciences*, **93**, 957–968.

68 Cabrera, M.A., Gonzalez, I., Fernandez, C., Navarro, C. and Bermejo, M. (2006) A topological substructural approach for the prediction of P-glycoprotein substrates. *Journal of Pharmaceutical Sciences*, **95**, 589–606.

69 Xue, Y., Yap, C.W., Sun, L.Z., Cao, Z.W., Wang, J.F. and Chen, Y.Z. (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *Journal of Chemical Information and Computer Sciences*, **44**, 1497–1505.

70 Crivori, P., Reinach, B., Pezzetta, D. and Poggesi, I. (2006) Computational models for identifying potential P-glycoprotein substrates and inhibitors. *Molecular Pharmaceutics*, **3**, 33.

71 Li, Y., Wang, Y.-H., Yang, L., Zhang, S.-W., Liu, C.-H. and Yang, S.-L. (2005) Comparison of steroid substrates and inhibitors of P-glycoprotein by 3D-QSAR analysis. *Journal of Molecular Structure*, **733**, 111–118.

72 Wang, Y.H., Li, Y., Yang, S.L. and Yang, L. (2005) Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *Journal of Chemical Information and Modeling*, **45**, 750–757.

73 Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R.P. and Song, Q. (2005) Boosting: an ensemble learning tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, **45**, 786–799.

74 Gerebtzoff, G., Li-Blatter, X., Fischer, H., Frentzel, A. and Seelig, A. (2004) Halogenation of drugs enhances membrane binding and permeation. *Chembiochem: A European Journal of Chemical Biology*, **5**, 676–684.

75 Ecker, G., Huber, M., Schmid, D. and Chiba, P. (1999) The importance of a nitrogen atom in modulators of multidrug resistance. *Molecular Pharmacology*, **56**, 791–796.

76 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23**, 3–25.

77 Didziapetris, R., Japertas, P., Avdeef, A. and Petrauskas, A. (2003) Classification analysis of P-glycoprotein substrate specificity. *Journal of Drug Targeting*, **11**, 391–406.

78 Sauna, Z.E., Andrus, M.B., Turner, T.M. and Ambudkar, S.V. (2004) Biochemical basis of polyvalency as a strategy for enhancing the efficacy of P-glycoprotein (ABCB1) modulators: stipiamide homodimers separated with defined-length spacers reverse drug efflux with greater efficacy. *Biochemistry*, **43**, 2262–2271.

79 Seelig, A. and Gerebtzoff, G. (2006) Enhancement of drug absorption by noncharged detergents through membrane and P-glycoprotein binding. *Expert Opinion on Drug Metabolism & Toxicology*, **2**, 733–752.

80 van de Waterbeemd, H., Camenisch, G., Folkers, G., Chretien, J.R. and Raevsky, O.A. (1998) Estimation of blood–brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. *Journal of Drug Targeting*, **6**, 151–165.

81 Seelig, A. and Landwojtowicz, E. (2000) Structure–activity relationship of P-glycoprotein substrates and modifiers. *European Journal of Pharmaceutical Sciences*, **12**, 31–40.