# Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method

Jörg K. Wegner* and Andreas Zell

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

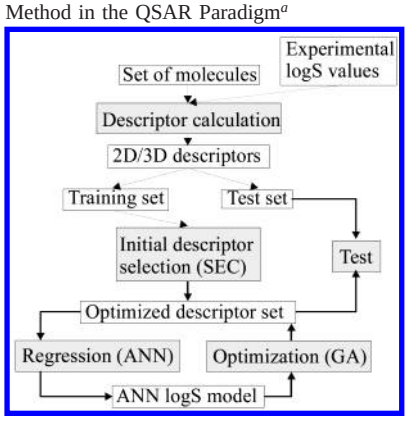
Received January 17, 2003

The paper describes a fast and flexible descriptor selection method using a genetic algorithm variant (GA-SEC). The relevance of the descriptors will be measured using Shannon entropy (SE) and differential Shannon entropy (DSE), which have very sparse memory requirements and allow the processing of huge data sets. A small quantity of the most important descriptors will be used automatically to build a value prediction model. The most important descriptors are not a linear combination of other descriptors, but transparent, pure descriptors. We used an artificial neural network (ANN) model to predict the aqueous solubility *logS* and the octanol/water partition coefficient *logP*. The *logS* data set was divided into a training set of 1016 compounds and a test set of 253 compounds. A correlation coefficient of 0.93 and an empirical standard deviation of 0.54 were achieved. The *logP* data set was divided into a training set of 1853 compounds and a test set of 138 compounds. A correlation coefficient of 0.92 and an empirical standard deviation of 0.44 were achieved.

## INTRODUCTION

Feature (descriptor) selection is an important task for quantitative structure activity relationship (QSAR) methods,[1−5] similarity/diversity analysis,[1,7] and library design.[8,9] The performance of descriptor sets for specific tasks, such as representative subset selection or diversity analysis, has been evaluated in a number of case studies.[3,4,6,10−13] It has been shown that the feature selection problem is *NP-complete*.[14] The common feature selection and feature extraction methods work with genetic algorithms (GA),[1,2] principal component analysis (PCA),[3] or with hybrid methods such as PCA-GA.[4] The disadvantages of PCA are the memory requirements and the missing transparency of the calculated eigenvectors. Input variables are a linear combination of all the original input features, which poses a transparency problem for optimizing specific descriptors. Other methods such as the CROatian Multiregression selection of descriptors (CROMsel)[5] can handle only a predefined number of descriptors to select. In the process of generating a logS[15] model there are several issues to take into account. First, there is the task to select relevant descriptors and second, the calculation of the regression model, e.g. an artificial neural network (ANN) model.[16] If the selected descriptors of the first step are not perfect, it is necessary to repeat these two steps until the model cannot be optimized any further. The selected descriptor set and the ANN model of the logS/logP value will be the fitness function for our optimization model. For the optimization a modified genetic algorithm (GA) is applied, which uses the Shannon entropy cliques (SEC) to generate an initial population, which speeds up the evolutionary optimization process. The developed inverse Shannon entropy (ISE) mutation operator yields a higher fitness value than the standard mutation operator. Scheme 1 shows the SEC algorithm, the NN regression, and the GA in the context of the QSAR paradigm.[17]

**Scheme 1.** SEC Descriptor Selection, ANN Regression, and GA Optimization Method in the QSAR Paradigm[a]




[a] For a selected descriptor set (SEC), a regression model (ANN) is calculated. The results are optimized (GA) until the optimum is reached.

