Feature Selection for Descriptor Based Classification Models. 2. Human Intestinal Absorption (HIA)

Jörg K. Wegner,* Holger Fröhlich, and Andreas Zell

Zentrum für Bioinformatik Tübingen (ZBIT), Universität Tübingen, Sand 1, D-72076 Tübingen, Germany

Received October 24, 2003

We show that the topological polar surface area (TPSA) descriptor and the radial distribution function (RDF) applied to electronic and steric atom properties, like the conjugated electrotopological state (CETS), are the most relevant features/descriptors for predicting the human intestinal absorption (HIA) out of a large set of 2934 features/descriptors. A HIA data set with 196 molecules with measured HIA values and 2934 features/descriptors were calculated using JOELib and MOE. We used an adaptive boosting algorithm to solve the binary classification problem (AdaBoost.M1) and Genetic Algorithms based on Shannon Entropy Cliques (GA-SEC) variants as hybrid feature selection algorithms. The selection of relevant features was applied with respect to the generalization ability of the classification model, avoiding a high variance for unseen molecules (overfitting).

INTRODUCTION

Oral bioavailability is an important factor for designing new drugs.[1,9−19] So the feature selection problem is not only an interesting topic in machine learning[20,21] but also in QSAR for understanding molecular properties and designing new drug candidates.[8,22−27] It was already shown that the polar surface area (PSA)[1] correlates inversely with the lipid penetration ability[63,28] which shows that the objective of the relevance of the PSA is 3-fold:

First, compounds with PSA > 140 Å^2 are less than 10% absorbed by humans. A PSA of 120−140 Å^2 thus sets an upper limit for PSA in the design of oral drugs. Second, although PSA depends on the conformation, the calculations from single low-energy conformations appear to give equally good correlations.[29] Third, PSA has also a relevance for predicting the blood-brain barrier (BB) penetration.[30]

To find other relevant descriptors we applied our Genetic Algorithm based on Shannon Entropy Cliques (GA-SEC),[8] described in detail in the first part of this two-part paper. In general the performance of descriptor subsets for specific tasks, such as representative subset selection or diversity analysis, has been evaluated in a number of case studies.[23−27,31−34]

Scheme 1 shows the modified QSAR paradigm[20,35] addressing the feature selection problem using a *filter* and *wrapper approach*. It can be seen that the *wrapper approach* contains two additional loops, the optimization loop for picking an optimal descriptor set and the validation loop for the model assessment.

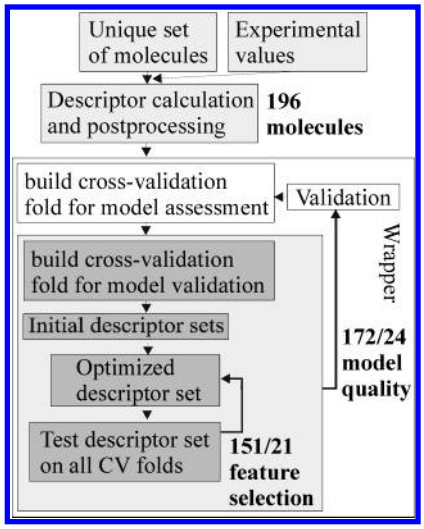
METHODS

Data Preparation. A set of 196 drugs and drug-like compounds was collected from data sets published by Wessel et al.,[22] Gohlke et al.,[36] Palm et al.,[28] Balon et al.,[65] Kansy et al.,[66] Yazdanian et al.,[67] and Yee,[68] and molecular structures were obtained from ALTANA Pharma, which could not be published. For further research we recommend the use of a huge public available collection of Human Intestinal Absorption values with molecular structures as SMILES code published by Waterbeemd, Lennernäs, and Artursson.[69]

Because the expected experimental error for the measured HIA values[28] for the data set with 196 molecules is 25% and 25% of the smallest and highest HIA values contains nearly 80% of all HIA values, we decided to reduce the regression problem to a more confidential binary classification problem.

We used a cross-validation (CV) set for a good model quality analogously to the work of Wessel et al.[22] and Gohlke

Scheme 1. Modified QSAR Paradigm[20,35] with Focus on Feature/Descriptor Selection[a]



[a] The inner loop is necessary for optimizing the feature set and the outer validation loop for assessing the model quality.

* Corresponding author phone: +49-7071-2976455; fax: +49-7071-29-5091; e-mail: wegnerj@informatik.uni-tuebingen.de.

10.1021/ci034233w CCC: $27.50
Published on Web 02/13/2004

et al.[36] All of the eight cross-validation sets for the model quality contained 151 training patterns and 21 test patterns. The feature selection wrapper was applied to these eight CV fold models. Additionally we used a second 8-fold cross-validation set for the model assessment to grant that we will not reduce the bias when optimizing the features and increase the variance for unseen molecules. All of the eight CV folds for the model assessment contain 172 training sets and (unseen) 24 test sets.

The molecules were prepared by using Corina[37] and the MOE All-Atom-Pair force field[3] for generating energy minimized structures. We checked for duplicate molecules using a hash code based on atom properties and a SMILES hash code described in detail in our theory paper.[70]

As descriptors we used the MOE descriptors[3] and the group contribution descriptors for the topological polar surface area (TPSA),[1] molar refractivity (MR), and LogP[38] under JOELib.[2]

We calculated the atom properties[2] van der Waals volume, mass, valence, electrogeometrical state index, electron affinity, electronegativity pauling, electrotopological state[7] ETS_i, graph potentials[39] GP_i, Gasteiger−Marsili partial charge[40] GM_i, intrinsic state I_i,[7] and electrogeometrical state[8] EGS_i to calculate descriptors which depend on atom properties.

