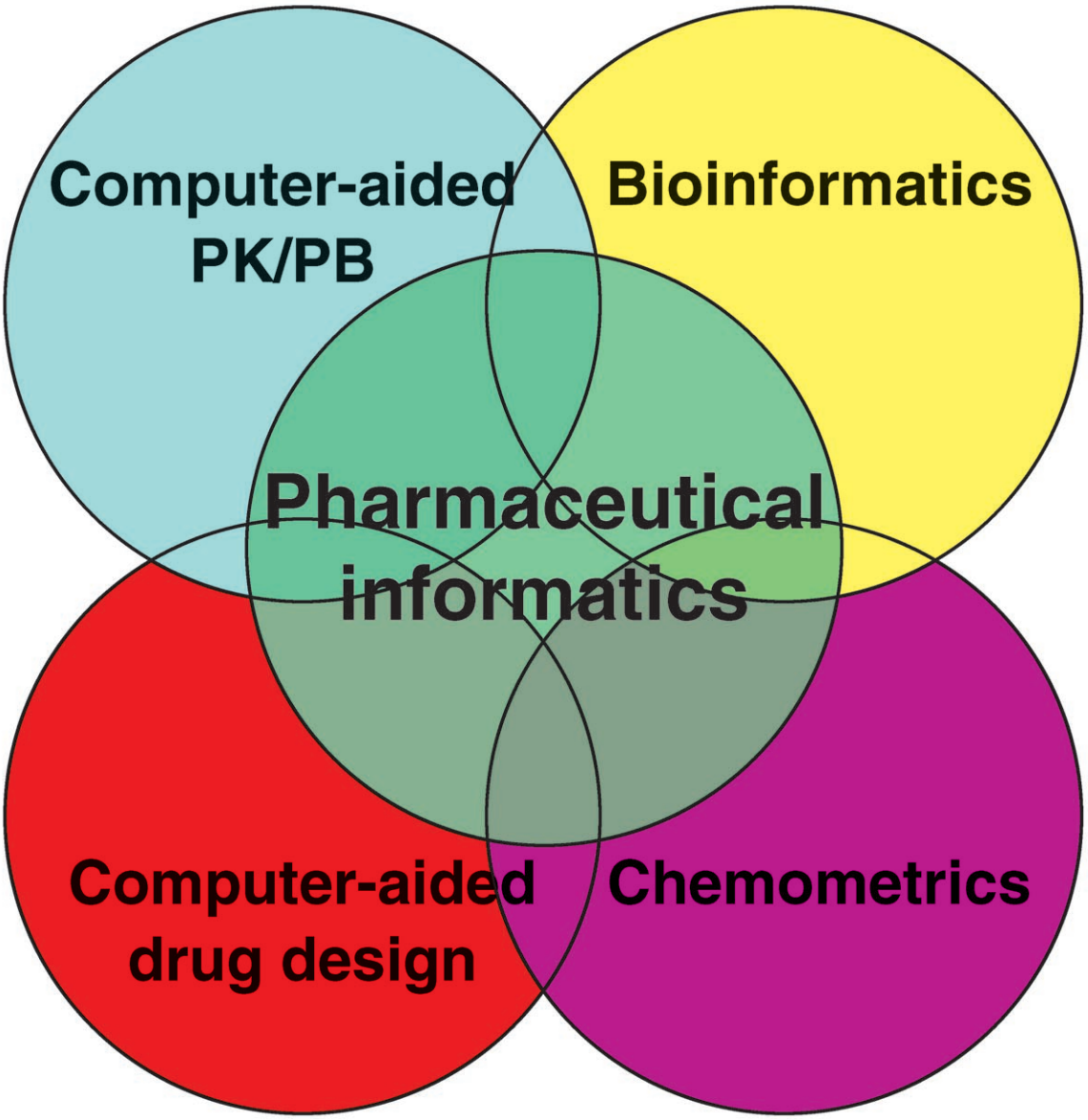
ADMET Property Predictions



 DOI: 10.1002/cmdc.200600155

DOI: 10.1002/cmdc.200600155

Prediction of ADMET Properties

Ulf Norinder*[a] and Christel A. S. Bergström[b]

This Review describes some of the approaches and techniques used today to derive in silico models for the prediction of ADMET properties. The article also discusses some of the fundamental requirements for deriving statistically sound and predictive ADMET relationships as well as some of the pitfalls and problems encountered during these investigations. It is the intension of the authors to make the reader aware of some of the challenges involved in deriving useful in silico ADMET models for drug development.

1. Introduction

By the use of genomics, proteomics, and bioinformatics, the possibility to identify and validate target proteins has recently improved. When the target has been identified, the search begins for a pharmacophore—a structural fragment that binds to the target and exerts a given effect with an acceptable therapeutic potency. After finding such a structure, the lead optimization is initiated. Computational chemistry (CC) and high-throughput screening (HTS) is used to synthesize new compounds and optimize them with regard to increased potency. The lead optimization is performed in cycles, and in the end the leads with the highest potency might be rather structurally diverse from the starting structure. The chemical library obtained can comprise several thousands of new structures. The synthesized library is experimentally examined for its developability through the use of rapid experimental techniques that measure such factors as stability, solubility, permeability, and toxicity. After these determinations, one to two candidate drugs (CD) are selected from the library for further development (Figure 1).

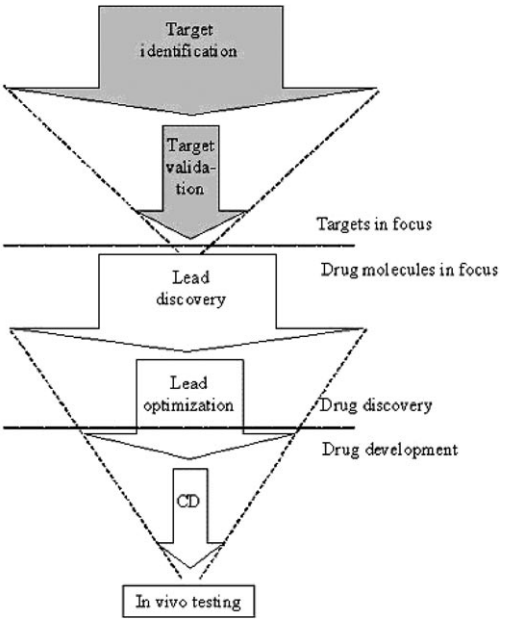


Figure 1. From target identification to candidate drug (CD): Target identification and validation are followed by lead discovery and lead optimization. The lead-optimization process is performed in cycles, and in the end of the lead-optimization process the developability of the compounds is traditionally investigated.

The increase in the number of new structures generated each year has not resulted in the expected increase in the number of marketed new drugs. This has been attributed, among other factors, to poor pharmacokinetic (PK) properties of the CDs.[1] Hence, reliable screening filters for factors such as absorption, distribution, metabolism, elimination/excretion, and toxicity (ADMET) are highly desirable.[2–4] Indeed, the effort that has been invested in the development of experimental absorption filters such as cell monolayers for permeability determinations[5,6] and the turbidimetric method for solubility measurements,[7] has lately resulted in a decrease in the attrition rate related to PK properties (Figure 2).[8] However, to allow an ADMET analysis of computationally designed druglike molecules prior to their chemical synthesis, computer-based filters for predicting PK properties are needed.

Additionally, in today's pharmaceutical research new challenges have been added for which additional considerations must be taken with respect to toxicological effects, such as avoiding interactions with hERG (human Ether-a-go-go-Related Gene) as well as potential interactions with cytochrome P450, related to avoidance of Phase 1 metabolism. A particular problem associated with predictions of toxicological effects is the lack of one well-defined and measurable target (endpoint) for which the same mechanism is involved in giving rise to the observed effects. On the contrary, even fairly similar compounds may exert their toxicity through different mechanisms.

From a development perspective, one of the first properties to evaluate is gastrointestinal absorption, as the extent to which a drug is absorbed through the intestine will determine

[a] *Prof. U. Norinder*
AstraZeneca Research and Development Södertälje
151 85 Södertälje (Sweden)
Fax: (+46) 8-553-240-42
E-mail: ulf.norinder@astrazeneca.com

[b] *Dr. C. A. S. Bergström*
Center of Pharmaceutical Informatics, Department of Pharmacy
Uppsala University, BMC P.O. Box 580, 751 23 Uppsala (Sweden)

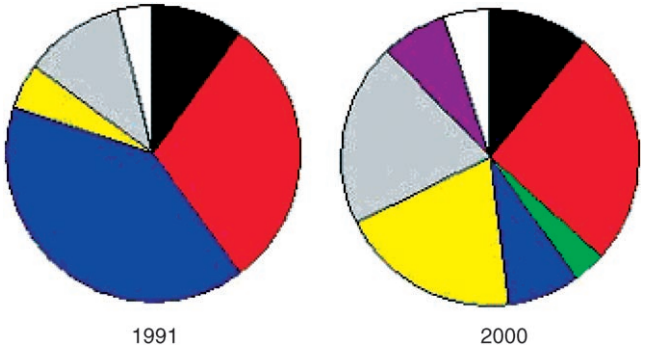


Figure 2. The following reasons for attrition in drug development in 1991 and 2000 were observed: clinical safety (black), efficacy (red), formulation (green), PK/bioavailability (blue), commercial (yellow), toxicology (gray), cost of goods (purple), and unknown/others (white). Note that formulation and cost of goods were only observed as reasons for attrition in 2000 and not 1991. Further, PK/bioavailability profiles of new drugs were largely improved during this decade. Finally, commercial reasons for attrition were over threefold greater in 2000 than in 1991.

if it is possible to administer the drug in the form of an oral dosage. This formulation is the most convenient, as it allows the patient to take the medication independently. Two of the main factors that influence intestinal absorption are the solubility of the compound in the gastrointestinal fluid and the permeability of the compound through the intestinal wall. The solubility will restrict absorption if the oral dose given is not soluble in 250 mL within the pH interval relevant in the gastrointestinal tract (pH 1 in the stomach and up to pH 8 in the colon).[9] Permeability will restrict absorption if the permeability coefficient through the enterocytes is low, thus leading to the transport of only a fraction of the compound in solution across the epithelium during the transit time in the small intestine. Both solubility and permeability are dependent on the physicochemical properties of the molecule, but unfortunately in an opposing manner (Figure 3). For instance, lipophilicity, which is the major driving force for permeability, is one of the most restricting properties for aqueous solubility.