## DATA SETS

For the analysis of the descriptor selection method we worked with the Huuskonen data set[18−23] with logS[15] results, where logS is the solubility at a temperature of 20−25 °C in mol/L. The data set was converted from SMILES[24,25] flat file representation to a structured data file (SDF)[26] using JOELib[27] (a Babel/OELib[28] successor). Seventeen of the 1033 molecules from the training set and five of the 258 molecules from the test set could not be converted successfully. The training set and the test set contained 1016 and 253 molecules, respectively. The smaller validation set with 21 molecules was converted completely. The partitioning of the data set is analogue to the work of Liu and So. [22]

Additionally we used the original logP data set of Wang[29,30] to show that the GA-SEC model building method

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

10.1021/ci034006u CCC: $25.00 
Published on Web 05/02/2003

is independent of the data set. The data set contains 1853 molecules in the training set and 138 molecules in the test set.

## METHODS

**Structure Representation.** The 3D coordinates were calculated by using Corina[31] and afterward the MOE all-atom-pair AMBER98 force field. The descriptors were calculated with MOE[32] and JOELib.[27] The implemented global topological charge index,[33] the Moreau-Broto autocorrelation, and the Burden-modified eigenvalues in JOELib used the following atom properties: Gasteiger-Marsili partial charges,[34] electronegativity (Pauling), graph potentials,[35] atom mass, van der Waals volume, electron affinity, and atom valence. Additionally the XlogP[29,30] descriptor was calculated for every molecule by using a XlogP-JOELib processing module. Altogether 199 descriptors were used for the Huuskonen data set experiments of the series1 and 230 descriptors for the series2.

For the data set of Wang we used additionally MolConnZ descriptors and the atom parameters intrinsic state *I*, electrotopological state *ETS*, electrogeometrical state *EGS*, conjugated topological distance *CTD*, and conjugated electrotopological state *CETS*. The atom parameters were used to calculate the autocorrelation and the Burden-modified eigenvalues with JOELib

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

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

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

$$CTD_i = \max\left(\{d_{ij,top} | a_i \text{ is conjugated}\}\right) \forall\, j$$

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

where $L_i$ is the principal quantum number, $\delta_i^{\,v}$ is the number of valence electrons, and $\delta_i$ is the number of sigma electrons of the $i$th atom $a_i$. $d_{ij,top}$ and $d_{ij,geom}$ are the topological and geometrical distances between the $i$th atom and $j$th atom. $k$ is the distance influence; we used $k = 2$. Altogether 332 descriptors were used in the Wang data set (experiments series3). The complete descriptor names lists and statistics for all series are available in the Supporting Information.

All descriptor values $d_i$ were normalized by using

$$\hat{d}_{i,m} = \frac{d_{i,m} - \overline{d_i}}{\sigma_i}$$

where $i$ is the index number of the descriptor and $m$ is the index number of the molecule, $\overline{d_i}$ is the average $d_{i,m}$ over all molecules, and $\sigma_i$ is the standard deviation of descriptor $d_i$.

**Theoretical Basis.** Shannon entropy (SE)[36] provides a connection between entropy and information content and was originally applied in digital communication technology to determine the amount of data that could be transmitted given a range of frequencies.[37] Shannon entropy is defined as

$$SE(i) = -\sum_{k=1}^{B} p_{i,k} \log_2 p_{i,k}$$

where $p_{i,k}$ is the probability of a data point or "count" $c_{i,k}$ to adopt a value within a specific data interval $k$ with $B$ bins. Here we chose $B = 20$. Thus, $p_{i,k}$ is calculated as

$$p_{i,k} = c_{i,k} / \sum_{n=1}^{N} c_{i,k,n}$$

where $N$ is the number of molecules. $SE(i)$ contains a logarithm to the base 2, which corresponds to a scale factor and permits the resulting $SE(i)$ to be considered as the number of bits necessary to capture the information contained within the descriptor variation. In this fashion, SE values for different data sets can be directly compared, provided a uniform binning scheme can be defined. As discussed in previous papers, SE values alone may not be sufficient to select descriptors with significant discriminatory power.[36] Furthermore this method is characterized by a strong tendency to oversample remote areas of the feature space and to produce unbalanced designs.[38]

