molecular pharmaceutics

articles

Combinatorial QSAR Modeling of Human Intestinal Absorption

Claudia Suenderhauf,[†] Felix Hammann,[†] Andreas Maunz,[‡] Christoph Helma,[‡,§] and Jörg Huwyler*[,†]

Division of Pharmaceutical Technology, Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50, CH-4056 Basel, Switzerland, Freiburger Zentrum für Datenanalyse und Modellbildung, University Freiburg, Hermann Herder Strasse 3a, D-70104 Freiburg, Germany, and In silico toxicology, Altkircherstrasse 3a, CH-4054 Basel, Switzerland

Received August 23, 2010; Revised Manuscript Received December 6, 2010; Accepted December 10, 2010

Abstract: Intestinal drug absorption in humans is a central topic in drug discovery. In this study, we use a broad selection of machine learning and statistical methods for the classification and numerical prediction of this key end point. Our data set is based on a selection of 458 small druglike compounds with FDA approval. Using easily available tools, we calculated one- to three-dimensional physicochemical descriptors and used various methods of feature selection (best-first backward selection, correlation analysis, and decision tree analysis). We then used decision tree induction (DTI), fragment-based lazy-learning (LAZAR), support vector machine classification, multilayer perceptrons, random forests, *k*-nearest neighbor and Naïve Bayes analysis to model absorption ratios and binary classification (well-absorbed and poorly absorbed compounds). Best performance for classification was seen with DTI using the chi-squared analysis interaction detector (CHAID) algorithm, yielding corrected classification rate of 88% (Matthews correlation coefficient of 75%). In numeric predictions, the multilayer perceptron performed best, achieving a root mean squared error of 25.823 and a coefficient of determination of 0.6. In line with current understanding is the importance of descriptors such as lipophilic partition coefficients (log *P*) and hydrogen bonding. However, we are able to highlight the utility of gravitational indices and moments of inertia, reflecting the role of structural symmetry in oral absorption. Our models are based on a diverse data set of marketed drugs representing a broad chemical space. These models therefore contribute substantially to the molecular understanding of human intestinal drug absorption and qualify for a generalized use in drug discovery and lead optimization.

Keywords: QSAR; intestinal drug absorption; molecular modeling; machine learning; decision tree induction; artificial neural network; support vector machines; naive Bayes; receiver operating characteristics

1. Introduction

Oral administration of drugs is a cost-effective and often preferred route of administration. It is warranted for many indications, and associated with high patient compliance. However, low intestinal absorption of a drug may limit its clinical application, except in settings where the compound's target lies within the gastrointestinal lumen (e.g., vancomycin, mesalazine).

Limiting factors for small molecule drug absorption include low solubility or chemical instability in the gastrointestinal tract (GIT),[1] high gastrointestinal metabolism, and poor intestinal membrane permeability. Absorption kinetics are highly dependent on a compound's solubility

* Corresponding author. Mailing address: University of Basel, Institute of Pharmaceutical Technology, Klingelbergstrasse 50, CH-4056 Basel, Switzerland. E-mail: joerg.huwyler@unibas.ch. Telephone: +41 61 267 15 13. Fax: +41 61 267 15 16.

[†] University of Basel.

[‡] University Freiburg.

[§] In silico toxicology.

10.1021/mp100279d © 2011 American Chemical Society
Published on Web 12/10/2010

and hence galenic formulation, which influences exact location of dosage form disintegration in the GIT.[2] Moreover, permeability is highly dependent on hydrogen bonding capacity and lipophilicity.[3,4] These features play key roles for the passage from the aqueous intestinal environment through the cellular barrier of the gut wall. Generally, the higher the hydrogen bonding capacity, the more energy the permeation will cost and consequently the poorer the molecule's absorption.

Current understanding indicates that passive diffusion (transcellular and paracellular) is the determining factor in drug absorption.[5] Active efflux and influx at the level of the enterocyte are regulated by several transport systems, like multidrug resistance-associated protein 3 (MRP3)[6] and other well-known efflux transporters like P-glycoprotein.[7−9] The possible effect of enterocytic cytochrome P_{450} (CYP_{450}) metabolism, albeit very small when compared to hepatic CYP_{450}, serves as an example where metabolic degradation of the parent substance results in lower plasma levels.[10,11] Peptide transporters (PEPT1, PEPT2) contribute considerably to active drug absorption as they are among other carrier systems responsible for the uptake of various peptide-like drugs, such as beta-lactam antibiotics, angiotensin converting enzyme inhibitors, antiviral drugs, and anticancer drugs.[12,13]

Traditionally, drugs were discovered by synthesizing and testing compounds in expensive and time-consuming multistep processes against a battery of in vivo biological screening tests. Besides the recent attempts to reduce animal tests, concerns regarding the reliability of such intra- and interspecies comparisons have been put forth.[14,15] It is desirable to have information on metabolic and toxicological behavior in humans as early as possible in drug design to avoid costly late stage failures. Recent advances in information technology and artificial intelligence can be adapted to address these requirements.[16] Under the assumption that similar structures show similar biological activities, extrapolation from characterized compounds to untested molecules seems to be feasible. This fundamental concept is known as dissimilarity principle and forms the basis for quantitative structure activity relationships (QSAR). Various statistical methods can be applied to create more or less predictive models from molecular data. They range from linear regression[17] to recent developments of machine learning, e.g. support vector machines (SVM)[18] or artificial neural networks (ANN).[19]

2. Materials and Methods

2.1. Data Set. The data set used for the present study is based on a list of FDA approved small molecule drugs ($n = 458$) for which experimental data were available and sufficiently documented.[20] Omissions were due to missing

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.

(2) Macheras, P.; Reppas, C.; Dressman, J. B. *Biopharmaceutics of orally Administred Dosage Forms*: Ellis Horwood Ltd.: Chichester, U.K., 1995.

(3) Winiwarter, S.; Ax, F.; Lennernas, H.; Hallberg, A.; Pettersson, C.; Karlen, A. Hydrogen bonding descriptors in the prediction of human in vivo intestinal permeability. *J. Mol. Graphics Modell.* **2003**, *21*, 273–287.

(4) Kerns, E. H.; Di, L. *Drug-like Properties: Concepts, Structure Design and Methods. From ADME to Toxicity Optimization*: Academic Press: 2008.

(5) Norinder, U.; Osterberg, T.; Artursson, P. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharm. Res.* **1997**, *14*, 1786–1791.

(6) Zimmermann, C.; Gutmann, H.; Hruz, P.; Gutzwiller, J. P.; Beglinger, C.; Drewe, J. Mapping of multidrug resistance gene 1 and multidrug resistance-associated protein isoform 1 to 5 mRNA expression along the human intestinal tract. *Drug Metab. Dispos.* **2005**, *33*, 219–224.