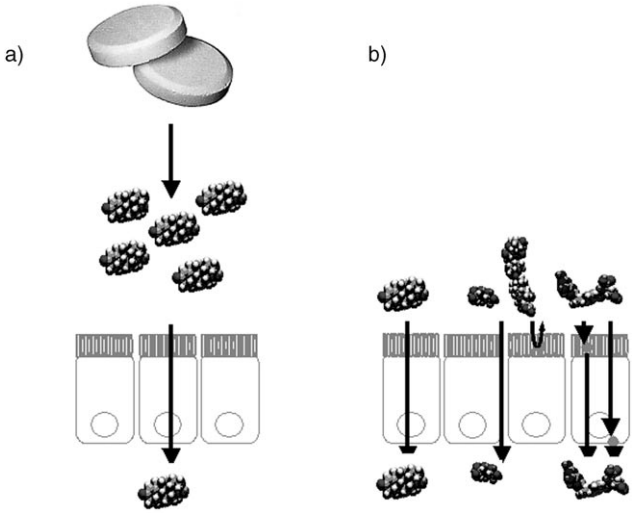
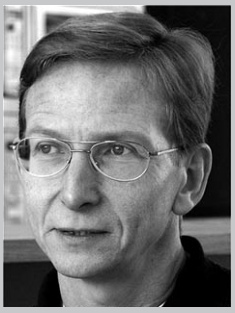


Figure 3. Molecular properties important for solubility and permeability: a) The tablet needs to dissolve in the GI tract to be able to permeate the intestinal wall. One of the main properties restricting solubility is hydrophobicity, which is a driving force for transcellular permeability. b) The following general properties can be extracted for permeability (from the left-hand side): the transcellular route is used by nonpolar, medium-sized ($M_w < 500$), and uncharged compounds; the paracellular route is used by compounds that are polar, small ($M_w < 180$), and charged; energy-dependent active-transport processes (transport efflux and influx proteins) are used by compounds in the medium to large size range, both polar and nonpolar. Furthermore, the compounds may be charged or uncharged.

1.1. Drug solubility

The aqueous solubility of the compound is dependent on both the intramolecular forces in the solid state and the intermolecular forces between the drug molecule and the surrounding intestinal fluid. The solubility will be poor if it is more energetically favorable for the molecules to bind to each other than to the water molecules; as a result, the drug molecules favor their solid form over dissolution into the aqueous fluid. However, poor solubility might also be a result of unfavorable bonding between water and drug molecules. Depending on which of these underlying properties is dominant, different

Ulf Norinder received his MS in chemical engineering in 1981 and PhD in organic chemistry in 1984 from Chalmers University of Technology in Gothenburg, Sweden. He is currently a principal scientist at AstraZeneca R&D Södertälje and is also an associate professor in the department of organic chemistry, Chalmers University of Technology, as well as an adjunct professor in computational pharmaceutics in the department of pharmacy, Uppsala University. His research interests include computer-assisted drug design and pattern recognition with special emphasis on multivariate data analysis.



Christel Bergström earned her PhD from Uppsala University, Sweden, in 2003 under the direction of Professor Per Artursson. Since 2003 she has been a researcher at the department of pharmacy and head of the program dedicated to pharmaceutical informatics at the centre of pharmaceutical informatics at Uppsala University. Her research is focused on in silico and in vitro models predicting absorption, distribution, metabolism, and excretion/elimination processes for drugs. She is currently involved in projects aimed at predicting aqueous drug solubility, active and passive transport processes of drugs, and blood–brain barrier permeation.

physicochemical properties will be important for the behavior of the molecule in water. Multivariate data analysis of melting point, a property that reflects the stability of the solid state, has shown that molecules proven to form stable crystals in general are small, rigid, and polar.[10] On the other hand, compounds that are hydrophobic, flexible, and large demand a larger cavity to be formed in the aqueous fluid in order to go into solution, and may have restricted solubility as a result of these properties. Models for prediction of solubility are further discussed in section 4.2, but the above-mentioned contrasts indicate that solubility is not a straightforward property to predict.

1.2. Intestinal permeability

A compound can permeate the intestinal wall by using the paracellular route (between the cells) or the transcellular route (through the cells) by passive diffusion. In general, small, hydrophilic and/or charged compounds, which cannot permeate the lipophilic cell membrane, diffuse through the aqueous pores. However, the pores cover less than 1‰ of the intestinal surface,[11] and this, in concert with the solute restriction by the tight junctions of the pores, largely limits the contribution of the paracellular pathway. Compounds that show a reasonable hydrophobicity ($\log D_{pH7.4} = 0$–2) and intermediate size (up to $M_w = 500$) are assumed to permeate the intestinal wall by passive transcellular diffusion. Even though the transport by the transcellular route seems to be a rather complex process that demands partitioning between lipophilic and hydrophilic milieus several times, the vast majority of druglike compounds use this pathway. Larger molecules with a large number of hydrogen bond donors and acceptors, sometimes in combination with a high lipophilicity value, may use active processes and transport proteins to get through the cells. However, the latter properties also increase the risk that the compound might be transported by efflux proteins, resulting in a secretion of the compound back to the intestinal lumen. Such efflux results in a lower drug concentration reaching the blood circulatory system and the site of action.

To conclude, two of the main factors that influence intestinal drug absorption are aqueous solubility and intestinal permeability. These characteristics are dependent on opposed physicochemical properties, resulting in difficulties in finding easily interpretable models for prediction of the drug-absorption process. Many computational solubility and permeability models have been developed so far, and a majority of these are either dataset restricted (for example, only a small volume of the druglike space has been included in the training of the model), or mechanism based (for example, valid for a specific transport route or transport protein). This indicates that first, the datasets used in the development of absorption models applicable in the drug-discovery process need to cover a large volume of the druglike space. Second, the development of pharmaceutical informatics tools is crucial for the extraction of correct information from combinations of all mechanism-based models that are available.

1.3. Toxicity

Structure–activity relationships in toxicology are based on the assumption that an adequate representation, that is, geometric and electronic, of the investigated structures will permit the derivation of a quantitative statistical model. This assumption is not unique for toxicological modeling but is true for all other areas of ADME modeling as well. However, in toxicology the situation is somewhat further complicated by the fact that toxicological effects may result from many different mechanisms. This, in turn, means that it is possible to establish good in silico models for congeneric series of molecules, but that more general models may be difficult to derive. Already in 1969, Corwin Hansch, the founder of modern QSAR, proposed that, in general, a biological and toxicological action for a congeneric series of structures could be described by the model:

$$\log(\text{activity}) = a(\pi) + b(\varepsilon) + c(S) + d \quad (1)$$

for which π, ε, and S are related to hydrophobic, electronic, and steric descriptions of the studied compounds, respectively.

Toxicological structure–activity investigations have been conducted over the years in areas such as nonspecific toxicity, aquatic toxicology, mutagenicity, and carcinogenicity as well as developmental toxicity and skin sensitization. For a recent article on the subject, see reference [12].

2. History and Development

Traditionally, the discovery setting has worked in serial with the primary focus set on the identification of new structures that show good pharmacological effects. After the screen for pharmacological effects, other important properties such as solubility, permeability, stability, metabolism, distribution, elimination, and toxicity are investigated one after each other. This is an ineffective, time-consuming drug-discovery process that does not necessarily result in the identification of the optimal drug molecule, due to the investigation of one property at a time. The pharmaceutical industry is currently working with experimental screens in a parallel setting, in which the above-mentioned properties are experimentally examined at the same time and thereafter evaluated. Hence, all properties affect the final decision on which compounds to pursue, leading to improved selection of the CDs. Further, the discovery setting is now moving into the virtual era, applying several virtual tools to further cut time and costs during the discovery process. By designing virtual compound libraries and testing these by virtual docking to targets and in silico models for ADMET properties, a prioritized library predicted to have a favorable pharmaceutical profile and acceptable pharmacological potency is computationally selected and thereafter synthesized. This scenario results in knowledge-based synthesis of fewer compounds with better properties than can be generated by both the serial and parallel setting. After the synthesis of the prioritized library, the potency and the developability of the compounds must be experimentally confirmed (Figure 4). Thus, methods for rapid and reliable experimental screening of

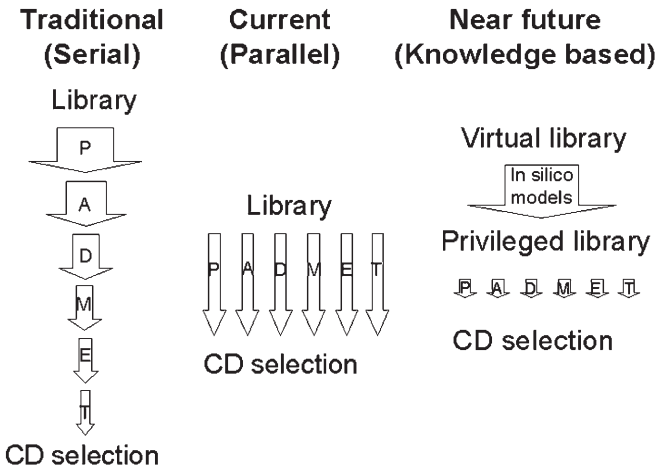


Figure 4. The traditional setting (left) applied in the candidate drug (CD) selection was a serial experimental testing of pharmacology (P) followed by the different ADMET properties, resulting in extended development times and difficulties in finding the optimal compound. Today the pharmaceutical industry applies a parallel setting (center) and moves toward the knowledge-based setting. In the parallel setting, both pharmacology and ADMET properties are experimentally evaluated simultaneously, and the complete profile can be used when selecting the CD. In the knowledge-based setting (right), a virtual library designed in the computer is primary evaluated through different in silico models for pharmacology and ADMET properties. A privileged library is synthesized based on the results from the virtual screening, and the compounds are thereafter experimentally tested.

these properties are warranted. Today, rapid methods have been devised for the screening of several of the ADMET properties at the expense of reliability,[7,13] resulting in a large number of false-positive results in the screens. By incorporating reliable computational and experimental screens, better leads will be produced, saving time and money during the discovery process. However, if the virtual-based discovery setting is to be successful, new computational tools need to be developed. The development of informatics tools applicable for pharmaceutical profiling and with the capacity to handle large databases with such diverse information as in silico, in vitro, and in vivo data as well as qualitative and quantitative information will be of utmost importance.

3. General Considerations

3.1. General terms

When trying to develop in silico models for the prediction of ADMET properties there is, in most cases, a trade-off between accuracy, speed, and, many times, transparency of the derived models. This is not always a significant problem, as the various models may be intended for different use, such as high-throughput in silico screening or for guidance and focusing. In reality this often means that rapidly computed descriptors, often of one- and two-dimensional nature, are used in the former type of model, whereas more computationally intensive, three-dimensional-based variables are employed (sometimes in conjunction with one- and two-dimensional representations) for the latter model type.

Schultz and Cronin[14] have put forward some rather basic requirements to derive statistically sound models:

1) A well-defined and measurable target;
2) A chemically and biologically diverse data set;
3) Physicochemical descriptors that are consistent with the modeled target;
4) Use of an appropriate statistical technique;
5) Where possible, a strong mechanistic basis.