We used also the atom properties conjugated environment C_i, conjugated topological distance CTD_i, conjugated electrotopological state $CETS_i$, and conjugated electrogeometrical state $CEGS_i$, which will be here described in more detail[8]

$$I_i = \frac{(2/L_i)^2 \delta_i^{\,v} + 1}{\delta_i} \tag{1}$$

$$ETS_i = I_i = \Delta I_{i,top} = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij,top} + 1)^k} \tag{2}$$

$$EGS_i = I_i + \Delta I_{i,geom} = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij,geom} + 1)^k} \tag{3}$$

$$CTD_i = \max(\{d_{ij,top} | C_i\}) \forall j \tag{4}$$

$$CETS_i = I_i + \Delta I_{i,conj} = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij,top} + 1)^{k/CTD_i}} \tag{5}$$

$$CEGS_i = I_i + \Delta I_{i,conf} = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij,geom} + 1)^{k/CTD_i}} \tag{6}$$

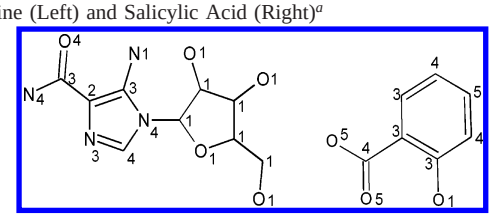
where L_i is the principal quantum number, $\delta_i^{\,v}$ is the number of valence electrons, δ_i is the number of sigma electrons of the ith atom a_i, $d_{ij,top}$ and $d_{ij,geom}$ are the topological and geometrical distances between the ith atom and the jth atom, and k is the distance influence, we used $k = 2$. Table 1 shows the SMARTS pattern[2,37] to assign the conjugation flag C_i to an atom.

The atom parameters intrinsic state I_i reflects the possible partitioning of nonsigma electrons influence along the paths starting from the considered atom i. E-state is related to electronegativity; $I_i - I_j$ is the electronegative gradient, but it is not a pure electronic descriptor. It is in fact a descriptor of atom polarity and steric accessibility.[7]

Table 1. SMARTS[41] Definitions for Assigning the Conjugated Atom Property Flag C_i[2]

SMARTS	description
a	aromatic atoms
=,#-,=*=,#*	all butadien analogues
[N,P,O,S]=,#*-[*;!H0]	α,β unsaturated (π effect)
=,#-[F,Cl,Br,I]	α,β unsaturated (σ effect)
=,#-[N,P,O,S;!H0]	α,β unsaturated (π effect, tautomer)

Scheme 2. The Conjugated Topological Distance CTD_i Descriptor for Acadesine (Left) and Salicylic Acid (Right)[a]



[a] Greater numbers represents atoms where the electronegativity gradient takes a higher effect because of a lower distance influence for the conjugated atom C_i (see also Table 1).

The conjugated topological distance CTD_i represents the number of the maximal path of conjugated atoms C_i to other connecting atoms, and Scheme 2 shows an example for acadesine and salicylic acid. So the conjugated electrotopological state $CETS_i$ reduces the distance influence for highly conjugated atoms and increments the effect of the "delocated" EN gradient.

The atom properties p_i and p_j for the atoms a_i and a_j were used to calculate the global topological charge index (GTCI)[42] and the eigenvalue descriptor: burden modified eigenvalues (BCUT).[7] Furthermore we calculated the Moreau-Broto autocorrelation $AC(d)$, which is a special case of the radial distribution (RDF) function[4−7]

$$g(r) = \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}}^{A} p_i p_j e^{-Bd^2} \tag{7}$$

$$AC(d) = g(r) \lim_{B \to \infty} \sum_{\substack{i,j \\ i \neq j}}^{A} (p_i p_j)_d \delta_{ij}$$

with

$$\delta_{ij} = \begin{cases} 1, \text{ if } d(a_i, a_j) = d \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

with $r \in \{r_{min}, r_{min} + r_{res}, \ldots, r_{min} + k r_{res} \leq r_{max}\}$, $d = r - r_{ij}$ where r_{max} and r_{min} are the minimum and maximum radius and r_{res} is the resolution used. B is an uncertainty parameter. If B is growing, the atom properties will be more located. For $B \to \infty$ the exponential term will migrate into δ_{ij} of the autocorrelation function. We can also see that the autocorrelation function is the RDF function with located atom properties where the bond lengths between the atoms shrink to a mean bond length, which can be simply expressed by the topological distance. We used $B \in \{5,25,100,200\}$, $r_{min} = 0.2$ Å, $r_{max} = 10$ Å, and $r_{res} = 0.2$ Å (or analogue $k = 50$) to calculate the RDF descriptors. Figure 1 shows the calculated RDF values for the $CETS_i$ atom property.

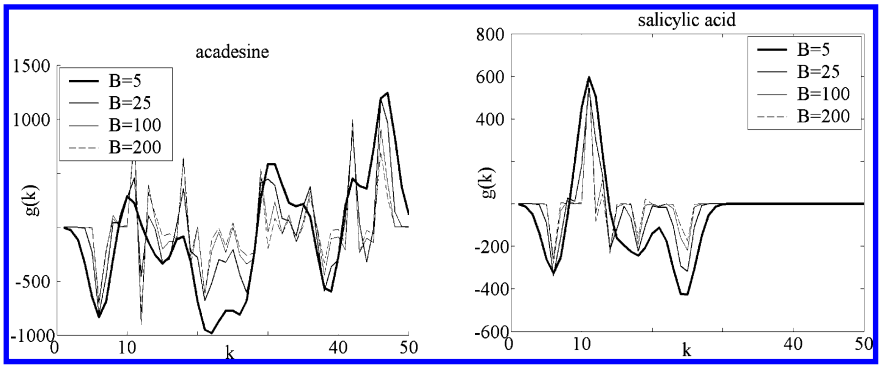


Figure 1. RDF descriptor using the $CETS_i$ atom property for acadesine and salicylic acid using different smoothing factors $B \in \{5,25,100,200\}$, $r_{min} = 0.2$ Å, $r_{max} = 10$ Å, and $r_{res} = 0.2$ Å (or analogue $k = 50$). The radius is $r = r_{min} + kr_{resolution}$. For $B = 0$ we obtain completely delocalized atom properties (straight line parallel to k axis).

Over all we calculated 3387 descriptors. Descriptors where missing values occurred were completely removed, mainly for the BCUT and autocorrelation descriptors with a great index. More sophisticated methods for handling missing values are described by Trigg.[43] Additionally all descriptors containing no information ($H_{SE} = 0$) were removed, and so we obtained 2934 descriptors, which were shown in Table 2. All descriptors were normalized by using the z-transformed descriptor distributions.[44]

Evaluation. The classification loss w_i of the cross-validation (CV) fold t can be calculated as

$$l_b(\vec{m}_j,c_j) = \begin{cases} 0 & f(m_j)c_j \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

$$w_t = \frac{\sum_{j=1}^{|M_t|} l_b(\vec{m}_j,c_j)}{|M_t|}$$

where l_b is the binary loss function, $\hat{f}(\vec{m}_j)$ is the estimated (predicted) value by the classification model, c_j is the true value, $\vec{m}_j$ is the molecule using a feature subset, and $|M_t|$ is the number of molecules in the CV fold t. The standard error of mean $\sigma_{\bar{k}}^{(k-1)}$ is calculated as[44]

$$\sigma_{\bar{k},R} = \frac{\sigma_{k,R}}{\sqrt{k}} = \frac{1}{\sqrt{k}}\sqrt{\frac{\sum_{t=1}^{k}\sum_{r=1}^{R}(c_t - \bar{c}_k)^2}{k-1}} \tag{10}$$

where $\bar{c}_k$ is the mean classification loss over all CV folds. Because we applied several GA runs R, we used

$$\sigma_{\bar{k},R} = \frac{\sigma_{k,R}}{\sqrt{k}} = \frac{1}{\sqrt{k}}\sqrt{\frac{\sum_{t=1}^{k}\sum_{r=1}^{R}(c_{t,r} - \bar{c}_{k,R})^2}{kR-1}} \tag{11}$$

where $c_{t,r}$ is the classification loss of the tth CV fold and the rth GA run and $\bar{c}_{k,R}$ is the mean classification loss over all CV folds and GA runs.