(7) Huwyler, J.; Wright, M. B.; Gutmann, H.; Drewe, J. Induction of cytochrome P450 3A4 and P-glycoprotein by the isoxazolylpenicillin antibiotic flucloxacillin. *Curr. Drug Metab.* **2006**, *7*, 119–126.

(8) Fricker, G.; Drewe, J.; Huwyler, J.; Gutmann, H.; Beglinger, C. Relevance of p-glycoprotein for the enteral absorption of cyclosporin A: in vitro-in vivo correlation. *Br. J. Pharmacol.* **1996**, *118*, 1841–1847.

(9) Dahan, A.; Amidon, G. L. Segmental dependent transport of low permeability compounds along the small intestine due to P-glycoprotein: the role of efflux transport in the oral absorption of BCS class III drugs. *Mol. Pharmaceutics* **2009**, *6*, 19–28.

(10) Benet, L. Z. The drug transporter-metabolism alliance: uncovering and defining the interplay. *Mol. Pharmaceutics* **2009**, *6*, 1631–1643.

(11) Christians, U.; Schmitz, V.; Haschke, M. Functional interactions between P-glycoprotein and CYP3A in drug metabolism. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 641–654.

(12) Brandsch, M.; Knutter, I.; Bosse-Doenecke, E. Pharmaceutical and pharmacological importance of peptide transporters. *J. Pharm. Pharmacol.* **2008**, *60*, 543–585.

(13) Meredith, D.; Price, R. A. Molecular modeling of PepT1—towards a structure. *J. Membr. Biol.* **2006**, *213*, 79–88.

(14) Kuijpers, M. H.; de Jong, W. Spontaneous hypertension in the fawn-hooded rat: a cardiovascular disease model. *J. Hypertens. Suppl.* **1986**, *4*, 41–44.

(15) Kararli, T. T. Comparison of the gastrointestinal anatomy, physiology, and biochemistry of humans and commonly used laboratory animals. *Biopharm. Drug Dispos.* **1995**, *16*, 351–380.

(16) Metcalfe, P. D.; Thomas, S. Challenges in the prediction and modeling of oral absorption and bioavailability. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 104–110.

(17) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernas, H.; Karlen, A. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* **1998**, *41*, 4939–4949.

(18) Hou, T.; Li, Y.; Zhang, W.; Wang, J. Recent developments of in silico predictions of intestinal absorption and oral bioavailability. *Comb. Chem. High Throughput Screening* **2009**, *12*, 497–506.

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

(20) Dollery, C. *Therapeutic Drugs*, 2nd ed.; Churchill Livingstone: Edinburgh, 1999.

Table 1. Ordinal Classes and Corresponding Absorption Values (% Abs)

class label	% Abs	no. of instances
true	$a \geq 80\%$	303
unknown	$30\% < a < 80\%$	82
false	$a \leq 30\%$	73

information. Intestinal absorption (% Abs) is defined as the percentage of the dose absorbed from the gastrointestinal tract following oral administration. This is not necessarily the same as the amount of drug reaching systemic circulation, which is also affected by presystemic metabolism (e.g., hepatic first-pass effect). The arithmetical mean was used wherever an absorption range was given. We also did not omit compounds that are known substrates of efflux transporters such as P-glycoprotein (e.g., digoxin) because insufficient information on specific absorption and excretion pathways was supplied in the original data source.

As classification algorithms require nominal class labels as end points, we recoded the numerical absorption ratios into three different ordinal classes ("TRUE", "UNKNOWN", "FALSE", see Table 1). Thresholds were determined so as to produce a sufficiently large number of instances for each class. To achieve a better separability, members of the class "UNKNOWN" were exempt from classification learning. This class, corresponding to moderately absorbed compounds, was clearly underrepresented in the data source and hence was not deemed suitable for modeling. For numerical predictions, the entire data set was used. Prestudy analysis (data not shown) indicated that data transformation, i.e. linearization, does not improve model performance.

Lastly, the data source provided generic drug names but no structural information. We therefore retrieved the corresponding structures from the National Library of Medicine database PubChem (http://pubchem.ncbi.nlm.nih.gov/). For salts, the counterion was removed prior to further processing.

2.2. Descriptors. Descriptors are (usually numerical) features derived from a chemical structure, e.g. molecular weight and lipid solubility. For this study, we employed the Chemical Development Kit (CDK, Version 1.2.3, http://cdk.sourceforge.net/) to calculate a large set of descriptors. An overview is given in Table 2, and a full list and short explanations are provided in the Supporting Information. The structural information retrieved from PubChem was two-dimensional. For certain descriptors, however, three-dimensional structures are required. We extrapolated these using OpenBabel (version 2.2.3, http://www.openbabel.org/) to perform a search of lowest energy conformers within the "Ghemical" force field (see below).[21]

2.3. ML Techniques. *Decision Tree Induction (DTI).* The basic principle of DTI is the splitting of data on attributes into smaller subsets to obtain higher purity in the resulting child nodes. This procedure is repeated until purity with regard to the property of interest reaches its optimum. How purity is measured depends on the algorithm used. In this study, the chi-squared automatic interactor detector (CHAID)[22] and the classification and regression tree algorithm (CART)[23] were employed. CART can also prune trees, i.e. simplify them through replacement of subtrees with a node representing the majority vote of a subtree. Trees were grown with a maximum depth of five nodes and a minimum five instances in the parent node and two instances in the child node. The Gini coefficient was used as a homogeneity measure for CART.

Table 2. Overview of Descriptors ($n = 80$) Used in This Study[a]

class	type
charge analysis	hydrogen bonding capacity
	charged partial surface descriptors
	partitioning coefficients
	molecular polarizability
constitutional	counts of atoms, rings, and bonds
	length over breadth descriptors
	gravitational indices
	moment of inertia
	molecular weight
topological	eccentric connectivity index
	weighted burden matrix
	Kier−Hall kappa shape indices
	Petitjean number and index
	Wiener path and polarity numbers
	Zagreb index

[a] A detailed listing of all features is given in the Supporting Information.

Random Forest (RF). In 2001, Breimann introduced the principle of random forests (RF),[24] extending the DTI principle by using not only one but multiple unpruned trees with randomly selected feature sets. Predictions are made by majority vote, and the number of trees per forest in this study was set to ten.

Artificial Neuronal Networks (ANN). The ANN paradigm[25] is an abstraction of biological networks of neurons. Instances are represented as vectors containing their features. Each feature is passed to one of the input neurons to which a weight is assigned. Based on these weights, input is passed to the output layer over a number of interspersed optional hidden layers. The output layer combines these signals to produce a result. Initially, weights are set to random values. As the network is repeatedly presented with training instances, these weights are adjusted so that the total output

(21) Hassinen, T.; Peräkylä, M. New energy terms for reduced protein models implemented in an off-lattice force field. *J. Comput. Chem.* **2001**, *22*, 1229–1242.