3.2. Datasets and models

One consideration to take into account in ADMET modeling is the availability of relevant and accurate datasets. In general there exists a relatively small number of them, especially public datasets, with a desirable quality of data, diversity of the investigated structures, and which are large enough to permit sufficient validation of the derived model. In the ADMET literature, especially within the areas of solubility, absorption, and permeability, it is common that models are derived on rather few compounds (fewer than 50). These models are usually rather local, and have a limited scope with respect to their predictive ability. Local models, however, are quite useful in many cases for advancing a particular project or set of compounds, but in one particular respect, a vast majority of the published models lack information with respect to the applicability domain in which they operate. Very few publications of ADMET models explicitly point out or discusses how the applicability domain of the derived model in question is established. Statistical models in general, including in silico ADMET models, should always have some protocol (measure) to determine if the prediction of a property for a particular compound is within, on the border of, or outside (perhaps also how far outside) the applicability domain of the model based on the chemical description employed. This aspect is further discussed in section 3.3.6 together with an approach on how to proactively use the information of outliers to further advance the model. Absorption and permeability models and the datasets on which they were based also have a particular problem with respect to active transport. In the past, datasets were modeled under the assumption that the absorption or permeation process was devoid of active transport, although later analysis showed that this was not entirely true. Compounds in datasets investigated today will probably be found later to be involved in active transport by transporters not yet identified. An extenuating circumstance is the fact that if a model with good statistics and good predictive ability is derived despite the fact that some compounds of the training set (i.e., the compounds used to derive the model) are involved in active transport, then two alternative explanations may emerge: 1) the amount of active transport of a particular compound is rather small (negligible), or 2) the derived model somehow encompasses the information also related to active transport, although this was not the intent from the start in most cases.

3.3. Statistical tools

3.3.1. Linear multivariate methods

The statistical methods most often employed for developing ADMET in silico structure–property relationships are linear multivariate methods such as multiple linear regression (MLR) or partial least squares (PLS). Although aimed at the same endpoint, namely to derive a statistically sound and predictive structure–property relationship, the underlying assumptions regarding the information contained in the independent variables (i.e., the chemical description of the investigated structures) are different for the two methods. With respect to MLR, the following should be considered:

1) MLR assumes each variable to be exact and relevant; that is, the information content in each variable is to be used 100% for developing the statistical model.
2) Strong colinear variables must be eliminated by removing all but one of the strongly correlated variables. Otherwise spurious chance correlation may result.
3) The number of variables cannot exceed the number of observations (for example, the number of measured ADMET property points) to be studied. A rule-of-thumb is that the number of variables used should not exceed a fourth of the number of observations.

Regarding PLS, the following applies:

1) The descriptors (variables) are not treated as exact and relevant but as consisting of two parts: one related to the dependent variable and the other not related (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 PLS uses internally only a few (usually fewer than 5–10) latent variables for the actual statistical analysis.
4) Again, a rule-of-thumb is that the number of latent variables used should not exceed a fourth of the number of observations.

The PLS model becomes identical to the MLR when the number of latent variables of a PLS-derived model becomes equal to the number of actual independent variables, something that rarely happens as a consequence of model validation. The regression coefficients of the MLR model are straightforward to interpret, whereas the PLS latent variables need to be retransformed back into original variable space in order to be interpreted in a similar manner. This also means that the PLS "regression" coefficients are dimensional dependent, that is, they depend on how many latent variables (PLS components) are used. However, as each PLS component explains a decreasing amount of variance, it is usually not that important if a PLS model is based on three or four components which also means that the PLS "regression" coefficients will not differ very much between the three- and four-component model.

3.3.2. Nonlinear multivariate methods

Although a majority of the published ADMET models are based on linear multivariate methods as discussed in section 3.3.1, other nonlinear methods have also been employed. The most commonly used nonlinear method in ADMET modeling is neural networks (NN). Back-propagation NNs have been used to model absorption, permeation, solubility, and toxicological effects. A particular problem for many NNs is the tendency for these networks to overtrain (see further discussions on model validation in section 3.3.4). This needs to be closely monitored to avoid the situation in which the derived model becomes an "encyclopedia", that is, the model can perfectly explain the variance of the investigated property of the compounds used to derive the model, yet can have poor predictive ability with respect to new compounds.

3.3.3. Dataset pretreatment

It is very important to give the variables used in the model development equal chance, regardless of their respective numerical scales, 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 auto-scaling, whereby the variance of each variable is adjusted to 1. Sometimes it is also desirable to center each of the variables with respect to their mean values.

3.3.4. Model validation

Stringent model validation is a cornerstone for the successful development of any statistical model. Without proper validation, the predictive ability of the derived model cannot be estimated. Likewise, the derived model may equally well be nothing more than a random model. There are a few standard techniques that should be employed to ensure proper validation:

1) Cross-validation is one technique for the internal validation of a proposed model. When using cross-validation, the training set is divided into groups, usually four to seven in number, and one group is removed from the set. The model is then derived using the rest of the training set. The dependent property of the compounds of the group left out is then predicted by the developed model. Each group is successively left out and predicted in the same manner as just described. The predicted residual error sum of squares (PRESS) is computed from all the predictions. The PRESS value is compared with the sums of squares for the dependent variable y (SSY):

$$\Sigma(y_{i,\,\mathrm{measured}} - y_{\mathrm{mean}})^2 \qquad (2)$$

A squared correlation coefficient (Q^2) is then defined as:

$$Q^2 = 1 - \mathrm{PRESS}/\mathrm{SSY} \quad (3)$$

A significant difference between Q^2 and the normal squared correlation coefficient (R^2) is that the former may also assume negative values, indicating that the model has worse predictive ability than using the mean value as predicted value for each compound. Q^2 should be >0.5 for the model to be considered to have reasonable practical predictive performance.

2) An external validation set should be used as an independent test of the predictive ability of a derived model.
3) The value of the dependent variable is randomly redistributed among the compounds (randomization of the dependent variable). A model is then derived based on the redistributed values and checked for its predictive performance using the methods outlined under points 1) and 2) above. This procedure is repeated typically between 50–100 times. There should be a clear separation in predictive ability between the model based on the "true" dependent values versus the models based on redistributed values.

3.3.5. Training and test set selection

It is certainly possible to chose a training set at random and also to derive a statistically sound and predictive model. Chances are, however, that the choice of training set compounds is somewhat skewed. This, in turn, probably means that many of the remaining compounds, the external test set, will fall outside of the applicability domain of the derived model and constitute outliers to the present model. For a model to show good predictive capability and to cover the investigated descriptor space in a good manner, the training set must be chosen with some care. There are several methods available for the selection of well-distributed training sets. Two such methods are exemplified here:

1) Experimental design methods of some appropriate complexity are one such choice. The number of compounds to be used for the training set depends on the chosen design scheme and the number of investigated independent variables (descriptors), but may typically range between 8 and 64.
2) Maximin methods, in which the aim is to maximize the closest (minimum) distance between two potential training set compounds in the investigated descriptor space is another. By maximizing the closest distance, all other distances between training set compounds are greater thus ensuring a rather uniform distribution of compounds comprising the final training set.

3.3.6. Applicability domain estimation

It is essential that the applicability domain of a derived model can be evaluated so that outliers to the model may be indicated. If an established statistical model is to be regarded as poor from a predictive point of view, this should be done based on correct reasons, namely, that the model has truly poor predictive ability and not from the fact that the model cannot estimate outliers to the model with acceptable accuracy. The latter case is probably the most common cause for statistical (ADMET) models to "fall from fame", especially those that can be accessed through internal or external web services. In many cases it is difficult if not impossible to find out about the compounds used as training set and/or the chemical description used in the model. Thus, many compounds outside the applicability domain of the model will be submitted. It is therefore of great importance to have an indication together with the prediction whether or not the compound is considered to fall inside or outside of this domain, that is, if the compound is an outlier or not. The outlier information, and possibly how far from the model the compound in question is, may in many cases be used in a more proactive way than just realizing that a number of compounds submitted to the model for prediction are, in fact, outliers to the *present* model. Thus, by analyzing the outliers, perhaps virtual compounds, from various points of view, for example, structural or synthetic, some of these compounds may later be synthesized and tested experimentally. The same compounds may then be incorporated into a revised model that will have a broader applicability domain. There are different methods to determine whether a particular compound is to be labeled as an outlier or not. In this section we describe two of these methods:

1) The first of these methods is the Mahalanobis distance. This distance in descriptor space measures how similar the investigated compound is to the training set compounds. The Mahalanobis distance is superior to the corresponding and more familiar Euclidian distance, as the former takes correlations between the variables into account; the Mahalanobis distance does not assume orthogonal descriptors as does the Euclidian distance.
2) The second method is related to the remaining amount of information present in the variables used to describe the compound that has not been used by the model. This method is more closely related to the PLS method (see section 3.3.1). Thus, if a particular compound contains much more unexplained variance (information) in the chemical descriptor variables than is the case for the training set compounds, it is likely that the compound in question has other properties not taken into account by the present model that will impact the true value for the investigated ADMET property. Therefore, the predicted value will most likely deviate substantially from the corresponding experimental value.

3.3.7. Calculation of descriptors

A large number of different descriptors have been used to model ADMET properties. 1D-, 2D-, and 3D-based computed chemical properties have been found useful for deriving statistically sound and predictive ADMET models. The choice of which type of descriptors, or combinations thereof, to use very much depends on the aim for the derived model. Is the model to be used for screening large sets of (virtual) compounds or for smaller sets of structures? How important is interpretability versus predictive accuracy and robustness of the prediction? How much computational time is allowed to be spent on each individual prediction? In fortunate cases, many of these considerations coincide; the model is robust, shows good predictive capability, and is based on rapidly computed descriptors that are easy to interpret from a mechanistic of physicochemical point of view. However, in most cases, there is a tradeoff between objectives. Depending on the priorities for the development of the particular model at hand, different sets of descriptors have to be employed. With these aspects in mind, it is usually quite useful to develop more than one model for the same ADMET property based on different sets of descriptors. This way, interpretability (incorporated into a model with acceptable although perhaps not the best predictive ability), robustness, speed, and accuracy can be achieved. For instance, the first kind of model can be used for understanding the important physicochemical properties that influence the particular ADMET property in question: how these physicochemical properties should be modified to achieve a suitable level for the investigated ADMET property. This, in turn, gives an indication of how new and improved compounds could be designed, and enables one to focus on promising regions of the chemical space of the model. Thus, instead of simulating a very large number of virtual compounds for prediction by the model, a much smaller number can be submitted. Subsequently, this smaller number of structures can then be submitted to a more robust and accurate yet more complex model for the final estimation of the ADMET property. In many cases, consensus or ensemble models, although more complex in nature, are useful for deriving in silico ADMET models with high levels of predictive accuracy and high degrees of robustness. These models may often be looked upon as "gray" and not "black" boxes, as each model can be interpreted, but the multitude of them makes the overall picture difficult to comprehend.

1D and 2D descriptors are generally much faster to compute than the corresponding 3D-based descriptors. Moreover, the possible problems associated with generating a reasonable 3D conformation for the investigated structure are eliminated.