The importance of descriptors for the best models $E_{best,\vec{d}_i}^{s,e}[relevance/model]$, picked from the population of the GA between the generation interval [s, e], is calculated as

$$d_i^{best}(H_{r,g,t}^{best}) = \begin{cases} 1 & \text{if } d_i\, H_r^{best}\,(P_{GA,g}\, CV_t) \\ 0 & \text{otherwise} \end{cases}$$

$$E_{best,\vec{d}i}^{s,e}[relevance/model] = \frac{\sum_{r=1}^{R}\sum_{t=1}^{k}\sum_{g=s}^{g\le e} d_i^{best}(H_{r,g,t}^{best})}{k \cdot R \cdot (e-s)} \tag{12}$$

where $d_i^{best}(P_{GA,g},H_r, CV_t)$ is a descriptor $\vec{d}_i$ occurring in the best model $H_{r,g,t}^{best}(P_{GA,g}, CV_t)$ of the GA run r, CV fold k, and the generation g.

Table 2. Descriptors Used for Selecting Descriptor Sets and Building the Classification Models[a]

descriptor class	number
auto correlation (d_{max}=3)	52
atom and group counts	48
bond counts	12
burden values (largest 5 eigenvalues)	65
energetic descriptors	13
global topological charge	14
radial distribution function	2600
solubility descriptors	7
topological polar surface area	2
molecular refractivity	2
surface descriptors	32
partial charge surface descriptors	30
graph indices	41
weight, volume and density	16
	2934

[a] The detailed list including references can be found in the Supporting Information.

RESULTS AND DISCUSSION

Initialization. For the recursive feature elimination (RFE)[45,46] all SVM parameter combinations for $C = \{2e-6, 2e-4, ..., 2e12\}$ and $g = \{2e-8, 2e-6, ..., 2e8\}$ were built by 8-fold CV. The best results were obtained for a cost $C = 32$ and the radial basis function (RBF) parameter $\gamma = 256$.[47]

For selecting the initial D_{cut} and SE_{cut} values for the *GA-SEC* algorithm variants we calculated all cliques with D_{cut}-{0;2;0;3;...;1.0} and SE_{cut} {3;4;...;20} where the maximum number of cliques was limited to 22e6 cliques. The complete list of cliques found is available in the Supporting Information.

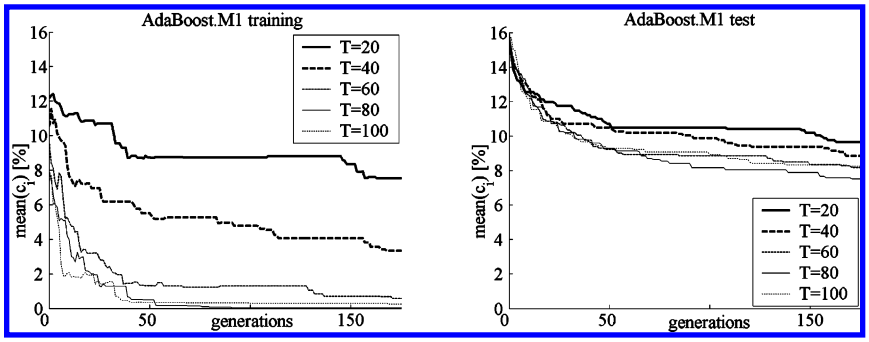


Figure 2. Training error (left) and test error (right) for AdaBoost.M1-DecisionStump using the *GA-SEC-MF* algorithm as wrapper.

We decided to use the moderate values, $D_{cut} = 0.4$, $SE_{cut} = 9$, $S_{Clique} = 6$, which results in $|G_d| = 6.1e6$ cliques that were found. In contrast to this small number the binomial coefficient for all combinations of size six out of 2934 features is approximately $8.8e17$. For the *GA-SE.MFC* variant we reduced the information content and divergence values to $D_{cut} = 0.3$ and $SE_{cut} = 3$, because the filter approaches with picking $n_{filter} = 44$ descriptors (1.5% of all descriptors) restricts the number of cliques to $|G_d| = 7198$.

We used a population size of 25 individuals. The GA used greedy overselection for picking individuals from the population, with $p_{top} = 0.9$ and $p_{get} = 0.5$, where p_{top} is the percentage of the population going into the top group and p_{get} is the likelihood that an individual will be picked from the top group. The best individual was always taken from the parent population into the child population. The one point crossover probability was $p_{cross} = 1.0$, and the mutation probability $p_{mut} = 1/|D| = 3.4e - 4$.

For all calculations an IBM *e*-series X440 server with 10GB memory and 8 Intel Xeon MP CPUs, 1.40 GHz was used running Red Hat Advanced server. The GA-SEC algorithms using the JOELib[2] library and the Weka data mining library[48] are completely written in Java and use SUN's JDK1.4.1_01-b01.

Number of Experts. It was shown that the generalization error ϵ_g for boosting, with high probability, is at most[49,50]

$$\epsilon_g(H) = \hat{\epsilon}(H) + \bar{O}\left(\sqrt{\frac{Td}{m}}\right) \qquad (13)$$

where $\hat{\epsilon}(H) = \hat{\Pr}[H(x) \neq y]$ is the empirical training error, d is the *VC dimension* (capacity) of the hypothesis space,[51] T is the number of combined experts, and m is the number of independent random examples. Although there exists an empirical upper bound formulation without containing T we applied a test series to find a suitable number of experts to use.[50]

Figure 2 shows the training and the test error using the *GA-SEC-MF* algorithm wrapping a AdaBoost.M1 classifier using a decision stump. A decision stump can be regarded as a one level decision tree, which tests all attributes and is very effective on two class problems.[48]

We can see that 80 experts lead to the lowest training and test error on the model cross-validation experiment using 8-folds on the test data set (inner loop in Scheme 1).