(22) Sonquist, J. A.; Morgan, J. N. *The detection of Interaction Effects*; Survey research center, University of Michigan: Ann Arbor, 1964; p 296.

(23) Breiman, L. *Classification and regression trees*, 1st ed.; Chapman & Hall/CRC: Boca Raton, 1984.

(24) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(25) McCulloch, W.; Pitts, W. A. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133.

of the network approximates the observed end point values associated with the instances. We used multilayer perceptrons (MP) with up to two hidden layers.

k-Nearest-Neighbor (KNN). KNN models are lazy-learning predictors, which assign new instances to the most common class of known instances in their immediate neighborhood.[26] The concept of neighborhood can be measured in various ways (e.g., Euclidian distance, city-block distance). We used the Euclidian distance measure.

Lazy Structure−Activity Relationships (LAZAR). The LAZAR engine[27] is a lazy-learning fragment-based predictor. Relevant fragments are determined by finding all linear fragments in the training data set (without size limits) and removing those that are statistically insignificant ($p < 0.95$) in the chi-square test. Relevant fragments are used to determine activity specific similarities with a weighted Tanimoto index. LAZAR generates predictions by calculating activity specific similarities for all compounds in the training data set and classifying the unknown compound with a modified *k*-nearest-neighbor (KNN) algorithm.

SVM. Support vector machines[26] are a recent linear classification paradigm, which has proven exceptionally useful for noisy (real-world) data. Models are built by fitting a rigid decision hyperplane with the greatest possible margin between classes. Nonlinear data can be handled by transposing the original feature space to higher dimensionalities using kernels.[28] In this study, we have chosen to use the polynomial kernel

$$K(x_i, x_j) = (\gamma x_i x_j + \text{const})^d \quad (1)$$

and the radial basis function (rbf) kernel

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \quad (2)$$

where γ and d are kept constant.

In SVM with rbf kernels, there are two learning meta-parameters that greatly influence performance (cost (c) and gamma (γ)). Determining these parameters is a heuristic optimization task. We used a grid-search based approach in a log-dimensional space using in-house software.

Naïve Bayes and Bayesian Networks. Under the assumption of variable independence and normal distribution, probabilities are represented as fractions of observation over observation times. They are summed up to compute an overall probability estimate of the end point.[29] Although these preassumptions are rarely given in real life data, Naïve Bayes predicted very strongly and frequently outperformed rather sophisticated machine learning paradigms. Reasons for this surprising efficacy were previously discussed.[30,31] Bayesian networks represent probability distributions as directed acyclic graphs, where each node represents an attributes probability. Predictions are made by summing up probabilities for each instance. For learning the networks presented here, we used the K2 algorithm.[32]

Cross-Validation. To avoid overfitting, we used *k*-fold cross-validation. Here, a data set is randomly divided into *k* subsets.[33] Of these, $k - 1$ sets are recombined to make up a training set, with the resulting model tested against the remaining instances. This process is repeated *k* times until all instances have served both as training and as test data, thereby making sure that no classes are left out. All models were built with $k = 10$, except for LAZAR where $k = n$ (leave-one-out (LOO) cross-validation). Other means of validation are possible (e.g., holdout validation) and have been used in the past (see Table 6).

Chemical Similarity. Structural similarity of compounds can be measured by calculating their proximity in a multidimensional space spanned by their descriptors. A highly dispersed cloud of instances is consequently more heterogeneous than one that is tightly packed. Commonly, similarity is assessed via its reciprocity, the dissimilarity coefficient. We used the Tanimoto coefficient[34] and considered all descriptors ($n = 80$).

$$sim(i,j) = \frac{\sum_{d=1}^{k} X_{di} X_{dj}}{\sum_{d=1}^{k} (X_{di})^2 + \sum_{d=1}^{k} (X_{dj})^2 - \sum_{d=1}^{k} (X_{di} X_{dj})} \quad (3)$$

Here, i and j represent two compounds with k descriptors of value X_d, and the mean similarity is calculated. The reciprocity approaches 1.0 as compound diversity increases.

Feature Reduction. With a great number of descriptors at one's disposal, it is necessary to restrict oneself to a small subset to avoid overfitting (i.e., creating overly complex models with very high predictive accuracy on training data by extracting too many parameters from the known data at the expense of not being able to predict unseen compounds). Ideally, one bases this selection on mechanistic knowledge of the process to be modeled. Alternatively, one may use statistical methods to pick out variables with a high

(26) Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 2002.
(27) Helma, C. Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Mol. Diversity* **2006**, *10*, 147–158.
(28) Aizerman, M.; Braverman, E.; Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition. *Autom. Remote Control* **1964**, *25*, 821–837.
(29) Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London* 1763, *53*, 13−20.
(30) Domingos, P.; Pazzani, M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* **1997**, *29*, 103–130.
(31) Zhang, H. The Optimality of Naive Bayes. *Proc. Seventeenth Fla. Artif. Intell. Res. Soc. Conf.* **2004**, 562–567.
(32) Cooper, G. F.; Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **1992**, *9*, 309–347.
(33) Kohavi, R. A. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. 14th Int. Jt. Conf. Artif. Intell.* **1995**, *2*, 1137–1143.
(34) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

independent explanatory power. In this study, we compare several approaches:

1. Best first feature selection (BFS) using a greedy hill-climbing algorithm.[35]
2. Linear correlation analysis (CFS) by performing linear regressions for every descriptor. The nine best correlating (by measure of R^2) were selected.
3. DTI splitting criteria (DTIS) were used as the final subset. Features were taken from DTI models produced beforehand.

Quality Measures. To estimate the predictive power of our classification models we used the corrected classification rate (CCR) and Matthews correlation coefficient (MCC). These measures give an indication of the performance of binary classification and consider specificity and sensitivity. Estimating predictive power using these formulas has the advantage of also being applicable to skewed data sets (where one class outnumbers greatly the other ones), giving a realistic evaluation of the models. Although these measures are redundant, results can only be converted into each other in presence of a confusion matrix (e.g., knowing the exact number of truly classified positives (TP), truly classified negatives (TN), falsely classified positives (FP) and falsely classified negatives (FN)). Unfortunately, in current studies, usually only one of these is used and thus interstudy comparisons are difficult. To ease comparison of our work with existing models, we always present both figures.

$$\mathrm{CCR} = \frac{1}{2}\left(\frac{T_\mathrm{N}}{N_\mathrm{N}} + \frac{T_\mathrm{P}}{N_\mathrm{P}}\right)$$

$$\mathrm{MCC} = \frac{T_\mathrm{P} \times T_\mathrm{N} - F_\mathrm{P} \times F_\mathrm{N}}{\sqrt{(T_\mathrm{P} + F_\mathrm{P})(T_\mathrm{P} + F_\mathrm{N})(T_\mathrm{N} + F_\mathrm{P})(T_\mathrm{N} + F_\mathrm{N})}}$$

T_N and T_P represent the number of true negative and true positive classified instances, respectively N_N and N_P the total number of negative and positive instances in the model.