1) 1D descriptors such as molecular weight, molar refractivity, and the number of atoms and bonds have been used to model permeability, absorption, solubility, and toxicological effects. These kinds of descriptors are usually rather easy to interpret.
2) There is a large number of 2D descriptors. Many of them are topological in nature, that is, they are computed from the connectivity of the investigated compound or, more specifically, from the mathematical graph that the structure represents. They often contain important information with respect to ADMET modeling. Some of the more well-known and frequently used topological variables are the Kier and Hall descriptors. However, these topological descriptors are often somewhat difficult to interpret with respect to the question: "How should the present structure be modified to improve the ADMET property presently investigated?" A particular subset of topological descriptors, the so-called electrotopological descriptors, is an exception with respect to interpretability. These kinds of descriptors are relatively easy to interpret in terms of hydrogen bonding, and quite a few published investigations report the electrotopological (or e-state) descriptors as useful for deriving good ADMET models.
3) In many cases descriptors derived based on 3D descriptors are superior to lower-dimensional descriptors because they capture important information such as internal hydrogen bonds and other potentially important, yet buried, functional groups, which are revealed only by using the actual 3D representation of the investigated compound. The 3D descriptors may also be easier to interpret than some of the previously mentioned variables. However, choosing the correct 3D conformation may, in some cases, cause problems depending on how rapidly the descriptors must be generated. Software is available for converting 2D structures into 3D structures (for example, Corina and Concord). Although successful in a vast majority of cases, both these programs sometimes fail during the conversion process, or the 3D conformation given is not reasonable for the particular modeling exercise. Certainly, some sort of conformational analysis would be desirable in many cases. For the 3D descriptors there is a large difference in complexity and computational speed, ranging from rapid calculations of various surfaces and volumes of a structure to high-level (e.g. ab initio) quantum-mechanics-based descriptors such as orbital energies, charges, polarizabilities, and multi-pole moments (Table).

In some cases, it is possible to go from more computationally demanding descriptors to more rapidly computed descriptors while preserving the information content from one descriptor matrix to the other.

4. Applications and Practical Examples

4.1. Physiological factors and experimental parameters influencing the accuracy of predictions of intestinal drug absorption

4.1.1. Solubility

The intestinal solubility of a compound is dependent on physicochemical properties of the molecule (discussed in sections 1.1 and 4.2), the location in the gastrointestinal (GI) tract, the general physiology, and the dosage form. By analyzing the descriptors in the Noyes–Whitney equation[15] the physiological

Table 1. Examples of commercial software available for prediction of ADMET-related properties.[a]

Software	Company	Dis[b]	Sol[c]	Per[d]	Tps[e]	OB[f]	HIA[g]	BBB[h]	Mtb[i]	Other PK[j]	Tox[k]
AbSolv			×	×							
ACD Solubility DB	ACD Labs		×								
ADME batches	PharmaAlgortihms		×				×				×
ADME boxes	PharmaAlgorithms		×	×	×	×					
Cerius^2	AccelRys		×				×	×	×	×	×
Cloe PK	Cyprotex			×			×		×		
GastroPlus	Simulations Plus	×	×	×		×	×		×	×	
iDEA PKexpress	Lion Biosciences			×			×		×	×	
KnowItAll ADME/Tox	Bio-Rad Laboratories		×			×	×	×	×		×
OraSpotter	ZyxBio		×		×		×				
PK-sim	Bayer Technology Services			×	×	×	×		×	×	
QikProp	Schrödinger		×	×				×			
QMPRPlus	Simulations Plus		×	×			×	×		×	×
SLIPPER	TimeTec		×				×				

[a] ×: Indicates property predicted in the reported software. [b] Dissolution. [c] Solubility. [d] Intestinal permeability. [e] Transporters. [f] Oral bioavailability. [g] Human intestinal absorption. [h] Blood–brain barrier permeability. [i] Metabolism. [j] Other pharmacokinetic properties. [k] Toxicity.

and pharmaceutical influence on dissolution becomes apparent:

$$dm/dt = D A(C_s)/h \quad (4)$$

in which C_s is the maximum amount of drug that can be dissolved in the fluid (the solubility value), A is the surface area of the undissolved compact, D is the diffusion coefficient in the intestinal fluid, and h is the height of the diffusion layer adjacent to the undissolved tablet. The diffusion coefficient of a molecule is dependent on the viscosity of the fluid; the higher the viscosity, the lower the diffusion coefficient, and thereby a smaller amount of compound will be dissolved per unit time. Furthermore, the larger the surface area of the undissolved compact and the higher the solubility of the compound, the more of the compound will be dissolved per unit time.

The pH conditions the GI tract vary from pH 1 in the stomach up to pH 8 in the colon. Thus, the solubility of protolytes, compounds with one or several ionizable groups, will be dependent on the location in the GI tract.[16] Compounds with an acidic functional group will show increased solubility at pH values above the pK_a, whereas the solubility of bases will improve at pH values below the pK_a value. For ampholytes the lowest solubility will be found at the isoelectric point, which is obtained at a pH value between the acidic and basic pK_a values. Another physiological factor that will influence the solubility is the ionic strength of the intestinal fluid. This will be dependent on food and fluid intake, and on absorption and secretion of fluid within the intestine.[17] In general, the solubility decreases with increased ionic strength as a result of the salting-out effect and/or the common-ion effect displayed by the counterions in the solution.[18,19] However, the presence of electrolytes can, in specific cases, improve the solubility.[20] This phenomenon is known as the salting-in effect and occurs when additives such as electrolytes loosen up the tight water structure, thereby driving the formation of solvent cavities for the drug molecule. Further, food induces the secretion of bile salts (surfactants secreted by the bile bladder), which may improve the solubility of poorly soluble compounds by acting as a wetting agent or by solubilization within the lipophilic core of bile salt micelles formed at higher bile salt concentrations.[21]

The in silico models derived for solubility are based on intrinsic solubility as their experimental input data. The intrinsic solubility is the solubility value determined for the neutral (i.e. uncharged) species of the compound and is generally determined at two pH units above the pK_a value of bases, and two pH units below the pK_a value for acids. Ampholytes are determined at their isoelectric point. Therefore, solubility values used for model development seldom reflect the apparent solubility observed in the intestinal fluids. Hence, the predicted values obtained from the models need to be transferred to an in vivo situation, for instance by use of the Henderson–Hasselbalch equation, which takes into account the pH dependence of solubility.[16]

4.1.2. Permeability

The rate and extent of intestinal permeation is dependent on physicochemical properties of the compound (see sections 1.2 and 4.3) and physiological factors. Drugs are mainly absorbed in the small intestine because it has a larger surface area and looser epithelium than the colon.[17] The permeation of the intestine may be affected by the presence of an aqueous boundary layer and mucus adjacent to the cells, but for a majority of substances the epithelial barrier is the most important barrier to drug absorption. The lipoidal cell membrane restricts the permeability of hydrophilic and charged compounds, whereas large molecules are restricted by the ordered structure of the lipid bilayer.

In the gastrointestinal tract, pH-dependent permeability is observed (see also section 4.1.1); the higher degree of ionization of the compound, the poorer the permeability. Other physiological factors that influence the permeability of compounds include the motility of the gastrointestinal tract, the expression of transport proteins, and the thickness of the

mucus layer adherent to the enterocytes. These factors influence the permeability as follows: the better the motility of the intestine, the smaller the unstirred water layer (UWL) adjacent to the cells. In general, peristalsis is so good in vivo that the UWL does not become the rate-limiting step in the absorption process. Furthermore, the extent to which the transport proteins is expressed will largely influence absorption. The fraction absorbed (FA) will either increase or decrease with high expression levels of the transporter depending on whether the transport protein is an influx protein, which transports compounds through the enterocytes into the bloodstream, or if it is an efflux protein, which transports compounds out from the cell back to the intestinal lumen. Finally, a thick mucus layer adjacent to the cells may slow the diffusion of the compound and become the rate-limiting step of the absorption process. Taken together, these physiological factors may result in large inter-individual variability in the permeability value, giving large standard deviations in the FA in vivo.

The in silico models derived for permeability are based on experimentally determined permeability values using different cell culture models. The most commonly used is the Caco-2 cell line, which is a human colon carcinoma cell line.[22,23] This cell line is inexpensive and easy to culture; these factors, in concert with its human origin, make it a popular cell model. However, the colonic epithelium is somewhat tighter than the small-intestinal epithelium, resulting in permeability values 1–2 orders of magnitude less than that observed in small-intestinal tissue. Despite this fact, the permeability ranking of the compounds is in good agreement with that obtained in the small intestine, and therefore the model is a valid tool for estimates of FA over the small-intestinal wall. Other cell lines used for determinations of permeability values are, for example, MDCK cells, which originate from canine kidney tissue[24] and 2/4/A1 cells, which originate from the rat small intestine.[25] The drawback with these cell lines is that they are not obtained from human tissues, and the MDCK cell line is further restricted by its kidney origin, resulting, for example, in an expression pattern of transporters that is different from that of the human small intestine. In vivo perfusion studies in humans can be used to determine intestinal drug permeability.[26] All the different experimental settings and protocols applied for permeability measurements will largely influence the permeability data obtained. It is therefore important that the experimental values used in the development of computational models are determined in a consistent manner, within the same laboratory using one experimental setting and one experimental protocol. Only then is the in silico model based on high-quality data, and the noise level minimized.

4.1.3. Fraction absorbed

Several computational absorption models based on human fraction absorbed (FA) data have been published.[27–30] These models should be interpreted with caution, owing to the fact that the datasets are compiled from a large number of literature sources of varying quality. The following facts must be taken into consideration:

1) Different experimental methods are used to determine the FA, resulting in a large variability in the numbers reported.
2) The influence of active transporters and the concentration dependence in vivo are not always clear.
3) It is not clear whether the FA is solubility-limited and/or permeability-limited, resulting in difficulties in obtaining a mechanistically transparent model.
4) The datasets obtained are often heavily biased toward compounds with high FA due to the fact that a majority of the compounds for which FA is known are commercially available compounds. Hence, these compounds are the results of years of discovery and development and they are expected to show a good absorption profile. However, this fact will influence the in silico models obtained. These will be rather good at sorting compounds as high FA, but poor in determining other classes such as intermediate or poor FA due to the lack of such compounds in the training sets.

It is not unusual that published FA data for the same compound vary largely. For example, FA can be reported as either 10% or 60%, generally classified as poor and intermediate FA, respectively. If such data are used for training the in silico model, the model will, to a large extent, be based on noise. This leads to poor external predictions and non-interpretable results. In our view, it is more relevant to estimate the FA based on in silico solubility and permeability screens.