A method termed differential Shannon entropy (DSE)[39] can be used to improve the significant discriminatory power and compare differences in information content and variance of molecular descriptors. In this paper the advantages of SE and DSE are combined. For the diversity calculation the differential Shannon entropy (DSE) is used,[39,40] which is defined as

$$DSE(i,j) = SE_{ij} - (SE_i + SE_j)/2$$

$SE_{ij}$ refers to the Shannon entropy calculated from a single histogram reflecting the distribution of two descriptor values over a complete molecular data set. $SE_i$ and $SE_j$ represent the SE values of all descriptor values when considered individually. The underlying idea of DSE calculations is that combinations of descriptor distributions in histograms with consistent binning schemes are not the sum of single distributions. Combining distributions requires renormalization of the data over a constant number of intervals and thus generates a new envelope for data representation. *DSE(i,j)* is the difference between the renormalized histogram of both distributions and the average of their independent histograms. If the envelope of the combined distributions shows increased spread or variability, a positive *DSE(i,j)* is observed. In contrast, if the envelope shows no increased variability, a negative *DSE(i,j)* value may be observed. Thus, even subtle differences in distributions and their value ranges can be quantified. The memory requirements[41] are $O(D \cdot B)$ for the descriptor binning and $O(D^2)$ for the selection method respectively, where $D$ is the number of all available descriptors. Since the memory requirements of both methods are independent of the number of molecules, they are suitable for analyzing huge data sets. For computing all *SE(i)* and *DSE(i,j)* values only two iteration steps over the complete data set are needed. One for getting the descriptor statistics

with minima/maxima of the descriptors and another for calculating the *SE(i)* and *DSE(i,j)* values.

Choosing $d$ maximally diverse descriptors out of a descriptor set of size $D$ requires the evaluation of

$$\binom{D}{d} = \frac{D!}{d!(D-d)!}$$

subsets.[41] To reduce the number of evaluation steps a matrix $M_{adj}$ is defined. $M_{SE}(i,j)$ contains the quadratic information content, and $DSE(i,j)$ is the diversity information for a descriptor pair. Both matrices are combined to $M_{adj}(i,j)$ in the following way

$M_{adj}(i,j)$ is 1 if $M_{SE}(i,j) > SE_{cut}$ and $DSE(i,j) > DSE_{cut}$,

otherwise $M_{adj}(i,j)$ is 0

where $SE_{cut}$ is the minimally allowed quadratic information content value and $DSE_{cut}$ is the minimally allowed differential Shannon entropy value. The $M_{adj}(i,j)$ matrix can be used to find descriptors with a high information content which will correlate little with other descriptors by using a maximum clique detection algorithm.[43,44] A maximum complete subgraph (clique) is a complete subgraph that is not contained in any other complete subgraph.[44] Complete means that every node of the clique is connected to every other node of the clique. In our case a clique will be the descriptor subset which has a high information content and where every descriptor is maximally diverse to any other descriptor in this clique. A subset-selection graph, $G_d$, is created from $M_{adj}(i,j)$ by using the Bron-Kerbosch (BK)[44] clique detection algorithm. We call the $G_d$ subsets of the $M_{adj}(i,j)$ matrix the Shannon entropy cliques (SEC). The run time of clique detection algorithms depends strongly on the edge/node density in graphs, and an overview was recently published.[45]

Every set bit in the GA genome represents a descriptor used. The fitness function of the GA takes the correlation coefficient of the test set $R^2_{ANN,Test}$ and the number of used descriptors into account. The penalty term (Figure 1) decrements the correlation coefficient (raw fitness), if the number of descriptors used increases.