For estimating numeric models we used root mean squared error (RMSE) and the coefficient of determination (R^2), which measures the statistical correlation between the predicted and the actual values. The R^2 ranges from 1 for perfectly correlated results to 0, for no correlation at all.

Receiver Operating Characteristics (ROC). Additionally, we analyzed our results with receiver operating characteristics (ROC) curves.[36] ROC graphs are commonly used to evaluate the correlation between binomial outcomes and a continuous variable. For comparison, the area under the ROC curve (AUC) is usually used, and we give it for all recoded numeric predictions. Values of the AUC lie within 0 and 1.0 and exhibit an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen negative instance.

Moreover, ROC curves are often used for optimization of models based on continuous variables. This can be done by determining optimal threshold values or cutoff points. A frequently used measure for this is the Youden index[37] (J), which is defined as

$$J = \text{sensitivity} + \text{specificity} - 1$$

This can be calculated for all points on the ROC curves, with maxima denoting thresholds of optimal sensitivity and specificity for binomial classifications where both outcomes are of equal interest. The geometrical interpretation of the Youden index is that its maximal values are the points on the curve which lie farthest from the chance line.

Comparison of Numerical Predictors and Classifiers. The results of numerical models cannot be compared directly to those of the classification paradigms.

One approach is the comparison of numerically predicted absorption with the classes from the original data set by means of receiver operating characteristics (ROC) and their areas under the curve (AUC).

Software Used. Molecular descriptors were generated with the open-source cheminformatics package Chemical Development Kit (CDK, version 1.2.3, 2009, http://sourceforge.net/projects/cdk).[38] For several descriptors, 3D structures had to be derived from SMILES representations by the Ghemical force field[21] (http://www.uku.fi/~thassine/projects/ghemical/). We used Weka[35] (version 3.6; http://www.cs.waikato.ac.nz/~ml/Weka/) for RF, SVM, ANN and KNN. We performed DTI with PASW Statistics version 18 for Windows (http://www.spss.com/statistics/) and linear correlation analysis with Gnu R (http://www.r-project.org/). LAZAR is available at http://www.in-silico.de/. Tanimoto coefficient calculations and grid screening for SVM meta-parameters were done with in-house software.

3. Results

3.1. Data Set. Drugs used for modeling and simulation cover a broad chemical range (Tanimoto dissimilarity: 0.702). This seems reasonable considering the data set consists of commercially available drugs and thereby exhibits certain similarities, e.g. druglike properties. The mean (±SEM) weight of molecules within the database was 346.1 (±8.3). The present data set exhibits a bimodal distribution with accumulation of compounds at 100%Abs and 0%Abs (Figure 1). This clearly reflects the two major routes of applications of common drugs (oral (high %Abs) or intravenous/topical administration (low %Abs)).

3.2. Feature Reduction. The descriptors selected by the different means of feature reduction are summarized in Table

(35) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutmann, P.; Witten, I. H. The WEKA Data mining Software: An Update. *SIGKDD Explorations* **2009**, *11*, 10–18.

(36) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861–874.

(37) Youden, W. J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35.

(38) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

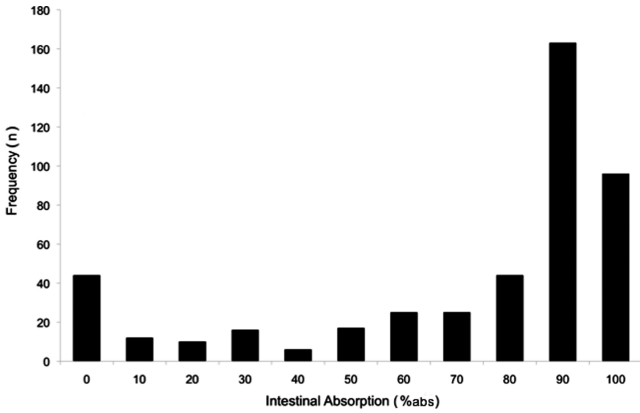

Figure 1. Histogram of measured % Abs. % Abs values retrieved from literature are given. In the present data set, well-absorbed compounds clearly outweigh badly or moderately absorbed ones.

Table 3. Feature Reduction Subsets[a]

BFS	CFS	DTIS
aLogP (LogKow aLogP2)	molar refractivity (AMR)	aLogP2 (LogKow aLogP2)
bCUTS (highest atom weighted)	bCUTS (lowest atom weighted)	H-bond donor count
H-bond donor count	H-bond acceptor count	H-bond acceptor count
H-bond acceptor count	gravitational index 4	molecular weight
gravitational index 4	gravitational index H1	longest aliphatic chain descriptor
moment of inertia descriptor (Z-axis)	gravitational index H2	tPSA
moment of inertia descriptor (XY-axis)	length over breadth descriptor (LOBMAX)	rNCS
pPSA3	pPSA2	
dPSA1	pPSA3	
dPSA3		
fNSA2		
fNSA3		
tPSA		
rHSA		

[a] In order to avoid overfitting, we applied three different selection methods to identify the most relevant physicochemical features in a complete descriptor set. The *best first algorithm* uses a greedy hill-climbing algorithm and revealed 13 features (BFS). A linear correlation of each feature with the end point was performed, and the nine descriptors with highest R^2 were used for modeling (CFS). For the final set, we used the seven splitting criteria revealed by classification and regression trees (CART), which was produced beforehand (DTIS). For a more detailed listing of all descriptors see the Supporting Information.

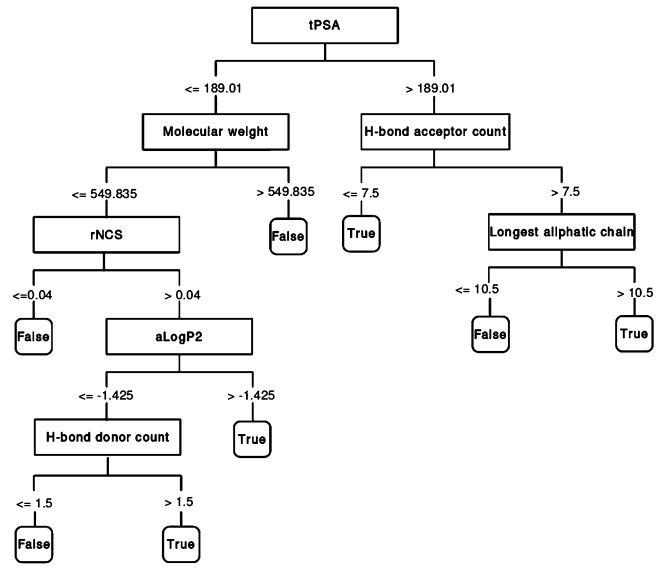

Figure 2. CART DTI. Classification of human oral absorption with CART (classification and regression tree) decision tree. The 10-fold cross-validated tree performed with corrected classification rate (CCR) of 0.84 (MCC: 0.70). The Gini coefficient was used as a homogeneity measure.