4.2. In silico solubility models

Modeling solubility represents perhaps a bigger challenge than modeling absorption and permeability. Why is this so? Some of the particular issues involved in trying to derive good statistical models for solubility are related to the quality and precision of the dependent variable, that is, the solubility values, the complexity (or lack thereof) and diversity of the compounds of the investigated datasets, the possible influence of the solid state for each of the studied compounds, and whether or not modeling solubility is fundamentally a linear or nonlinear problem. With respect to the first issue, namely the quality (precision) of published solubility values, it must be recognized that these values stem from a variety of experimental procedures that make comparisons between sets of measurements rather difficult. It is not uncommon to find published values for a particular compound that differ by as much as a factor of 10! This, in turn, certainly makes modeling solubility a difficult problem. Many of the publications on modeling solubility contain a large number of compounds, but in many cases a majority of these structures are rather simple, non-druglike molecules in which the structural complexity with respect to functional groups and ring systems is somewhat limited. Such datasets are easier to model and for which to derive good quantitative structure–solubility relationships. Also, it has been recognized for many years that the solid state of each of the investigated compounds may very well play an important role in the success or failure of the modeling attempt. The difficulty here lies in the fact that it is rather difficult to obtain a theoretical estimate of the solid phase within reasonable com-

putation time and with satisfactory precision. Nevertheless, many attempts have been made, and many articles have been published over the years on how to model solubility. In this section, some of these recently published works will be described and commented upon to illustrate the present status of field:

1) A well-known paper is that by Huuskonen.[31] In this investigation, a back-propagation artificial neural network (ANN) was used as statistical engine and e-state descriptors to parameterize the chemical structures. The investigation was based on 1297 compounds, also known as the "Huuskonen dataset", and used a large training set, a randomly chosen test set, and a second (external) test set consisting of 21 compounds. A model with good statistical quality was developed (see Table 2). Notable in this investigation is the use of the dataset-specific "test" set where, in this case, according to the author: "The network architecture and the training end point giving the highest coefficient of determination, r^2_{pred}, and the lowest standard error *s* for the predictions of the test set were then used". This means that the randomly chosen test set is, in fact, a validation set for the training of the neural network and the only "true" external test set is the 21-compound set. A somewhat larger external test set is desirable to evaluate the predictive ability of the derived model in question more extensively. The statistical results are presented in Table 2.
2) Several other investigations of solubility using the Huuskonen dataset and other datasets using ANNs and various other neural network methods such as Bayesian NNs and Kohonen's self-organizing NNs have been published the last few years (see Table 2 for results and references).
3) Jorgensen and Duffy published a recent review of predictions of solubility focused on drugs.[32]
4) Consensus modeling using ANNs have been published by Manallack and co-workers.[33] They used BCUT variables with diagonal elements consisting of charges, hydrogen bonding acceptor and donor ability, and polarizability.

Many, not to say an overwhelming majority, of the investigations published on the prediction of aqueous solubility of drugs (and other compounds) have identified the most important (influential) factors to be related to hydrogen bonding, polarizability or polarity as well as hydrophobicity expressed through terms such as e-state indices, hydrogen bonding terms, and the log *P* variable.

Lately, consensus modeling has come into play as a useful tool to obtain robust models with good predictive ability. By using this approach the weakness of one particular model is compensated by the other models, thus obtaining a much more robust behavior for the ensemble of models.

However, there is a problem with the presently derived models apart from the accuracy of experimental data as discussed earlier. Although at first sight these models appear to be quite respectable statistical models with rather good predictive ability, they are not optimal for predicting the solubility of drugs. Why is this?

An investigation by Norinder, Lidén and Boström[34] will be used to illustrate the situation, but again, this is a general deficiency among the published models for predicting aqueous solubility. The statistics for the model is appreciable (see Table 2, Norinder; PLS, and also a plot of experimental versus calculated solubility in Figure 5).

However, closer inspection of the solubility range relevant for most drugs, −6 to −3, reveals a rather different picture (Figure 6). For the accurate prediction of such entities the derived model is not very useful. This is, however, the situation that investigators are faced with when trying to derive models for accurately predicting drug solubility that can be of valuable practical use for medicinal chemists, biologists, pharmacol-

Table 2. Summary of different methods and models for the Huuskonen aqueous solubility dataset.

		Training Set			Test Set			Test Set 2			
Model	Type	n[a]	R^2[b]	s[c]	n	R^2	s	n	R^2	s	Ref.
Gasteiger	MLR	797	0.79	0.93	496	0.82	0.79	21	0.56	1.20	[64]
	ANN40-8-1	797	0.93	0.50	496	0.92	0.59	21	0.85	0.77	
Liu	ANN7-2-1	1033	0.86	0.70	258	0.86	0.71	21	0.79	0.93	[65]
Tetko	MLR	879	0.86	0.75	412	0.85	0.81	21	0.77	0.99	[66]
	ANN33-4-1	879	0.94	0.47	412	0.91	0.60	21	0.90	0.64	
Huuskonen	MLR	884	0.89	0.67	413	0.88	0.71	21	0.83	0.88	[31]
	ANN30-12-1	884	0.94	0.47	413	0.88	0.60	21	0.91	0.63	
Wegner	ANN9-15-1	1016	0.94	0.52	253	0.93	0.54	21	0.82	0.79	[67]
Norinder	PLS	800	0.87	0.69	497	0.93	0.58	21	0.80	0.82	unpublished work
Norinder	RDS/ensemble	800	0.97	0.35	497	0.95	0.51	21	0.87	0.67	unpublished work

Model	Type	n	Accuracy [%][d]	n	Accuracy [%]	n	Accuracy [%]	
Norinder	RDS/ classification	800	82.10	497	80.30	21	0.83	unpublished work
Norinder	RDS/ classification/ ensemble	800	98.00	497	86.90	21	0.91	unpublished work

[a] Number of compounds. [b] Squared correlation coefficient. [c] Standard error. [d] Percentage of compounds correctly classified into the three classes: good, medium, and poor.

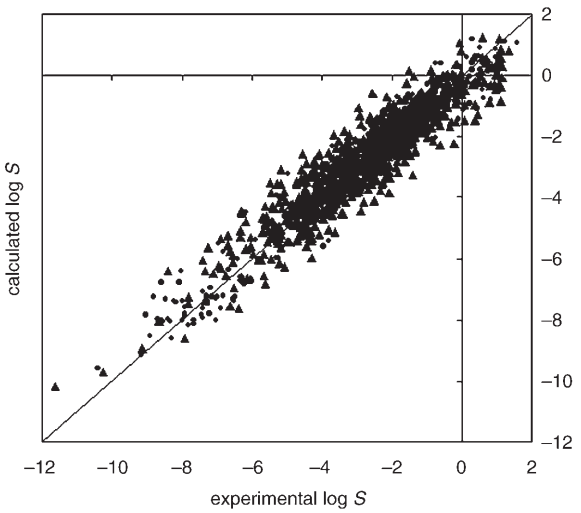


Figure 5. Model of the Huuskonen aqueous solubility dataset using PLS (reference [34]). Training set: ▲, test set: ●. The plot shows the "deceptively" good performance of the developed model with respect to use for predicting aqueous solubility for new potential drug compounds (see also Figure 6).

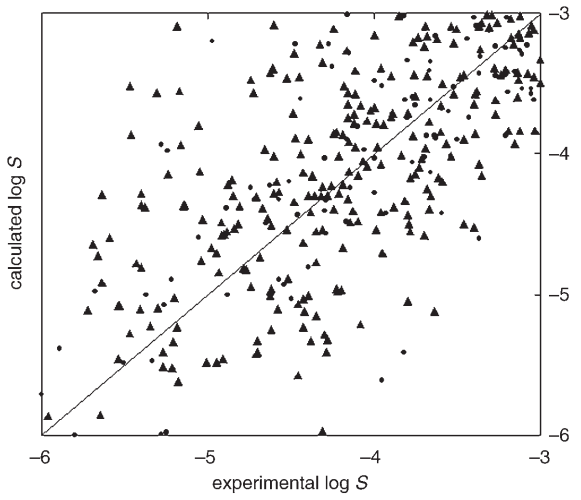


Figure 6. Close-up of the area of aqueous solubility that is of interest from a drug-development perspective (reference [34]). Training set: ▲, test set: ●. The graph shows the "true" or limited performance of the developed solubility model with respect to predictive capability for new compounds.

ogists, and others in trying to advance research projects to arrive at compounds with reasonable solubility. Using consensus or ensemble modeling instead of a single model usually improves the situation somewhat, as exemplified by a rule-based ensemble model using 2D parameters on the Huuskonen dataset (Table 2, Norinder; RDS/classification/ensemble).[34] Sometimes, depending on the targeted use of the model as well as the precision of the experimental data, it is more useful to bin the range of solubility into two or three bins (categories). This approach is exemplified on the same dataset in which three categories ($\log S$: good, > -2; medium, -2 to -4; poor, < -4) were used. The results of a single-model approach as well as an ensemble modeling (50 models) are reported in Table 2.

4.3. In silico models of permeability and FA

4.3.1. Descriptors used for permeability predictions

Response parameters in the study of permeability-related absorption can include permeability through a cell monolayer such as that of Caco-2, MDCK, and 2/4/A1 cells, the permeability coefficients obtained from Ussing experiments, the effective permeability in the intestine, and the FA of the dose. Permeability models predicting intestinal absorption are generally models of transcellular passive diffusion, and descriptors of hydrophobicity, hydrophilicity, and size have proven important (see Table 3). Hydrophobic descriptors can be regarded as measures of distribution capacity into the membrane, hydrophilic descriptors as desolvation restriction when the compound partitions from the intestinal aqueous fluid into the hydrophobic membrane, and size reflects the steric hindrance to diffusion through the membrane.[35] The $\log P_{oct}$ descriptor has been used historically to predict membrane permeability, and hence it is incorporated into a large number of the models developed. For noncomplex datasets, properties such as $\log P_{oct}$, polar surface area (PSA), and hydrogen bond counts have each been used as a single predictor of permeability.[36–39] However, lipophilicity can be regarded as a composed property that is largely dependent on the size and hydrophilicity of the compound. Thus, the use of these two components might be regarded sounder than $\log P_{oct}$. Indeed, the use of molecular weight and the number of hydrogen bonds have been shown to predict permeability of a smaller dataset better than the use of $\log P_{oct}$.[40]

The introduction of more complex datasets used for model development has pointed at the need for several descriptors and multivariate data analysis (Table 3). For instance, combinations of PSA and NPSA (nonpolar surface area) were able to predict the permeability of a series of peptides when PSA alone failed.[41] Moreover, the introduction of larger structures and structures with greater flexibility showed that the partitioned total surface areas (PTSAs, the surface area of the molecule occupied by a specific atom) and/or descriptors related to the flexibility of the molecule are also needed in the permeability predictions.[42,43]

Electrotopological indices have been used to predict permeability computationally (Table 3). The electrotopological descriptors are not always easily understood, even though they can be attributed to describe hydrophobicity, hydrophilicity, and size. Other typical 2D-generated descriptors are related to dispersion forces, polarizability, solute molar volume, and hydrogen bonding acidity and basicity.[44–47] Descriptors such as $\log P_{oct}$/$\log D_{oct}$, polarizability, polarity, Lewis base and acid strength, and the number and strength of hydrogen bond donors/acceptors obtained from quantum mechanics have also been correlated to permeability.[42,48,49] These descriptors did show high accuracy in the prediction, even though less complex and more rapidly calculated descriptors were almost as accurate. Thus, as quantum mechanics descriptors do not outperforming the fragment-based descriptors with respect to accuracy, they will not be usable in the drug-discovery setting until such calculations become faster.