In contrast to this very optimistic result we see in Figure 3 that this great number of experts leads to a high variance on the CV model assessment experiment, using also 8-folds on the validation data set (model assessment loop; outer loop in Scheme 1). Taking the run time also into account, we decided to use only 20 experts.

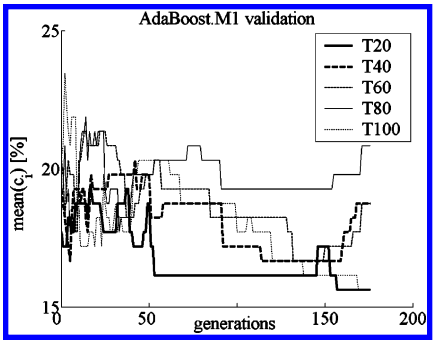


Figure 3. Validation error for AdaBoost.M1-DecisionStump using the *GA-SEC-MF* algorithm as wrapper.

Table 3. Initialization Parameters for the GA-SEC Algorithm Used in the HIA Experiments[8]

$P_{GA}(0)$ initialization method (filter)	description
GA-SEC	$D_{cut} = 0.4$; $SE_{cut} = 9$; $S_{Clique} = 6$
GA-SEC-MF	$D_{cut} = 0.4$; $SE_{cut} = 9$; $S_{Clique} = 6$; $p_{Filter} = 0.005$; (best 15 descriptors); n_{Filter} (pick two of the best descriptors of every filter approach)
GA-SEC−PMF	$D_{cut} = 0.3$; $SE_{cut} = 3$; $S_{Clique} = 6$; $n_{Filter} = 44$ (1.5% of the best descriptors)
GA-MF	p_{Filter} (best 15 descriptors); n_{Filter} (pick two of the best descriptors of every filter approach)

Table 4. Classification Loss $\bar{c}_k$ and Standard Error $\sigma_k^{(k-1)}$ of Mean Using 8-Fold CV

	classification model	no. of features	$\bar{c}_k\ \sigma_k^{(k-1)}$
A	no feature selection AdaBoost.M1-DS (T=20)	2934	20.41 ± 2.97
B	no feature selection SVM (RBF, γ=256, C=32)	2934	17.85 ± 3.10
C	RFE (RBF, γ=256, C=32)	20	32.10 ± 6.58
D	RFE (RBF, γ=256, C=32)	50	23.42 ± 4.68
E	RFE (RBF, γ=256, C=32)	100	20.92 ± 3.61

GA-SEC Variants. Our hybrid feature selection algorithm can be divided into two groups. One, in which we ignore the class information (*GA-SEC*), which is suitable for classification and clustering problems. And the other group, taking the class information into account (*GA-SEC-MF*, *GA-SEC-PMF*, and *GA-MF*), which is suitable for classification problems.

Table 4 shows the results for the AdaBoost.M1-DecisionStump algorithm[48−50] and the support vector machine (SVM) algorithm using a radial basis function (RBF) kernel[47] without using any feature selection algorithm. We see that the SVM

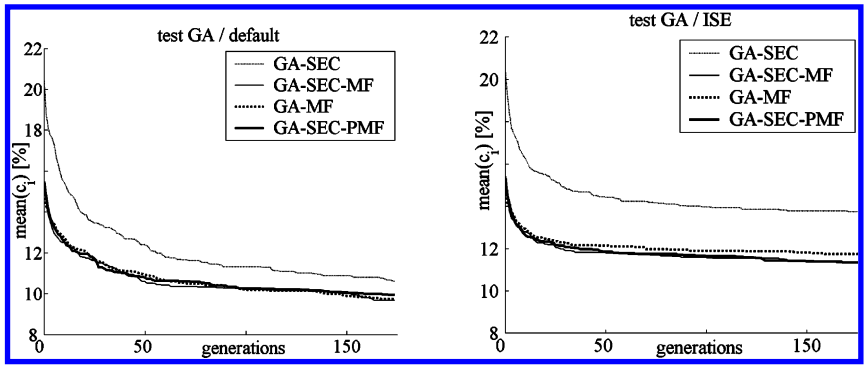


Figure 4. Classification loss $\bar{c}_{k,R}$ for $R = 3$ GA experiments for the 8-fold CV test set (inner model validation loop in Scheme 1).

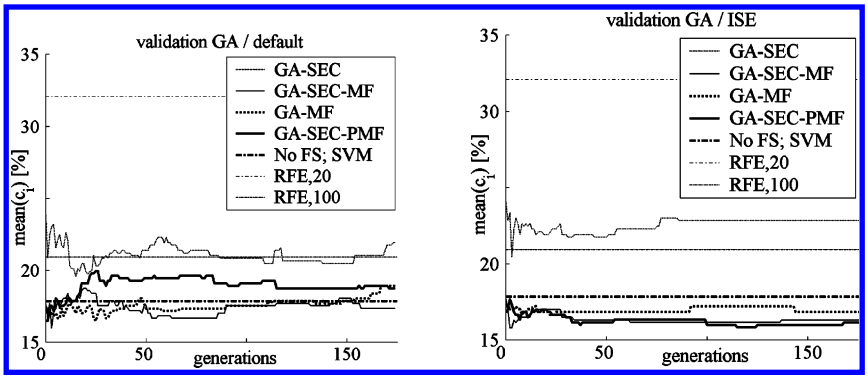


Figure 5. Classification loss $\sigma_{k,R}^{(k-1)}$ for $R = 3$ GA experiments for the 8-fold CV validation set (outer model assessment loop in Scheme 1).