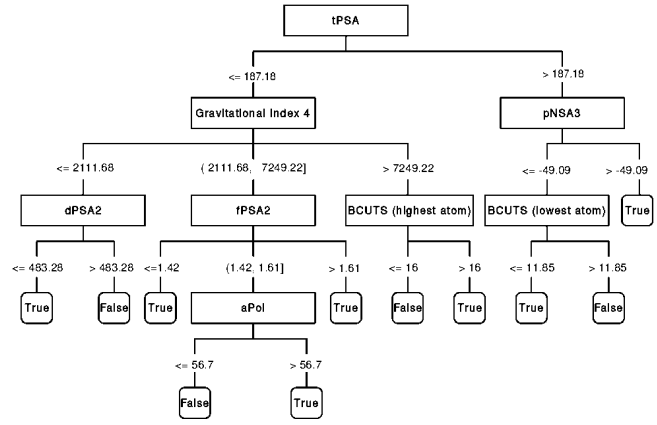

Figure 3. CHAID DTI. Decision tree with chi-squared interactor detector (CHAID). A maximum depth of five nodes and a minimum five cases in the parent and two cases in the child node were allowed for tree growth. Splitting criteria (boxes) and corresponding cutoff values are given. Leaves are depicted as rounded boxes. Predictions achieved a corrected classification rate (CCR) of 0.88 (MCC: 0.75). The whole descriptor set was used ($n = 80$) for decision tree induction.

3. All methods selected descriptors from the charged partial surface area (cPSA) subset, partition coefficients and hydrogen bonding capacity, reflecting well-known properties of drug absorption. Strikingly, measures of molecular symmetry and mass distribution (gravitational indices, moments of inertia, longest aliphatic chains) are singled out as well. To our knowledge, compound shape is not widely used in modeling these end points.

3.3. Model Performance. *3.3.1. Classification Models.* The most effective classification model in our study was built with DTI and the CHAID algorithm (CCR, 0.88; MCC, 0.75; Figure 2), followed closely by the CART algorithm (CCR, 0.84; MCC, 0.70; Figure 3). Of all methods applied to reduced feature sets, Bayesian techniques performed best (using the BFS subset, CCR, 0.81; MCC, 0.62). Other

Table 4. Performance of Classification Models[a]

method	specificity	sensitivity	CCR	MCC
Whole Feature Set				
LAZAR	0.438	0.974	0.706	0.529
CART	0.740	0.947	0.843	0.698
CHAID	0.822	0.944	**0.883**	**0.751**
random forest	0.589	0.947	0.768	0.583
BFS				
SVM rbf	0.644	0.934	0.789	0.597
SVM polynomial	0.521	0.974	0.747	0.596
multilayer perceptron	0.534	0.974	0.754	0.607
KNN	0.534	0.974	0.754	0.607
naive Bayes	0.685	0.934	0.809	0.629
Bayesian nets	0.699	0.934	**0.816**	**0.639**
CFS				
SVM rbf	0.575	0.974	0.774	0.639
SVM polynomial	0.521	0.967	0.744	0.578
multilayer perceptron	0.589	0.970	0.780	**0.641**
KNN	0.603	0.931	0.767	0.558
naive Bayes	0.521	0.931	0.726	0.492
Bayesian nets	0.616	0.957	**0.787**	0.628
DTIS				
SVM rbf	0.479	0.970	0.725	0.553
SVM polynomial	0.521	0.974	0.747	0.596
multilayer perceptron	0.589	0.964	0.776	0.623
KNN	0.589	0.947	0.768	0.583
naive Bayes	0.671	0.941	0.806	0.632
Bayesian nets	0.685	0.941	**0.813**	**0.643**

[a] We used the ordinal classes "*true*" (≥80% Abs) and "*false*" (≤ 30% Abs) for classification. Compounds of the class *"unknown"* (30% < % Abs <80%) were omitted from learning to achieve better separability for classifiers. Apart from sensitivity and specificity, highest values of corrected classification rates (CCR) and Matthews correlation coefficients (MCC) are in boldface. Results are shown for the whole data set, best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS).

paradigms such as SVM did not achieve similar performance with any of the feature sets. Classification models are summarized in Table 4.

3.3.2. Numerical Models. The multilayer perceptron yielded the strongest numerical predictions (using the CFS subset, RMSE, 25.823; R^2, 0.600). Performance is illustrated in Figure 4. SVMs with the rbf kernel achieved comparable efficacy on the DTIS subset (RMSE of 26.953; R^2, 0.590). Other methods did not perform as well (Table 5a).

3.3.3. Recoding of Numerical Predictions into Classes. As a means of salvaging predictive power from the rather mediocre numerical models, we attempted to recode the predicted %Abs values into classes based on a retrospective analysis using ROC curves (Figure 5). For each model, optimal thresholds were selected by determining the one with the highest Youden index (J). Instances were recoded into the two-class case according to these thresholds. Models and their performance after recoding are summarized in Table 5c. The SVM model with the rbf kernel was the most precise (CCR, 0.72; MCC, 0.47; using the BFS subset), outperforming the MP model.

4. Discussion

4.1. Data Set. Our data set of 458 substances covers a broad range of small molecule drugs, as indicated by the high value of dissimilarity within the descriptor space employed. The distribution of absorption ratios is bimodal with a small peak at the low end of the spectrum and a larger one for highly absorbed molecules (Figure 1). This reflects the desire to bring to market orally administrable drugs. Models based on this data set should therefore best be applied in late-stage drug development. Even though the data set is unbalanced, the sensitivity of our models is not impaired.

Wang et al.[39] proposed to use drug subsets with similar pharmacological targets when modeling human intestinal absorption. While this approach may be of use in late-stage optimization, we feel that general physiological features cannot be deduced when examining such restricted data sets.

Our models intentionally disregard mechanistic minutiae of intestinal absorption (e.g., transcellular versus paracellular pathways) as we are predicting the final end point of human intestinal absorption and not specific pathways. Members of our group have shown the validity of this approach for even more complex end points.[50] Furthermore, the original data source does not provide sufficient information on the specific absorption kinetics and metabolism on the level of the intestinal epithelium.