Table 3. Quantitative in silico models based on Caco-2 permeability values or human fraction absorbed (FA) data.[a]

Response	Descriptor Type	Statistical Method	R^2[b]	n_{tr}[c]	n_{te}[d]	Ref.
Caco-2 P_{app}[e]	number of H bonds	LR[j]	0.94	10	0	[36]
Caco-2 P_{app}	PWASA[h]	LR	0.98	11	0	[38]
Caco-2 P_{app}	PSA[i]	SR[k]	0.96	9	0	[68]
Caco-2 P_{app}	molecular surface areas	MLR[l]	0.96	19	0	[41]
Caco-2 P_{app}	solute and solvation related	MLR	0.86	30	8	[69]
Caco-2 P_{app}	PSA, lipophilicity, size, and flexibility	MLR	0.71	77	23	[70]
Caco-2 P_{app}	H bond capacity, lipophilicity, and size	MLR	0.71	33	12	[71]
Caco-2 P_{app}	H bond strength and electrostatics	PLS[m]	0.85	9	8	[48]
Caco-2 P_{app}	H bond capacity, lipophilicity, size, and flexibility	PLS	0.80	16	0	[72]
Caco-2 P_{app}	H bond capacity and lipophilicity	PLS	0.92	11	0	[73]
Caco-2 P_{app}	size, surface tension, and dielectric constant	PLS	0.90	16	0	[74]
Caco-2 P_{app}	electrotopological indices	PLS	0.71	17	10	[42]
Caco-2 P_{app}	H bond strength and electrostatics	PLS	0.87	17	10	[42]
Caco-2 P_{app}	surface areas	PLS	0.93	17	10	[42]
Caco-2 P_{app}	electrotopological indices	PLS	0.91	9	8	[75]
Caco-2 P_{app}	surface areas	PLS	0.93	13	10	[55]
Caco-2 P_{app}	H bond capacity, PSA, and charge	PLS	0.83	20	10	[76]
Caco-2 P_{app}	H bond capacity, charge, polarizability, and dipole moment	NN[n]	0.62	87	0	[77]
Caco-2 P_c[f]	PSA	SR	0.91	9	0	[39]
Caco-2 actp[g] (peptides)	size, electrostatics, and flexibility	PLS	0.75	20	0	[78]
Caco-2 actp (peptides)	electrotopological indices	PLS	0.92	20	0	[78]
Caco-2 P_{app}	H bond capacity, charge, PSA, and electrotopological indices	PLS	0.88	20	10	[76]
FA	PSA	SR	0.94	20	0	[37]
FA	PSA	SR	0.91	20	0	[68]
FA	structural fragments	MLR	0.79	417	50	[28]
FA	H bond capacity, lipophilicity, size, and flexibility	PLS	0.50	85	0	[72]
FA	H bond capacity and lipophilicity	PLS	0.93	74	0	[73]
FA	electrotopological indices	PLS	0.83	13	7	[75]
FA	H bond capacity, size, and flexibility	NN	0.87	76	10	[50]
FA	H bond capacity, flexibility, and hydrophobicity	NN	0.86	76	10	[29]

[a] Compilation of descriptors, size of datasets, statistical models used, and accuracy of published in silico absorption models. Several classification models can be found in the literature which are regarded as qualitative models and are therefore not reported. Caco-2 and FA data were selected for the compilation, as these are the main responses used in the development of computational models. However, other responses such as permeability in 2/4/A1 cell monolayers, artificial membranes, and the MDCK cell line have also been used in computational model development. [b] Coefficient of determination. [c] Number of compounds in training set. [d] Number of compounds in test set. [e] Apparent permeability. [f] Cellular permeability. [g] Active transport. [h] Polar water accessible surface area. [i] Polar surface area. [j] Linear regression. [k] Sigmoidal regression. [l] Multiple linear regression. [m] Partial least squares projection to latent structures. [n] Neural network.

4.3.2. Factors that influence the accuracy of computational permeability models

Most published models are based on experimentally determined permeability data in Caco-2 cell monolayers. However, models based on FA (human intestinal absorption) have also been developed. The descriptors used in these models are of the same type as found in the cell-based models. However, the response parameters used generally show large variability depending on the methodology used to determine the FA in humans and the inter-individual variability (see section 4.1.3), and hence the accuracy of the model obtained is heavily influenced. Even for datasets in which the compounds were carefully selected to use only passive diffusion to permeate the intestinal cell membrane,[50] it later became evident that some of the compounds included also have an active component as part of their transport mechanism. The quality of the response parameters can also vary for the datasets used in permeability models based on cell lines. Permeability values obtained for the same compound using the same cell line in different laboratories will differ in their absolute numbers due to effects of cell culture protocols and experimental procedures during the measurements. Hence, the dataset used for training and evaluation should be determined within the same laboratory using the same experimental protocol. However, classification models might be based on compiled data, as measurements in the different laboratories, in general, will result in the same ranking of compounds; the compounds will be correctly sorted as poor-, intermediate-, or high-permeability compounds even though the absolute numbers between the laboratories may differ greatly.

Other important factors that influence the accuracy and applicability of the model are the chemical diversity of the training set used in the model development, the statistical tools used in the development, and the transport mechanisms included in the response parameter. These will influence the models as follows: to be generally applicable and highly accurate in the prediction of drug permeability, the training set used should cover a large volume of the druglike space. If a model applicable for a specific therapeutic class is warranted, the training set should be focused on this region of the druglike space. In any of these scenarios, the most important fact

to bear in mind is that the training set should be representative for the type of compounds that are to be predicted; if a model is to predict the permeability of drugs, then druglike molecules must be used in the model development. Regarding the statistical tool used, it is important to select a statistical and mathematical tool that is sound. Hence, the data has to be pre-analyzed so that linear versus nonlinear methods are correctly selected. Finally, it is difficult to obtain transparent and interpretable models if all different kinds of transport routes are included in the measured permeability value. Ideally, separate models are developed for passive transcellular diffusion, passive paracellular diffusion, and for each of the transport proteins that can be used. After the establishment of these models, pharmaceutical informatics tools are used to extract the information on the apparent permeability through the intestinal wall.

In plots of permeability versus FA, different cell models will result in largely different slopes and ranges of the respective permeability curve. All the cell models have relatively steep slopes as a common feature, as exemplified in Figure 7. The 2/4/A1 cell line has the steepest slope and highest apparent permeability values of the two cell lines which is in good agreement with the values obtained in human perfusion studies.[25] The steep slopes of these model systems result in that the in silico models based on these data are good at discriminating high permeability from low permeability. However, a small difference in predicted permeability relative to the experimental value in the region of the slope may shift the compound from being predicted as having intermediate permeability to being either highly or poorly permeable. Hence, the predictions in the mid-range of the permeability values are much more difficult to interpret and draw conclusions from regarding further development.

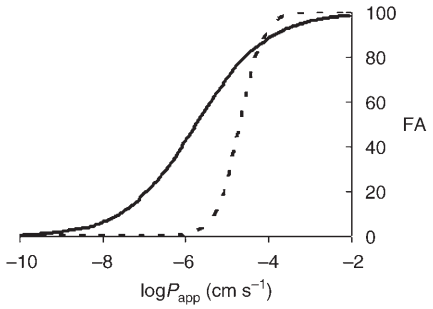


Figure 7. Permeability versus human fraction absorbed (FA): The range and slope of the apparent permeability values obtained from different cell models used for in vitro studies of absorption differ largely, as exemplified with Caco-2 (——) and 2/4/A1 (-----) cell-permeability values. Adapted from Ref. [76] with permission, Copyright© American Chemical Society, 2005.

4.4. A computer-based biopharmaceutical classification system

The biopharmaceutics classification system (BCS) is one way of gathering information on drug absorption.[51] According to the BCS, compounds can be sorted into four classes depending on their solubility and permeability: class I compounds, high solubility and high permeability; class II compounds, poor solubility and high permeability; class III compounds, high solubility and poor permeability; and class IV compounds, poor solubility and poor permeability. High solubility is defined as the maximum oral dose given being soluble in 250 mL within a pH range of 1–7.5; otherwise compounds are classified as having low solubility. High permeability is defined as $\geq$90% absorbed, any value less than this is considered low.[9] If a compound is categorized as a class I compound, no further clinical studies need to be performed after minor changes in the formulation. Various cutoff values for the BCS have been previously applied as qualitative screening tools for drug absorption in drug discovery and development.[9,52,53] Recently, a semi-experimental study using published solubility data in combination with FA data predicted from the calculated logP_{oct} correctly sorted 65% of a series of 29 compounds.[54] If a computer-based BCS with high accuracy in the prediction of the absorption characteristics were devised, it would be possible to sort compounds in accordance with their developability in terms of absorption prior to synthesis. Such virtual tools applied in early drug discovery would result in a decreased number of CDs with formulation problems.

In a recent study we used a BCS with six classes, according to which the solubility was classified as either "low" or "high" in accordance with the cutoff values set by the U.S. Food and Drug Administration (FDA), and the permeability was classified as "low" (FA $<$ 20%), "intermediate" (20% $<$ FA $<$ 80%) or "high" (FA $>$ 80%).[55] This classification was chosen because we believe it provides a better tool for ranking compound absorption in drug discovery than the stricter permeability classification provided by the FDA. Experimental determinations of the Caco-2 permeability and intrinsic solubility were performed in-house, and PLS in silico models based on PTSAs were derived. In comparison with the experimentally determined data, the combination of the two in silico models resulted in 87% of the compounds being sorted into the correct class. The compounds included in a reference test set given by the FDA were correctly sorted with an accuracy of 77%. In summary, these results indicate that more sophisticated in silico models that combine computational analysis of solubility and permeability can successfully estimate the absorption process both qualitatively and quantitatively.[55]