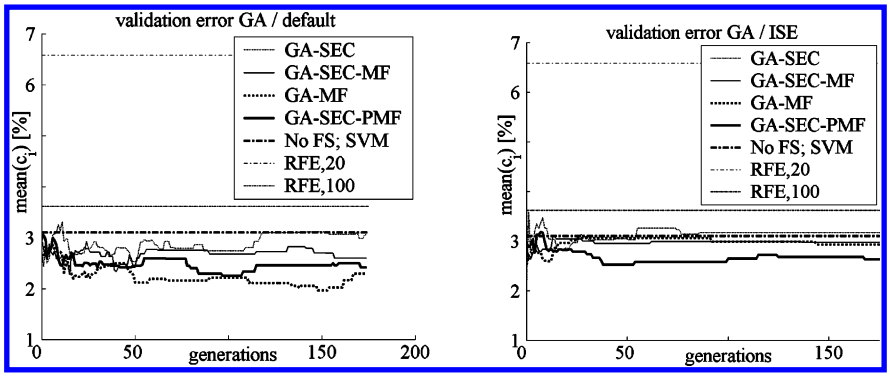


Figure 6. Standard error of mean $\sigma_{k,R}^{(k-1)}$ for $R = 3$ GA experiments for the 8-fold CV validation set (outer model assessment loop in Scheme 1).

for this data set was able to calculate a better model using all features, than AdaBoost.M1. Additionally Table 3 shows the results for the deterministic recursive feature elimination (RFE) algorithm.[45,46] The results for using 20, 50, and 100 features are all inferior to the results without using any feature selection algorithm. In contrast to our hybrid feature selection algorithms the number of features to be selected by the RFE must be defined.

The plain *GA-SEC* algorithm using only diverse descriptor sets without any problem relevant class information leads to the highest number of wrongly classified molecules/instances on the 8-fold CV test set (Figure 4).

The GA wrapper using the *SE.MFC* initialization and the ISE mutation operator leads to the models with the smallest number of wrongly classified molecules on the 8-fold CV validation set (Figure 5). We can also see that the GA using no class information (*GA-SEC*) is inferior to all GA variants on the 8-fold CV validation set and obtains already comparable results to the RFE with 100 selected features.

The *GA-SEC-MF* algorithm leads furthermore to a small standard error of mean $\sigma_{k,R}^{(k-1)}$ on the 8 CV validation set (Figure 6). The algorithms *GA-SEC-MF*/*default*, *GA-SEC-PMF*/*ISE*, and *GA-MF*/*ISE* were able to find better results than all RFE feature selection variants and are even better than using a SVM without using feature selection. The *GA-SEC-PMF/ISE* algorithm leads to the best result with a classification loss $\bar{c}_{k,R}$ of 15.83 ± 2.73.

Table 5 contains the classification loss $\bar{c}_{k,R}$ and the standard error of mean $\sigma_{k,R}^{(k-1)}$ for $R = 3$ GA experiments, for the bests results found, for the generations 3, 120, and 175. The number of descriptors used for this generation can be found in Table 6.

Relevant Descriptors. We analyzed the descriptors used for the best models found by the GA for the *GA-SEC-PMF*/*ISE* algorithm in more detail. When analyzing all descriptors over all generations (Table 7) and all descriptors from the models with the lowest classification loss in the generation

Table 5. Classification Loss $\bar{c}_{k,R}$ and Standard Error of Mean $\sigma_{k,R}^{(k-1)}$ for $R = 3$ GA Experiments[a]

	$g = 3$	$g = 120$	$g = 174$
initialization/mutation	$\bar{c}_{k,R} \pm \sigma_{k,R}^{(k-1)}$	$\bar{c}_{k,R} \pm \sigma_{k,R}^{(k-1)}$	$\bar{c}_{k,R} \pm \sigma_{k,R}^{(k-1)}$
GA-SEC/default	22.22 ± 2.71	21.94 ± 3.27	22.50 ± 3.19
GA-SEC-MF/default	**15.97 ± 2.56**	**17.71 ± 2.72**	**17.36 ± 2.59**
GA-SEC-PMF/default	**17.54 ± 2.64**	18.75 ± 2.46	18.75 ± 2.42
GA-MF/default	**17.01 ± 2.78**	17.88 ± 2.10	18.90 ± 2.30
GA-SEC/ISE	21.62 ± 2.96	23.70 ± 3.30	23.70 ± 3.30
GA-SEC-MF/ISE	**16.67 ± 2.61**	**17.71 ± 2.32**	18.23 ± 2.83
GA-SEC-PMF/ISE	**17.33 ± 3.03**	**15.83 ± 2.72**	**16.17 ± 2.63**
GA-MF/ISE	**17.21 ± 2.79**	**17.21 ± 2.99**	**16.85 ± 2.93**

[a] Results in bold letters are better than all RFE results using feature selection and plain SVM results.

Table 6. Mean Number of Descriptors $\bar{d}^*$ of the Best Feature Sets (Models) and the Standard Deviation $\sigma^{(d^*-1)}$ for $R = 3$ GA Experiments[a]

	$g = 3$	$g = 120$	$g = 174$
initialization/mutation	$\bar{d}^* \pm \sigma^{(d^*-1)}$	$\bar{d}^* \pm \sigma^{(d^*-1)}$	$\bar{d}^* \pm \sigma^{(d^*-1)}$
GA-SEC/default	7.73 ± 1.67	33.00 ± 8.94	37.47 ± 12.50
GA-SEC-MF/default	**8.16 ± 2.75**	**26.38 ± 7.97**	**30.46 ± 8.35**
GA-SEC-PMF/default	**8.30 ± 2.57**	24.42 ± 8.95	28.04 ± 10.37
GA-MF/default	**5.54 ± 1.79**	25.29 ± 9.01	28.13 ± 10.39
GA-SEC/ISE	10.69±1.62	34.06 ± 8.90	36.18 ± 11.24
GA-SEC-MF/ISE	**11.00 ± 2.67**	**34.13 ± 9.75**	36.13 ± 10.83
GA-SEC-PMF/ISE	**10.12 ± 2.64**	**28.36 ± 8.02**	**31.44 ± 9.60**
GA-MF/ISE	**7.30 ± 2.20**	**23.26 ± 7.51**	**26.69 ± 7.99**

[a] Results in bold letters are better than all RFE results using feature selection and SVM results using no feature selection.

Table 7. 10 Most Relevant Descriptors for the *GA-SEC-PMF* Initialization Using the ISE Mutation Operator for $s = 0$ and $e = 175$[a]

$E_{best,\vec{d}_i}^{s,e}$ [relevance/ model]	descriptor $\vec{d}_i$
86.7	TPSA (JOELib)
67.1	TPSA (MOE)
53.7	PEOE_VSA_POL
53.7	RDF_B5.0:Conjugated_electrotopological_state_index:0
32.5	Auto_correlation:Electrogeometrical_state_index:1
32.1	RDF_B100.0:Gasteiger_Marsili:1
27.8	RDF_B200.0:Atom_valence:0
27.3	RDF_B100.0:Electrotopological_state_index:1
26.8	RDF_B100.0:Electron_affinity:2
25.0	RDF_B25.0:Electron_affinity:2

[a] The mean value for descriptors per model is 24.90 with a standard deviation of 8.81. At all 345 different descriptors were used for the best models.

interval [115,125] (Table 8), we see that the ranking of the descriptors does differ not for the four most relevant descriptors: *TPSA (JOELib)*, *TPSA (MOE)*, *PEOE_VSA_POL*, *RDF_B5.0:Conjugated_electrotopological_state_index:0*. The complete lists of the most relevant descriptors for the intervals [0,175] and [115,125] can be found in the Supporting Information.