4.2. Feature Sets. All feature sets include descriptors of PSA. Palm et al. demonstrated its correlation with human intestinal absorption and the CACO-2 cell model.[40] In models of Winiwarter et al.,[17] polar surface area (PSA) was emphasized as one of the most important parameters to predict drug permeability. However, it was reported that an excellent sigmoidal relationship with high correlation could be established between the absorbed fraction after oral drug administration to humans and PSA.[41] In line with other groups[42,43] we clearly disapprove of this approach. As a single feature, PSA is not a reliable criterion to judge poor or good absorption. Seven descriptors of the cPSA set appear in the BFS set of features. It is important to note that while all of these concern charge analysis, they are distinctly different. Nonetheless, we performed an additional analysis and found low intercorrelation between these features (r_{avg} 0.59 ± 0.06 SEM). Highest r_{sig} (0.90) is seen for dPSA3

(39) Wang, Z.; Yan, A.; Yuan, Q.; Gasteiger, J. Explorations into modeling human oral bioavailability. *Eur. J. Med. Chem.* **2008**, *43*, 2442–2452.

(40) Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32–39.

(41) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997**, *14*, 568–571.

(42) Grass, G. M.; Sinko, P. J. Effect of diverse datasets on the predictive capability of ADME models in drug discovery. *Drug Discovery Today* **2001**, *6*, 54–61.

(43) Hou, T.; Wang, J.; Zhang, W.; Xu, X. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.

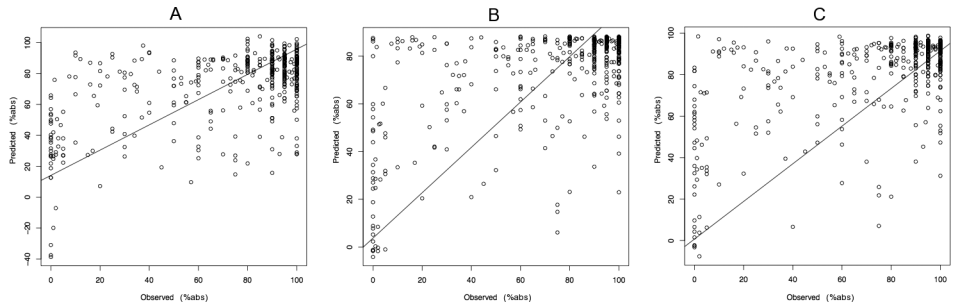

Figure 4. Scatterplot of predicted vs observed % Abs values. Best models on numeric % Abs values are given. (A) Multilayer perceptron on best first feature set (BFS). (B) Multilayer perceptron on linear correlation analysis set (CFS). (C) Support vector machines with rbf kernel on decision tree splitting criteria (DTIS).

and fNSA2, both of which are derived from central descriptors of the cPSA set. Specifically, fNSA2 puts charge into relation with molecular topology while dPSA3 weighs positive against negative charge contributions. Therefore, both contribute distinct molecular information to the models and hence have not been removed from the final data set. Additionally, any intercorrelation is penalized by cross-validation and does not introduce an overfitting bias.

In concordance with current understanding, the feature selection paradigms singled out descriptors of lipophilicity, charge (e.g., aPol), hydrogen bonding descriptors, and molecular weight (as selected in CART DTI trees). These features are already known from studies of human jejunal permeability (log P_{eff})[17,44] and deconvolution studies of human absorption rate constants.[45] Zhao et al. further demonstrated that hydrogen bonding is the rate-limiting step in absorption kinetics.[46]

Repeated inclusion of gravitational indices (CHAID, BFS, CFS), moments of inertia (BFS), length over breadth (CFS), and longest aliphatic chain (BFS, CART) indicate the importance of molecular mass distribution and geometry in modeling oral absorption. This seems reasonable as smaller molecules have better passive permeation capability than compounds with long aliphatic motifs. Moreover, BCUT descriptors[47] were selected by two paradigms (CHAID, BFS). These features are defined as eigenvalues of modified connectivity matrices with frequent application in drug discovery.[48] Their discriminatory power for aqueous solubility is well-known[49] and therefore confirms the importance of this physicochemical property in absorption kinetics.

Best-first feature selection and linear correlation analysis are two commonly used means of reducing dimensionality of the descriptor space. The use of features selected from DTI models learned from the same data is rather unusual. We consider this a valid approach in that the feature set provides a mechanistical theory which the models created *a posteriori* seek to verify. There is no unreasonable flow or leakage of information into the learning process (as in an overfitting bias) compared to reducing features using the other paradigms.

4.3. Individual Models. The DTI algorithms provided the strongest models (Table 4). Other paradigms, such as SVM classifiers, showed far weaker performance. These observations are in line with other studies.[50,51] DTI often outperforms other machine learning methods when moderately sized and skewed data sets are used. Table 6 gives a summary of recent modeling attempts. Comparable perfomance is achieved only by work based on DTI and Gaussian kernels such as Obrezanova et al.[52] who, however, fail to cross-validate their models. In terms of accuracy, our models (DTI) are only rivalled by work by Shen,[53] which, again, is not cross-validated. Remarkably, many studies choose not to employ cross-validation, resulting in accuracy measures which

(44) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
(45) Linnankoski, J.; Makela, J. M.; Ranta, V. P.; Urtti, A.; Yliperttula, M. Computational prediction of oral drug absorption based on absorption rate constants in humans. *J. Med. Chem.* **2006**, *49*, 3674–3681.
(46) Zhao, Y. H.; Abraham, M. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Beck, G.; Sherborne, B.; Cooper, I. Rate-limited steps of human oral absorption and QSAR studies. *Pharm. Res.* **2002**, *19*, 1446–1457.
(47) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.
(48) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.
(49) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.
(50) Hammann, F.; Gutmann, H.; Vogt, N.; Helma, C.; Drewe, J. Prediction of adverse drug reactions using decision tree modeling. *Clin. Pharmacol. Ther.* **2010**, *88*, 52–59.
(51) Hammann, F.; Gutmann, H.; Baumann, U.; Helma, C.; Drewe, J. Classification of cytochrome p(450) activities using machine learning methods. *Mol. Pharmaceutics* **2009**, *6*, 1920–1926.
(52) Obrezanova, O.; Segall, M. D. Gaussian processes for classification: QSAR modeling of ADMET and target activity. *J. Chem. Inf. Model.* **2010**, *50*, 1053–1061.
(53) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034–1041.