4.5. In silico toxicity models

Toxicology is a rather different matter compared with the other ADME disciplines because many different mechanisms may be involved. Thus, although they appear to be rather similar, the compounds of the investigated dataset may be subject to different toxicological mechanisms that, in turn, give rise to different types of toxicological responses. A large number of papers have been published over the years with proposed models (relationships) that relate molecular structure to a toxicological endpoint of some sort. Three good published starting points with respect to the present state of in silico toxicology statistical modeling are reported by Green,[56] Schultz et al.,[57]

and Dearden.[58] The work by Green is an update on the various software that is present for prediction toxicology, such as DEREK, OncoLogic, HazardExpert, COMPACT, multi-CASE, and TOPCAT, whereas the report by Schultz et al. focuses on quantitative structure–activity relationships (QSARs) in toxicology. Toxicological endpoints that are referred to in this investigation are acute and aquatic toxicity, receptor-mediated toxicity, mutagenicity and carcinogenicity, skin sensitization, and skin and eye irritation. The article by Dearden deals with software as well as references to some specific toxicological QSAR investigations related to endpoints such as cytotoxicity, drug resistance, and skin permeability. Useful reading for the development of QSARs in toxicology from a historical perspective was published by Schultz, Cronin, and co-workers.[12] Within the area of modeling QSARs, including pharmacophore approaches, several articles have appeared in recent years. A more QSAR-related article for cytochrome P450s was published by Lewis et al.[59] Relationships between binding affinities related to various binding site interactions such as hydrogen bonding and π–π stacking but also to parameters related to hydrophobicity, (i.e. log *P* and log *D*) have been developed. An extensive review article related to QSARs of cytochrome P450s was recently published by Hansch and co-workers.[60] A large number of P450 endpoints and datasets for which QSARs have been investigated are presented in this review article. A slight drawback with many of the P450 datasets in this review is that they are relatively small in size. Typically, many P450 datasets contain 7–15 compounds, and the largest investigated dataset contains only 28 members. Although useful for elucidating important properties and possibly rendering some mechanistic insight in fortunate cases, the resulting statistical models are rather local in character with a small applicability domain. The practical use of these models for predicting the behavior of new and virtual sets of compounds may therefore be of limited value. Lately, additional considerations with respect to hERG have entered into the drug-development scenario, owing to the severe consequences associated with hERG interaction such as QT interval prolongation. Avoiding interactions with hERG has become a top priority for many pharmaceutical companies as a consequence of the increased attention to this issue by the FDA and regulatory agencies in other countries. Only a few studies on hERG SARs have been published so far, and much work is currently being conducted to identify properties and/or structural entities that cause hERG channel inhibition. One model of hERG inhibition based on the KcsA crystal structure has been published, while the other models are ligand-based, with 3D QSAR techniques such as CoMFA, CoMSIA, and Catalyst. Recently, 2D QSAR descriptions using both more traditional variables as well as holograms have been used to derive models for hERG inhibition. For a recent minireview, see the work by Norinder.[61] Again, the publicly available training sets for developing in silico models for hERG are rather limited in size, which restrain these models with respect to predictive ability for estimating inhibition of new compounds.

An interesting article published by Stouch et al. addresses some cases in which ADME/Tox models fail and the reasons for these failures.[62] In some cases, the failure is related to the intended use of the in silico model and the expectations of the users of the model. In other cases, failures are related to development aspects of the model, such as choice of statistical tool and description of the investigated structures, as well as limited model validation. The work by Feng and co-workers has benchmarked some different sets of descriptors, such as constitutional descriptors (CONS), topological information indices (TI), BCUT parameters, and some fragment (fingerprint) descriptors (FRAG), as well as statistical methods, such as recursive partitioning (RP), artificial neural networks (ANN), and partial least squares (PLS) on four different datasets with different toxicological endpoints.[63] They found that three combinations, BCUT and RP, FRAG and PLS, and FRAG and RP worked better than expected, whereas two combinations, BCUT and NN together with TI and RP worked somewhat worse than expected. The fact that fragment (fingerprint) descriptors seem to work well is not too surprising, as the concept of toxicophores has been used for quite some time in explaining the toxicological behavior of compounds. At the same time, however, the authors of the article also state that for large datasets, there is a clear need for the development of new descriptors and/or statistical methods.

5. Future Development and Conclusions

To improve solubility, permeability, and toxicity predictions further, a number of actions are needed. First, as mentioned above, focus should be set on the datasets used for the training of the in silico models. The compounds included in the model development and validation need to be representative for the application of the model. Hence, if a general in silico model is to be developed, a large dataset (i.e. hundreds of compounds) with a chemical diversity covering the volume of the druglike space should be used. On the other hand, if a model applicable for the prediction of a specific subset is warranted, focus should be set on this region of the druglike space to improve the accuracy of the model. Second, the experimental setting needs to be standardized and the experimental values used in the model development should be consistently determined using one type of assay. Only high-quality data should be incorporated to minimize the effect of noise on the model. Third, the models should be simplified as much as possible. In our opinion, in terms of permeability, it is therefore better to develop several mechanism-based models that reveal, for example, the extent of the passive transcellular and/or paracellular transport and eventual binding to important transport proteins. Finally, to extract information from such different models to transfer the computational predictions to approximations of the in vivo behavior, new data-mining tools need to be devised (Figure 8). The need for such tools for pharmaceutical informatics is exemplified by the absorption process per se. The extent to which a compound is absorbed will be dependent on its dissolution rate, stability (chemical and enzymatic), solubility, and permeability (passive transcellular component, passive paracellular component, active influx, and active efflux). For each component in the ADMET screen,

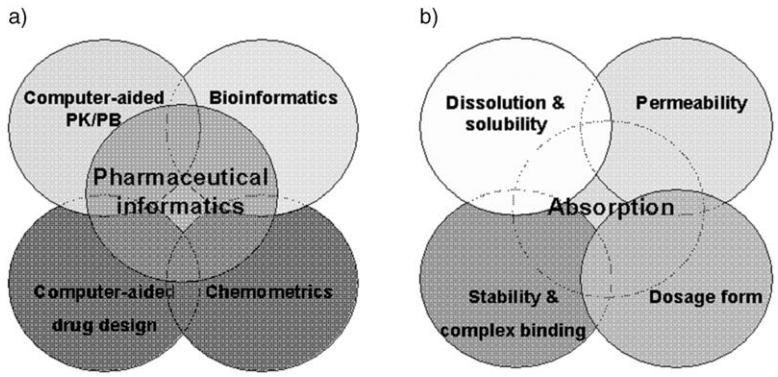


Figure 8. a) To improve the drug-discovery setting, the development of informatics tools suitable for virtual pharmaceutical screening are highly desirable. Such tools need to be able to extract important information related to each of the main areas investigated during the drug-discovery and early development process (PK/PB = pharmacokinetics/pharmacodynamics). b) Each of the main areas are further divided into properties with subgroups. Absorption is one of the properties included in computer-aided PK/PB, and, as an example, the subgroups of absorption are shown. These subgroups may cooperate, counteract, or be independent of each other. Furthermore, both qualitative and quantitative information is compiled in the screening, further stressing the importance of the development of software specific for this application.

the same scenario is valid; that is, a large number of in silico models need to be devised to predict each of the ADMET components. Hence, one of the future challenges will be the development of user-friendly, transparent, and rapid data-mining tools that allow pharmaceutical informatics to be performed in early drug discovery. If such computational tools are devised and highly accurate in silico models of ADMET properties applicable to the druglike space are developed, the prerequisites for a successful virtual drug-discovery setting are given.

Acknowledgements

C.B. acknowledges financial support from the Knut and Alice Wallenberg foundation and the Swedish Fund for Research without Animal Experiments.

Keywords: absorption · ADMET · in silico · predictions · solubility

[1] "Managing the drug discovery/development interface": T. Kennedy, *Drug Discovery Today* **1997**, *2*, 436 – 444.
[2] "Progress in computational methods for the prediction of ADMET properties": D. E. Clark, P. D. Grootenhuis, *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 382 – 390.
[3] "Computational approaches to the understanding of ADMET properties and problems": S. Modi, *Drug Discovery Today* **2003**, *8*, 621 – 623.
[4] "ADMET in silico modelling: towards prediction paradise?": H. van de Waterbeemd, E. Gifford, *Nat. Rev. Drug Discovery* **2003**, *2*, 192 – 204.
[5] "Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells": P. Artursson, J. Karlsson, *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880 – 885.
[6] "Intestinal drug absorption and metabolism in cell cultures: Caco-2 and beyond": P. Artursson, R. T. Borchardt, *Pharm. Res.* **1997**, *14*, 1655 – 1658.
[7] "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings": C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeny, *Adv. Drug Delivery Rev.* **1997**, *23*, 3 – 25.
[8] "Can the pharmaceutical industry reduce attrition rates?": I. Kola, J. Landis, *Nat. Rev. Drug Discovery* **2004**, *3*, 711 – 716.
[9] "Drug-like properties and the causes of poor solubility and poor permeability": C. A. Lipinski, *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235 – 249.
[10] "Molecular descriptors influencing melting point and their role in classification of solid drugs": C. A. S. Bergström, U. Norinder, K. Luthman, P. Artursson, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177 – 1185.
[11] "Contribution of solvent drag through intercellular junctions to absorption of nutrients by the small intestine of the rat": J. R. Pappenheimer, K. Z. Reiss, *J. Membr. Biol.* **1987**, *100*, 123 – 136.
[12] "Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective": T. W. Schultz, M. T. D. Cronin, J. D. Walker, A. O. Aptula, *J. Mol. Struct.* **2003**, *622*, 1 – 22.
[13] "Physicochemical high-throughput screening: parallel artificial membrane permeation assay in the description of passive absorption processes": M. Kansy, F. Senner, K. Gubernator, *J. Med. Chem.* **1998**, *41*, 1007 – 1010.
[14] "Pitfalls in QSAR": T. W. Schultz, M. T. D. Cronin, *J. Mol. Struct.* **2003**, *622*, 39 – 51.
[15] "The rate of solution of solid substances in their own solutions": A. A. Noyes, W. R. Whitney, *J. Am. Chem. Soc.* **1897**, *19*, 930 – 934.
[16] "The calculation of the hydrogen number of the blood from the free and bound carbon dioxide of the same and the binding of oxygen by the blood as a function of the hydrogen number": K. A. Hasselbalch, *Biochem. Z.* **1916**, *78*, 112 – 144.
[17] "Comparison of the gastrointestinal anatomy, physiology, and biochemistry of humans and commonly used laboratory animals": T. T. Kararli, *Biopharm. Drug Dispos.* **1995**, *16*, 351 – 380.
[18] "Common ion equilibriums of hydrochloride salts and the Setschenow equation": J. B. Bogardus, *J. Pharm. Sci.* **1982**, *71*, 588 – 590.
[19] "Aqueous solubility of diclofenac diethylamine in the presence of pharmaceutical additives: a comparative study with diclofenac sodium": E. Khalil, S. Najjar, A. Sallam, *Drug Dev. Ind. Pharm.* **2000**, *26*, 375 – 381.
[20] "Mechanism of protein salting in and salting out by divalent cation salts: balance between hydration and salt binding": T. Arakawa, S. N. Timasheff, *Biochemistry* **1984**, *23*, 5912 – 5923.
[21] "Physicochemical and physiological mechanisms for the effects of food on drug absorption: the role of lipids and pH": W. N. Charman, C. J. Porter, S. Mithani, J. B. Dressman, *J. Pharm. Sci.* **1997**, *86*, 269 – 282.
[22] "Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability": I. J. Hidalgo, T. J. Raub, R. T. Borchardt, *Gastroenterology* **1989**, *96*, 736 – 749.
[23] "Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells": P. Artursson, *J. Pharm. Sci.* **1990**, *79*, 476 – 482.
[24] "MDCK (Madin–Darby canine kidney) cells: A tool for membrane permeability screening": J. D. Irvine, L. Takahashi, K. Lockhart, J. Cheong, J. W. Tolan, H. E. Selick, J. R. Grove, *J. Pharm. Sci.* **1999**, *88*, 28 – 33.
[25] "A conditionally immortalized epithelial cell line for studies of intestinal drug transport": S. Tavelin, V. Milovic, G. Ocklind, S. Olsson, P. Artursson, *J. Pharmacol. Exp. Ther.* **1999**, *290*, 1212 – 1221.
[26] "Regional jejunal perfusion, a new in vivo approach to study oral drug absorption in man": H. Lennernäs, Ö. Ahrenstedt, R. Hällgren, L. Knutson, M. Ryde, L. Paalzow, *Pharm. Res.* **1992**, *9*, 1243 – 1251.
[27] "Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors": Y. H. Zhao, J. Le, M. H. Abraham, A. Hersey, P. J.