The high relevance of the membrane permeation rate on the polar surface area (TPSA) and the total polar van der Waals surface area (PEOE_VSA_POL) is consistent with the literature.[9,28,29,52] The TPSA[1] in MOE[3] is slightly inferior to the TPSA descriptor in JOELib,[2] which can be explained by the missing contribution patterns for aliphatic sulfur, aromatic sulfur, sulfone, and phosphorane groups (Table 9).

Table 8. 10 Most Relevant Descriptors for the *GA-SEC-PMF* Initialization Using the ISE Mutation Operator for $s = 115$ and $e = 125$[a]

$E_{best,\vec{d}_i}^{s,e}$ [relevance/ model]	descriptor $\vec{d}_i$
87.5	TPSA (JOELib)
66.7	TPSA (MOE)
54.2	PEOE_VSA_POL
54.2	RDF_B5.0:Conjugated_electrotopological_state_index:0
37.5	RDF_B100.0:Gasteiger_Marsili:1
33.0	Auto_correlation:Electrogeometrical_state_index:1
33.0	RDF_B100.0:Electron_affinity:2
33.0	RDF_B200.0:Atom_valence:0
30.0	RDF_B100.0:Electrotopological_state_index:1
29.0	RDF_B200.0:Conjugated_topological_distance:1

[a] The mean value for descriptors per model is 27.75 with a standard deviation of 7.64. At all 245 different descriptors where used for the best models

Table 9. Group Contribution Patterns Missing in the Topological Polar Surface Area (TPSA)[1] in MOE[3]

entry	description	SMARTS	contribution
33	aliph. sulfur	[S](-*)-*	25.30
36	arom. sulfur	[s](:*):*	19.21
38	sulfone	[S](-*)(-*)(=*)=*	8.38
42	phosphorane	[P](-*)(-*)(-*)=*	9.81

When using descriptors two aspects should be mentioned. First, the same data processing workflow should be used to calculate descriptors, because different implementations can use different expert systems for assigning the hybridization, aromaticity, implicite hydrogens, and finally the atom type. Second, the descriptor calculation algorithm applied after the expert systems should use always the same algorithm or here the same SMARTS based group contribution patterns.

Additionally atom property based descriptors were often selected for the best models to improve the performance of these models. It must be mentioned that nothing can be said about the relevance of these descriptors alone (which corresponds to plain *feature selection filter approaches*).[70] It can be only said that all following descriptors help to improve the performance of the models (using descriptor sets) and the often selected TPSA descriptor, which occurs in ∼87% of all of the ***best*** models built. As already mentioned the ***best*** model is the model picked out of a population size of 25 individuals (models) and was repeated three times (R=3), with different random generator seeds. So we can interpret the relevance of descriptor sets only statistically not deterministically, because it is not possible to calculate the full combinatorial descriptor subsets which can occur (*combinatorial optimization problem*).[70]

It is not recommended to use a subjective and unmotivated smaller set of descriptors without any objective ranking method. This will always lead to the subjective expected result, that the final model uses only this kind of descriptor. But the not used descriptors cannot be inserted afterward into the already obtained ranking of subsets by a *feature selection wrapper approach*. In contrast to that, *feature selection filter approaches*, which depends only on a single descriptor, can always be combined. Because there already exists a lot of filter approaches, the question is to use which one. We used six filter approaches to initialize our GA-SEC algorithm.

Table 10. 10 Best Correlation Rankings Found by Karlén et al.[11] To Predict the HIA Using a PLS Ranking[53−59]

R^2	descriptor $\vec{d}_i$
0.76	PSA
0.75	HBD
0.70	number of HBD and HBA
0.52	LogD6.5
0.47	LogD4.7
0.42	LogP
0.41	LogD5.5
0.41	HBA
0.37	dipole moment (DM)
0.36	ClogP

The importance of RDF values with a radius $r = r_{min} \approx 0$ for the $CETS_i$ atom property can be interpreted as a global $CETS_i - CETS_j$ interaction function, molecular distribution function $MDF = RDF(r \approx 0)$, for a molecule; which decreases exponentially fast depending on the smoothing parameter B and their atom−atom-distance r_{ij}. When interpreting formulas (1), (5), and (7) this is a kind of a globally "delocalized" EN gradient or global polarization descriptor.

Additionally we have found that the RDF descriptors are within the 100 most important descriptors, using all kinds of atom properties and smoothing parameters, but only with a radius $r \leq 1$ Å, or more exactly $k \leq 4$ with $r_{min} = 0.2$ Å, $r_{res} = 0.2$ Å, and $r = r_{min} + kr_{res}$. The complete lists of all relevant descriptors can be found in the Supporting Information.

DISCUSSION

The severe comparison to other feature ranking results with chemical data sets is difficult, because often the data sets are not publicly available, the used data processing workflow can differ, and the applied algorithms and all of their parameters are not always available. The recently published NIPS benchmark data sets[64] are helpful for pure machine learning algorithms but not for the QSAR area. In our companion paper we have published two huge QSAR feature selection benchmark data sets and the actual open source library JOELib[2] contains all mentioned atom properties and descriptor calculation algorithms. So the focus of this discussion lies on the topic of the number of descriptors and fingerprints used and a sloppy generalization ability discussion.

Karlén et al.[11] used a partial-least-squares (PLS)[53−59] analysis to rank descriptors to predict the HIA (Table 10). They used also only a set of 14 theoretical descriptors. Bajorath et al.[23,24] used a PCA-GA algorithm to identify the most relevant descriptors to distinguish between seven different targets (including diverse sets of enzyme inhibitors, receptor agonists, and antagonists) (Table 11). The ranking was calculated by interpreting the results of the Jarvis-Patrick (JP) clustering algorithm

$$S_1 = \frac{C_p}{10C_m + C_s} \tag{14}$$

where C_s are singletons, C_p are pure classes, and C_m are mixed classes penalized by a factor of 10.

Brown and Martin[31] presented an extensive comparison for four different data sets using seven different descriptor species and six different clustering methods, where Ward's clustering method outperforms the Jarvis-Patrick clustering. The ranking found is shown in Table 12.