Table 5. Performance of Numeric Models, Recoded Numeric Models, and Recoded Numeric Models after ROC Threshold Optimization

(a) Performance of Numeric Models[a]

method	R^2	RMSE
	BFS	
SVM rbf	0.574	27.648
SVM polynomial	0.561	27.807
multilayer perceptron	**0.590**	**26.390**
	CFS	
SVM rbf	0.559	27.828
SVM polynomial	0.546	28.535
multilayer perceptron	**0.600**	**25.823**
	DTIS	
SVM rbf	**0.590**	26.953
SVM polynomial	0.544	28.773
multilayer perceptron	0.588	**26.099**

(b) Initial Performance of Recoded Numeric Models[b]

method	specificity	sensitivity	CCR	MCC	AUC
		BFS			
SVM rbf	0.164	0.875	**0.519**	0.590	0.786
SVM polynomial	0.137	0.835	0.486	0.546	0.786
multilayer perceptron	0.329	0.686	0.508	**0.715**	0.746
		CFS			
SVM rbf	0.205	0.818	**0.512**	0.586	0.755
SVM polynomial	0.110	0.871	0.490	0.496	0.767
multilayer perceptron	0.301	0.719	0.510	**0.655**	0.757
		DTIS			
SVM rbf	0.205	0.871	**0.538**	0.579	0.780
SVM polynomial	0.110	0.875	0.492	0.479	0.773
multilayer perceptron	0.205	0.637	0.421	**0.681**	0.774

(c) Performance of Recoded Numeric Models after ROC Threshold Optimization[c]

method	Th_{opt}	specificity	sensitivity	CCR	MCC
		BFS			
SVM rbf	82.0	0.541	0.836	0.689	0.398
SVM polynomial	79.6	0.613	0.842	0.727	**0.464**
multilayer perceptron	78.5	0.723	0.739	**0.731**	0.443
		CFS			
SVM rbf	78.9	0.574	0.842	0.708	0.430
SVM polynomial	78.0	0.526	0.901	0.714	**0.471**
multilayer perceptron	68.4	0.538	0.896	**0.717**	0.468
		DTIS			
SVM rbf	83.9	0.570	0.809	0.689	0.391
SVM polynomial	83.4	0.602	0.754	0.678	0.358
multilayer perceptron	72.3	0.619	0.840	**0.729**	**0.461**

[a] Prediction of % Abs was assessed using support vector machines with kernels and multilayer perceptron. Methods were applied on reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS). As quality measures root mean squared error (RMSE) and correlation coefficient (R^2) are given. Best results are given in boldface. [b] Performance of numeric models is shown after recoding into classification scale. We applied cutoff values from initially set ordinal classes. For performance measurement only positive class ($\geq$80% Abs) and negative class ($\leq$30% Abs) was used. Compounds classified as unknown were omitted. The corrected classification rate (CCR), Matthews correlation coefficient (MCC), specificity, sensitivity and area under the ROC curve (AUC) are indicated for each model. Best results for coefficients are given in boldface for all reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS). [c] Optimal cut points for thresholds were determined using the maximization of the Youden indices in receiver operating characteristics analysis. Specificity, sensitivity, CCR and MCC for all models are indicated at the optimal threshold (Th_{opt}). The best results are given in boldface for all reduced feature sets: best first feature selection (BFS), linear correlation analysis (CFS) and decision tree splitting criteria (DTIS).

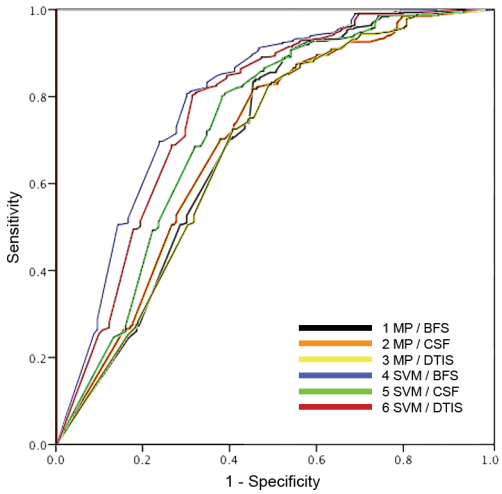

Figure 5. ROC curves of recoded numeric models. ROC curves are shown of best recoded models according to highest achieved corrected classification rates (CCR) and Matthews correlation coefficients (MCC). Multilayer perceptron on best first feature set (BFS) (1), linear correlation analysis set (CFS) (2) and decision tree splitting criteria (DTIS) (3). Support vector machines with rbf kernel on BFS (4), CFS (5) and DTIS (6).

overestimate their power in unseen data. Hou et al.[54] reported very strong models in a comparable context using SVM. Predictions achieved MCCs of up to 0.89. Their data set, however, had been artificially expanded by extrapolating %Abs values from bioavailability data. Moreover, the inclusion of redundant descriptors may have led to overfitting, as has been reported previously.[43]

It is worth noting that the PSA descriptors of CDK also contain the fragment-based method (tPSA) introduced by Ertl et al.[55] Inclusion of different PSA paradigms does not introduce redundancy. The method established by Ertl et al. considers molecule fragments, which might only be exposed to the environment when drugs are dissolved in the aqueous intestinal lumen. Other implementations focus on charge and total molecular surface area.

On reduced features, Naïve Bayes and Bayesian nets performed well. Their impressive results seem surprising at first glance, especially as these paradigms assume independence of variables. Although this is rarely given in real world data, dependence can be minimized by eliminating redundant and therefore nonindependent features. It can be argued that with the splitting of data into test and training sets, the independence bias is not equally distributed over both sets. Predictions of unseen data should then be interpreted with caution.[31] This assumption holds less well in the case of randomly performed cross-validation. The quality measures presented here might therefore give a more realistic estimate of the predictive power. Provided that the compound of interest fits the chemical space analyzed in this study, Naïve Bayes models should classify it correctly.

The numeric models presented here exhibit low predictive power as visualized in Figure 4. This might be caused by the bimodal distribution of the data set. The clustering of compounds around low and high levels of absorption reflects the two major galenic classes of drug: orally and intravenously/topically administered compounds. Hence, the instance space is not entirely covered (Figure 1). Regression models are likely to perform badly on such data. Indeed, the R^2 values achieved range from 0.544 to 0.600, confirming that linear regression models are a completely inappropriate method type for the present data set. This holds even in the case where compounds are grouped together (such that a group has similar activities), because of the linear model's constant slope. It would be more appropriate to perform binning of instances into two classes and analyze them by classification. Because focusing on two classes can improve numeric models, we stress the importance of choosing appropriate algorithms for the data set at hand.

(54) Hou, T.; Wang, J.; Li, Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408–2415.
(55) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
(56) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Butina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749–784.
(57) Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113–119.
(58) Bai, J. P.; Utis, A.; Crippen, G.; He, H. D.; Fischer, V.; Tullman, R.; Yin, H. Q.; Hsu, C. P.; Jiang, L.; Hwang, K. K. Use of classification regression tree in predicting oral absorption in humans. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2061–2069.
(59) Liu, H. X.; Hu, R. J.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 33–46.
(60) Jones, R.; Connolly, P. C.; Klamt, A.; Diedenhofen, M. Use of surface charges from DFT calculations to predict intestinal absorption. *J. Chem. Inf. Model.* **2005**, *45*, 1337–1342.
(61) Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Heyden, Y. V. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J. Pharm. Biomed. Anal.* **2005**, *39*, 91–103.
(62) Iyer, M.; Tseng, Y. J.; Senese, C. L.; Liu, J.; Hopfinger, A. J. Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol. Pharmaceutics* **2007**, *4*, 218–231.
(63) Yan, A.; Wang, Z.; Cai, Z. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. *Int. J. Mol. Sci.* **2008**, *9*, 1961–1976.
(64) Reynolds, D. P.; Lanevskij, K.; Japertas, P.; Didziapetris, R.; Petrauskas, A. Ionization-specific analysis of human intestinal absorption. *J. Pharm. Sci.* **2009**, *98*, 4039–4054.
(65) Guerra, A.; Campillo, N. E.; Paez, J. A. Neural computational prediction of oral drug absorption based on CODES 2D descriptors. *Eur. J. Med. Chem.* **2010**, *45*, 930–940.