Eddershaw, C. N. Luscombe, D. Boutina, G. Beck, B. Sherborne, I. Cooper, J. A. Platts, *J. Pharm. Sci.* **2001**, *90*, 749 – 784.
[28] "ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans": G. Klopman, L. R. Stefan, R. D. Saiakhov, *Eur. J. Pharm. Sci.* **2002**, *17*, 253 – 263.
[29] "Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures": T. Niwa, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113 – 119.
[30] "A topological sub-structural approach for predicting human intestinal absorption of drugs": M. A. Perez, M. B. Sanz, L. R. Torres, R. G. Avalos, M. P. Gonzalez, H. G. Diaz, *Eur. J. Med. Chem.* **2004**, *39*, 905 – 916.
[31] "Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology": J. Huuskonen, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773 – 777.
[32] "Prediction of drug solubility from structure": W. L. Jorgensen, E. M. Duffy, *Adv. Drug Delivery Rev.* **2002**, *54*, 355 – 366.
[33] "A consensus neural network-based technique for discriminating soluble and poorly soluble compounds": D. T. Manallack, B. G. Tehan, E. Gancia, B. D. Hudson, M. G. Ford, D. J. Livingstone, D. C. Whitley, W. R. Pitt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674 – 679.
[34] "Prediction of aqueous solubility using rule-based systems": U. Norinder, P. Lidén, H. Boström, (RDS, www.compumine.com) and ensemble modeling, unpublished results.
[35] "Simulation of water transport through a lipid membrane": S. J. Marrink, H. J. C. Berendsen, *J. Phys. Chem.* **1994**, *98*, 4155 – 4168.
[36] "The influence of peptide structure on transport across Caco-2 cells. II. Peptide bond modification which results in improved permeability": R. A. Conradi, A. R. Hilgers, N. F. Ho, P. S. Burton, *Pharm. Res.* **1992**, *9*, 435 – 439.
[37] "Polar molecular surface properties predict the intestinal absorption of drugs in humans": K. Palm, P. Stenberg, K. Luthman, P. Artursson, *Pharm. Res.* **1997**, *14*, 568 – 571.
[38] "Predicting drug absorption from molecular surface properties based on molecular dynamics simulations": L. Hjorth Krarup, I. Thøger Christensen, L. Hovgaard, S. Frokjaer, *Pharm. Res.* **1998**, *15*, 972 – 978.
[39] "Evaluation of dynamic polar molecular surface area as predictor of drug absorption: Comparison with other computational and experimental predictors": K. Palm, K. Luthman, A. L. Ungell, G. Strandlund, F. Beigi, P. Lundahl, P. Artursson, *J. Med. Chem.* **1998**, *41*, 5382 – 5392.
[40] "Estimation of permeability by passive diffusion through Caco-2 cell monolayers using the drugs' lipophilicity and molecular weight": G. Camenisch, J. Alsenz, H. van de Waterbeemd, G. Folkers, *Eur. J. Pharm. Sci.* **1998**, *6*, 313 – 319.
[41] "Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties": P. Stenberg, K. Luthman, P. Artursson, *Pharm. Res.* **1999**, *16*, 205 – 212.
[42] "Experimental and computational screening models for the prediction of intestinal drug absorption": P. Stenberg, U. Norinder, K. Luthman, P. Artursson, *J. Med. Chem.* **2001**, *44*, 1927 – 1937.
[43] "Molecular properties that influence the oral bioavailability of drug candidates": D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615 – 2623.
[44] "Solubility properties in biological media 9: prediction of solubility and partition of organic nonelectrolytes in blood and tissues from solvatochromic parameters": M. J. Kamlet, R. M. Doherty, V. Fiserova-Bergerova, P. W. Carr, M. H. Abraham, R. W. Taft, *J. Pharm. Sci.* **1987**, *76*, 14 – 17.
[45] "Molecular factors influencing drug transfer across the blood–brain barrier": J. A. Gratton, M. H. Abraham, M. W. Bradbury, H. S. Chadha, *J. Pharm. Pharmacol.* **1997**, *49*, 1211 – 1216.
[46] "On the mechanism of human intestinal absorption": M. H. Abraham, Y. H. Zhao, J. Le, A. Hersey, C. N. Luscombe, D. P. Reynolds, G. Beck, B. Sherborne, I. Cooper, *Eur. J. Med. Chem.* **2002**, **37**, 595 – 605.
[47] "SLIPPER-2001—Software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity": O. A. Raevsky, S. V. Trepalin, H. P. Trepalina, V. A. Gerasimenko, O. E. Raevskaja, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540 – 549.
[48] "Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parameterization and PLS statistics": U. Norinder, T. Österberg, P. Artursson, *Pharm. Res.* **1997**, *14*, 1786 – 1791.
[49] "Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parameterization and PLS statistics": U. Norinder, T. Österberg, P. Artursson, *Eur. J. Pharm. Sci.* **1999**, *8*, 49 – 56.
[50] "Prediction of human intestinal absorption of drug compounds from molecular structure": M. D. Wessel, P. C. Jurs, J. W. Tolan, S. M. Muskal, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726 – 735.
[51] "A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability": G. L. Amidon, H. Lennernäs, V. P. Shah, J. R. Crison, *J. Pharm. Res.* **1995**, *12*, 413 – 420.
[52] "HT29-MTX/Caco-2 co-cultures as an in vitro model for the intestinal epithelium: in vitro–in vivo correlation with permeability data from rats and humans": E. Walter, S. Janich, B. J. Roessler, J. M. Hilfinger, G. L. J. Amidon, *Pharm. Sci.* **1996**, *85*, 1070 – 1076.
[53] "Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach": S. Winiwarter, N. M. Bonham, F. Ax, A. Hallberg, H. Lennernäs, A. Karlén, *J. Med. Chem.* **1998**, *41*, 4939 – 4949.
[54] "Molecular properties of WHO essential drugs and provisional biopharmaceutical classification": N. A. Kasim, M. Whitehouse, C. Ramachandran, M. Bermejo, H. Lennernäs, A. S. Hussain, H. E. Junginger, S. A. Stavchansky, K. K. Midha, V. P. Shah, G. L. Amidon, *Mol. Pharm.* **2004**, *1*, 85 – 96.
[55] "Absorption classification of oral drugs based on molecular surface properties": C. A. S. Bergström, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, P. Artursson, *J. Med. Chem.* **2003**, *46*, 558 – 570.
[56] "Computer systems for the prediction of toxicity: An update": N. Green, *Adv. Drug Delivery Rev.* **2002**, *54*, 417 – 431.
[57] "The present status of QSAR in toxicology": T. W. Schultz, M. T. D. Cronin, T. I. Netzeva, *J. Mol. Struct.* **2003**, *622*, 23 – 38.
[58] "In silico prediction of drug toxicity": J. C. Dearden, *J. Comput.-Aided Mol. Des.* **2003**, *17*, 119 – 127.
[59] "Quantitative structure–activity relationships (QSARs) within substrates of human cytochromes P450 involved in drug metabolism": D. F. V. Lewis, S. Modi, M. Dickins, *Drug Metab. Drug Interact.* **2001**, *18*, 221 – 242.
[60] "QSAR of cytochromes P450": C. Hansch, S. B. Mekapati, A. Karup, R. P. Verma, *Drug Metab. Rev.* **2004**, *36*, 105 – 156.
[61] "In silico modelling of ADMET—a minireview of work 2000 to 2004": U. Norinder, *SAR QSAR Environ. Res.* **2005**, *16*, 1 – 11.
[62] "In silico ADME/Tox: why models fail": T. R. Stouch, J. R. Kenyon, S. R. Johnson, X.-Q. Chen, A. Doweyko, Y. Li, *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83 – 92.
[63] "Predictive toxicology: benchmarking molecular descriptors and statistical methods": J. Feng, L. Lurati, H. Ouyang, T. Robinson, Y. Wang, S. Yuan, S. S Young, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1463 – 1470.
[64] "Prediction of aqueous solubility of organic compounds based on a 3D structure representation": A. Yan, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429 – 434.
[65] "Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility": R. Liu S.-S. So, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633 – 1639.
[66] "Estimation of aqueous solubility of chemical compounds using e-state indices": I. V. Tetko, V. Yu. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488 – 1493.
[67] "Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method": J. K. Wegner, A. Zell, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077 – 1084.
[68] "Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties": P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.* **2000**, *43*, 3714 – 3717.
[69] "Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis": A. Kulkarni, Y. Han, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 331 – 342.
[70] "ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties": T. J. Hou, W. Zhang, K. Xia, X. B. Qiao, X. J. Xu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1585 – 1600.
[71] "A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell cluture": P. Y. Marrero, P. M. A.

Cabrera, Z. V. Romero, D. H. Gonzalez, F. Torrens, *J. Pharm. Pharm. Sci.* **2004**, *7*, 186 – 199.
[72] "Toward minimalistic modeling of oral drug absorption": T. I. Oprea, J. Gottfries, *J. Mol. Graph. Model.* **1999**, *17*, 261 – 274.
[73] "Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics": T. Österberg, U. Norinder, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1408 – 1411.
[74] "Prediction of drug transport processes using simple parameters and PLS statistics. The use of ACD/log *P* and ACD/ChemSketch descriptors": T. Österberg, U. Norinder, *Eur. J. Pharm. Sci.* **2001**, *12*, 327 – 337.
[75] "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": U. Norinder, T. Österberg, *J. Pharm. Sci.* **2001**, *90*, 1076 – 1085.
[76] "Exploring the role of different drug transport routes in permeability screening": P. Matsson, C. A. S. Bergström, N. Nagahara, S. Tavelin, U. Norinder, P. Artursson, *J. Med. Chem.* **2005**, *48*, 604 – 613.
[77] "Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network": S. Fujiwara, F. Yamashita, M. Hashida, *Int. J. Pharm.* **2002**, *237*, 95 – 105.
[78] "Two- and three-dimensional QSAR of carrier-mediated transport of β-lactam antibiotics in Caco-2 cells": S. Wanchana, F. Yamashita, H. Hara, S. Fujiwara, M. Akamatsu, M. Hashida, *J. Pharm. Sci.* **2004**, *93*, 3057 – 3065.

Received: June 26, 2006