Hence Karlén et al.[11] and Brown/Martin[31] used a *filter approach* to find relevant descriptors, and Bajorath et al.[23,24] used a GA *wrapper approach*.[20] Because all of these experiments did not use a separate validation data set or cross-validation, the generalization ability of these experiments must be seen critically. Additionally, we believe that it must be seriously distinguished between descriptor sets of similarly small size and all possible descriptors of one species (Table 12). Especially to avoid the overrepresentation of chemical similarity and getting inferior similarities, e.g. when the comparison based on 2D fingerprints is applied to the whole molecule and not to the substituents alone.[60] So the good performance of the MACCS key in contrast to the PPP pairs can be also interpreted by the smaller number of features used, especially if no combinatorial optimization algorithm is used to validate these descriptor (sub-) sets (outer model assessment loop in Scheme 1).[8,20,23−27,61,62] Furthermore it must be seriously distinguished between feature rankings using only 10 or 20 features and feature rankings showing a subset out of, e.g. 2934 features.

CONCLUSIONS

It was shown that our new hybrid feature selection algorithms are useful methods, which were able to find better results than boosting without feature selection and the recursive feature elimination (RFE) algorithm. Furthermore we did not need to define the number of features to be selected, and we were able to calculate the expectation values for the relevance of descriptors for the best models found. Hence we have a confidence measure with respect to the generalization ability of the classification model and with

Table 11. Preferred Descriptor Sets Found by Bajorath et al.[23,24] To Distinguish between Seven Target Types Using a PCA-GA Algorithm

preferred descriptor set[23]	no. of features	JP ranking: S_1
A_aro,f_c=o,LP,vsa_pol	4	1.78
A_nc,RPC-,KierA3,MACCS (166)	169	1.47
ChiO_C,chi1,a_nCl,PC+,PEOE_VSA+2,apol,MACCS (166)	172	1.44
A_aro,PEOE_VSA+4,PEOE_VSA_PNEG,PEOE_VSA_POL,vsa_pol	5	1.44
A_aro,b_double,PEOE_VSA-6, PEOE_VSA_PNEG,I_so2nh2,vsa_acc	6	1.44

preferred descriptor set[24]	no. of features	JP ranking: S_1
B_ar, SS (57),HB-a	59	1.01
B_1rotR,$^1\chi$,PEOE_PC+,SS (57),$^1\kappa$, $^2\kappa$, $^3\kappa$	63	0.96
B_1rotR,b_ar, SS (57), HB-a	60	0.95
$^1\chi$,PEOE_PC+,SS (57),HB-a, CMR	61	0.86
B_ar, PEOE_PC+, SS (57), $^2\kappa$, $^3\kappa$	61	0.84

Table 12. Brown's and Martin's[31] Ranking for Four Different Data Sets Using Ward's Clustering Method

descriptor set	no. of features	ranking
MACCS	153	1
Unity 2D	988	2
Unity 3D	336	3
PPP pairs	2048	4

respect to the used descriptors, because we used two separate CV folds (Scheme 1). Additionally we were able to compare descriptors, e.g. TPSA[1] of MOE[3] and JOELib,[2] which can be useful to find differences and improve them. Finally, we have shown that our conjugated electrotopological state (CETS)[8] descriptor is beside the topological polar surface area (TPSA)[1] a relevant descriptor for predicting the human intestinal absorption (HIA). Furthermore, the molecular distribution function $MDF = RDF(r \approx 0)$ seems to be a highly relevant descriptor for predicting the HIA. The *GA-SEC-PMF/ISE* algorithm leads to the best result with a classification loss of 15.83 ± 2.73.

We conclude that the feature selection problem[21] is not only an interesting topic for the machine learning community but also valuable for the QSAR community.

ACKNOWLEDGMENT

We thank ALTANA Pharma AG, Konstanz, Germany for financial support and providing us with the HIA data set. Stephen Jelfs, University of Sheffield, ChemoInformatics Group of Prof. Dr. Willett/Dr. Gillet for implementing the group contribution (GC) algorithm for molar refractivity (MR),[38] polar surface area (PSA),[1] and LogP[38] for JOELib.[2]

Note Added after ASAP Posting. This article was released ASAP on February 13, 2004 with equation 7 being incomplete. The correct version was posted on March 4, 2004.