Table 6. Summary of Recent Modeling Attempts of Human Intestinal Absorption[a]

work	*N*	CV	paradigm	R^2	RMSE	CCR	MCC	accuracy	sensitivity	specificity
Zhao et al., 2001[56]	241	no	regression	0.74	14				0.95	0.72
Niwa et al., 2003[57]	86	no	ANN (general regression)		22.8					
Niwa et al., 2003[57]	86	no	ANN (probabilistic)			0.75	0.612	80%	1	0.5
Bai et al., 2004[58]	1260	no	DTI (CART)					79−86%		
Liu et al., 2005[59]	169	yes	SVM (Gaussian kernel)	0.73	14.08				0.98	0.66
Jones et al., 2005[60]	241	no	kernel		22				0.9	0.46
Deconinck et al., 2005[61]	141	yes	DTI (CART)					65%	0.89	0
Iyer et al., 2006[62]	188	no	membrane-interaction QSAR	0.68						
Hou et al., 2007[43]	455	no	genetic programming				0.836		1	0.64
Yan et al., 2008[63]	552	no	PLS	0.83	18.18					
Yan et al., 2008[63]	552	no	SVM (rbf kernel)	0.89	16.53					
Reynolds et al., 2009[64]	567	no	nonlinear regression	0.84	35					
Obrezanova et al., 2010[52]	260	no	kernel					91%		
Obrezanova et al., 2010[52]	260	no	DTI (unspecified algorithm)					85%		
Shen et al., 2010[53]	578	no	SVM (polynomial)			0.928	0.909	98%	0.998	0.859
Shen et al., 2010[53]	578	no	SVM (rbf)			0.948	0.932	98%	0.998	0.897
Guerra et al., 2010[65]	202	yes	ANN					73%		
Suenderhauf et al., 2010	458	yes	DTI (CHAID)			0.883	0.752	92%	0.944	0.882
Suenderhauf et al., 2010	458	yes	DTI (CART)			0.843	0.698	91%	0.947	0.740
Suenderhauf et al., 2010	458	yes	ANN (numeric)	0.6	25.823					
Suenderhauf et al., 2010	458	yes	ANN (recoded)			0.717	0.468	79%	0.896	0.538

[a] Representative models for human intestinal absorption are summarized, indicating size of data set used (*n*), the use of cross-validation (CV), paradigm and algorithm used and performance estimates (coefficient of determination (R^2), root mean squared error (RMSE), corrected classification rate (CCR), Matthews correlation coefficient (MCC), accuracy, sensitivity and specificity). Accuracy was calculated as true hits (true positives and true negatives) divided by *n*.

It also should be noted that we succeeded in producing models of high accuracy (up to 92% with DTI) without specifically incorporating the influence of efflux transporters such as P-gp. This indicates that these are not major influencing factors. The poorer performance of numerical models therefore seems to be intrinsic to the paradigms.

4.4. Numeric vs Classification. Comparison of classification and numeric models is not straightforward. Numeric measures of accuracy are RMSE or R^2. In a classification system, confusion matrices and corresponding quality measures (CCR, MCC) are used. Both model types therefore cannot be compared directly. In an attempt to do so indirectly, we translated numeric predictions into a classification scale. As expected, performance of numeric models was worse when compared to genuine classification. We analyzed predictions with ROC graphs (Figure 5) to estimate predictive power for the well-absorbed class. Models achieved a reasonable sensitivity, which is reflected in high AUC values (Table 5b). In other words, numerical models tended to generally overestimate absorption ratios. By determining optimal cutoff values using the best Youden index, we markedly improved models in terms of specificity. Treated in this fashion, we state that numeric and classification models perform comparatively strong (Table 4, 5c).

5. Conclusions

While intestinal absorption of drugs in humans is mostly governed by passive diffusion, it is potentially influenced by several other factors. Our models can be seen as a combined endpoint analysis as they disregard, among other aspects, the location of absorption and specifics of administration (e.g., galenics, counterions). Performance, however, is not impeded by these generalizations. The data set comprises the entire range of absorption ratios and has a great chemical diversity in the descriptor space employed.

Although we used differing approaches to reduce features, certain descriptors were present in all sets. Descriptors of charge and lipophilicity reflect the current understanding of drug absorption in humans. Our models show the importance of molecular shape and complexity on absorption. Small size, little branching, and equal distribution of mass (as represented by descriptors of the moments of inertia) seem to be of advantage in oral absorption.

The advantages of computational methods in the prediction of oral absorption have been described previously, e.g. Norinder et al.[5] In clinical practice, drugs fall into just two categories with little overlap: those which are orally administrable and those for which other routes of administration have to be used (for example intravenous injection or topical application). Therefore, specific numerical values, such as absorption ratios, are often considered to be of less importance than classification.

Abbreviations Used

% Abs, absorption ratios; ANN, artificial neural network; AUC, area under the ROC curve; BFS, best first feature selection; CART, classification and regression tree algorithm; CCR, corrected classification rate; CFS, linear correlation feature set; CHAID, chi-squared automatic interactor detector; CYP_{450}, cytochrome P_{450}; DTIS, decision tree induction

feature set; GIT, gastrointestinal tract; KNN, *k*-nearest-neighbor; logP, partition coefficient; LOO, leave-one-out; MCC, Matthews correlation coefficient; MP, multilayer perceptrons; MRP, multidrug resistance-associated protein; (c)PSA, (charged) partial surface area; QSAR, quantitative structure−activity relationship; R^2, correlation coefficient; rbf, radial basis function; RF, random forests; RMSE, root mean squared error; ROC, receiver operating characteristics; SVM, support vector machines.

Acknowledgment. Andreas Maunz and Christoph Helma are supported by the EU FP7 Project OpenTox (Contract Number Health-F5- 2008-200787). Claudia Suenderhauf is supported by the Swiss National Foundation (Grant No. 323530-119218).

Supporting Information Available: Database of compounds used in the present work with corresponding numeric and nominal end points and SMILES codes. Observed % Abs ratios and corresponding predicted values of all numeric prediction models. This material is available free of charge via the Internet at http://pubs.acs.org.

MP100279D