Supporting Information Available: The complete lists of descriptors used, of the cliques obtained for $D_{cut} \in \{0.2;0.3;...1.0\}$ and $SE_{cut} \in \{3;4;...20\}$ and the most relevant descriptors for the intervals [0,175] and [115,125] of the *GA-SEC-PMF* algorithm. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.
(2) JOELib, http://joelib.sourceforge.net/.
(3) MOE (Molecular Operating Environment), Chemical Computing Group Inc., 2003.
(4) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vibrat. Spectrosc.* **1999**, *19*, 151−164.
(5) Hemmer, M. C.; Gasteiger, J. Prediction of Three-Dimensional Molecular Structures Using Information from Infrared Spectra. *Anal. Chim. Acta* **2000**, *420*, 145−154.
(6) Gasteiger, J. A Hierachy of Structure Representations. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1034−1061, ISBN 3-527-30680-3.
(7) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: D-69469 Weinheim, Germany, 2000.
(8) Wegner, J.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(9) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(10) Mandagere, A. K.; Thompson, T. N.; Hwang, K.-K. Graphical Model for Estimating Oral Bioavailability of Drugs in Humans and Other Species from Their Caco-2 Permeability and in Vitro Liver Enzyme Metabolic Stability Rates. *J. Med. Chem.* **2002**, *45*, 304−311.
(11) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernäs, H.; Karlén, 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.
(12) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Copper, 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.
(13) 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.
(14) Raevsky, O. A.; Schaper, K.-J.; Artursson, P.; McFarland, J. W. A Novel Approach for Prediction of Intestinal Absorption of Drugs in Humans based on Hydrogen Bond Descriptors and Structural Similarity. *Quantum. Struct.-Act. Relat.* **2002**, *20*, 402−413.
(15) Deretey, E.; Feher, M.; Schmidt, J. M. Rapid Prediction of Human Intestinal Absorption (HIA). *Quantum. Struct.-Act. Relat.* **2002**, *21*, 493−506.
(16) Zmuidinavicius, D.; DidziaPetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification Structure- -Activity Relations (C−SAR) in Prediction of Human Intestinal Absorption. *J. Pharm. Sci.* **2003**, *92*, 621−633.
(17) 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.
(18) Yoshida, F.; Topliss, G. QSAR Model for Drug Human Oral Bioavailability. *J. Med. Chem.* **2000**, *43*, 2575−2585.
(19) Mannhold, R. Octanol/Water Partition Coefficients. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1300−1313, ISBN 3-527-30680-3.
(20) Kohavi, R. Wrappers for Performance Enhancement and Oblivious Decision Graphs, *Dissertation*, Stanford university, 1995.
(21) Davies, S.; Russell, S. Np-completeness of searches for smallest possible feature sets, *Proceedings of the 1994 AAAI Fall Symposium on Relevance.* AAAI Press: New Orleans, 1994, pp 37−39.
(22) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(23) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 669−704.
(24) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(25) Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. Genetic Algorithm Applied to the Selection of Factors in Principal Component-Artificial Neural Networks: Application to QSAR Study of Calcium Channel Antagonist Activity of 1,4-Dihydropyridines (Nifedipine Analogous). *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328−1334.
(26) Baumann, K.; Albert, H.; Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part I. search algorithm, theory and simulations. *J. Chemom.* **2002**, *16*, 339−350.
(27) Baumann, K.; Korff, M.; Albert, H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part ii. practical applications. *J. Chemom.* **2002**, *16*, 351−360.
(28) Palm, K.; Stenburg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharma. Res.* **1997**, *14*, 568−571.
(29) Clark, D. E. Rapid calculation of polar surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.
(30) Clark, D. E. Rapid calculation of polar surface area and its application to the prediction of transport phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.
(31) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(32) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(33) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(34) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(35) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: D-69469 Weinheim, Germany, 2000; ISBN 3-52-29913-0.
(36) Gohlke, H.; Dullweber, F.; Kamm, W.; März, J.; Kissel, T.; Klebe, G. Prediction of Human Intestinal Absorption using a combined 'Simulated Annealing/Back-propagation Neural Network' Approach. *Rational Approaches Drug Des.* **2001**, 261−270.
(37) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D-Space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030−1037.
(38) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(39) Walters, W. P.; Yalkowsky, S. H. ESCHER-A Computer Program for the Determination of External Rotational Symmetry Numbers from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1015−1017.
(40) Gasteiger, J.; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, 3181−3184.
(41) Bush, B. L.; Sheridan, R. P. PATTY: A Programmable Atom Typer and Language for Automatic Classification of Atoms in Molecular Databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.
(42) Gálvez, J.; Garcia-Domenech, R.; Julián-Ortiz, V. D.; Soler, R. Topological Approach to Analgesia. *J. Chem. Inf. Comput. Sci.* **1994**, *14*, 1198−1203.
(43) Trigg, L. Designing Similarity Functions, *Dissertation*, University of Waikato, New Zealand, 1997.
(44) Altman, D. G. *Practical statistics for medical research*; Chapmann & Hall/CRC, New York, U.S.A., 1991; ISBN 0-412-27630-5.
(45) Weston, J.; Elisseeff, A.; Schölkopf, B.; Tipping, M. Use of the Zero-Norm with Linear Models an Weston, J.; Elisseeff, A.; Schölkopf, B.; Tipping, M. Use of the Zero-Norm with Linear Models and Kernel Methods'. *J. Machine Learning Res. (*special issue on Variable and Feature Selection) **2003**, *3*, 1439−1461.
(46) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389−422.
(47) Schöllkopf, B. Support Vector Learning, *Dissertation*, University of Berlin, Oldenbourg Verlag: Germany, 1997.
(48) Witten, I. H.; Eibe, F. Data Mining: *Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann: 1999; ISBN 1-55860-552-5.
(49) Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, 1995; pp 23−37.
(50) Freund, Y.; Schapire, R. A short introduction to boosting. *J. Jpn. Soc. Artif. Intel.* **1999**, *14*, 771−780.
(51) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, U.S.A., 1995.
(52) Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and Computational Screening Models for the Prediction of Intestinal Drug Absorption. *J. Med. Chem.* **2001**, *44*, 1927−1937.
(53) Cho, S. J.; Cummins, D.; Bentley, J.; Andrews, C. W.; Tropsha, A. An Alternative to 3D QSAR: Application of Genetic Algorithms and Partial Least Squares to Variable Selection of Topological Indices. Submitted for publication in *J. Comput.-Aided Mol. Des.*
(54) Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. Multivariate Structure−Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−137.
(55) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(56) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(57) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure−Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k-Nearest-Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(58) Stanton, D. T. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, 1423−1433.
(59) Eriksson, L.; Antti, H.; Holmes, E.; Johansson, E.; Lundstedt, T.; Shockcor, J.; Wold, S. Partial Least Squares (PLS) in Cheminformatics. *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1134−1166, ISBN 3-527-30680-3.
(60) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Molecular Diversity* **1999**, *4*, 1−22.
(61) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(62) Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. 2D QSAR Modeling and Preliminary Database Searching for Dopamine Transporter Inhibitors Using Genetic Algorithm Variable Selection of Molconn Z Descriptors. *J. Med. Chem.* **2000**, *43*, 4151−4159.
(63) Artursson, P.; Bergström, C. A. S. Intestinal Absorption: the Role of Polar Surface Area. In *Drug bioavailability*; Waterbeemd, H., Lennernäs, H., Artursson, P., Eds.; Wiley-VCH: Weinheim, Germany, 2003; pp 341−357, ISBN 3-527-30438-X.
(64) Neural Information Processing Systems Conference (NIPS) − Feature Selection Challenge, 2003, http://www.nipsfsc.ecs.soton.ac.uk/.
(65) Balon, K.; Riebesehl, B. U.; Müller, B. W. Drug Liposome Partitioning as a Tool for the Prediction of Human Passive Intestinal Absorption. *Pharm. Res.* **1999**, *16*, 882−888.
(66) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J. Med. Chem.* **1998**, *41*, 1007−1010.
(67) Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating Partitioning and Caco-2 Cell Permeability of Structurally Diverse Small Molecular Weight Compounds. *Pharm. Res.* **1998**, *15*, 1490−1494.
(68) Yee, S. In Vitro Permeability Across Caco-2 Cells (Colonic) Can Predict In Vivo (Small Intestinal) Absorption in Man-Fact or Myth. *Pharm. Res.* **1997**, *14*, 763−766.
(69) *Drug bioavailability*; Waterbeemd, H., Lennernäs, H., Artursson, P., Eds.; Wiley-VCH: Weinheim, Germany, 2003; pp 341−357, ISBN 3-527-30438-X.
(70) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature selection for Descriptor based Classification Models: Part I-Theory and GA-SEC Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.

CI034233W