The fitness function is defined as

$$\overline{TP} = \frac{1}{N}\sum_{n=1}^{N} T_n \sum_{n=1}^{N} P_n$$

$$\overline{TT} = \frac{1}{N}\sum_{n=1}^{N} T_n \sum_{n=1}^{N} T_n$$

$$\overline{PP} = \frac{1}{N}\sum_{n=1}^{N} P_n \sum_{n=1}^{N} P_n$$

$$R = \frac{\sum_{n=1}^{N} T_n P_n - \overline{TP}}{\sqrt{\left(\sum_{n=1}^{N} T_n T_n - \overline{TT}\right)\cdot\left(\sum_{n=1}^{N} P_n P_n - \overline{PP}\right)}}$$

$$f_{raw} = R^2_{ANN,Test}$$

$$f = f_{raw} - k f_{dec}(d,D) = f_{raw} - k\frac{e^{d/D} - 1}{(e-1)}$$

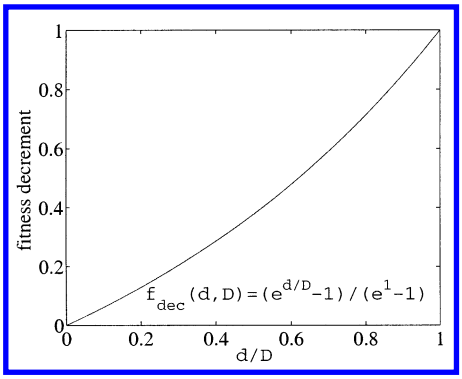



**Figure 1.** Fitness decrement of the $R^2_{ANN,Test}$ to penalize the increasing number of descriptors used.

where $T_n$ are the $N$ true values of the experimental logS/logP values and $P_n$ are the $N$ predicted values of the regression, $f$ is the fitness, $d$ is the number of descriptors used, and $D$ is the number of all available descriptors. $f_{dec}$ is a penalizing term for the increasing number of descriptors used, $k$ is the penalty relevance factor, we here set $k = 1$. $R$ is the correlation coefficient.

We used a population size of 70 individuals for series1 and 200 individuals for series2/3. 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 two point (series1) and one point (series2/3) crossover probability was $p_{cross} = 0.9$, the mutation probability $p_{mut} = 0.1$ for series1 and $p_{mut} = 0.0239$ for series2. For the mutation we tested two methods. First, the standard mutation operator of the genetic algorithm. Second, the inverse Shannon entropy (ISE) mutation operator. In the ISE muation operator the mutation probability of descriptor values in one GA individual is proportional to the Inverse Shannon Entropy values, which is defined as ISE(i) = 1/SE(i).

For the neural network regression analysis presented in this paper the Java interface JavaNNS[46] for the SNNS kernel[47] was used. We worked with a simple fully connected net with one hidden layer. The neurons used a logistic activation function. The fast and stable resilient back-propagation (Rprop)[16] learning function was the training algorithm.

The number of hidden neurons was set to 15 and for the number of training cycles a constant value of 1500 was chosen. This number is based on two arguments. First, Figure 2 shows that a small number of training cycles allows more generation evaluations of the genetic algorithm and leads to results with a smaller standard deviation. Second, Figure 3 shows that a value of 1500 training cycles leads to the best regression results for the series3 data set. Figure 3 shows the $R^2_{ANN,Test}$ values of the GA-SEC individuals with the highest fitness value, not the individuals with the highest $R^2_{ANN,Test}$ values. This explains the very small peak after 10 h (Figure 3) for the evaluation of 2000 training cycles, although we used an elitistic GA strategy.

## RESULTS AND DISCUSSION

**Initialization.** For the initial population of 70 and 200 cliques, respectively, the Shannon Entropy Cliques (SEC)

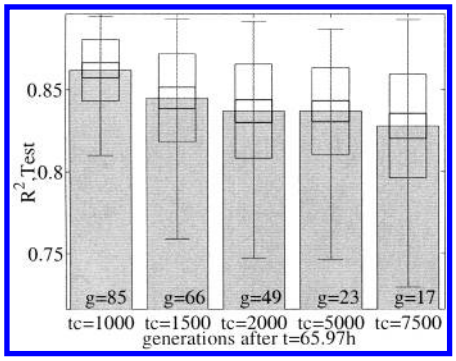



**Figure 2.** The outer lines show the minima and maxima of the $R^2_{ANN,Test}$ values of the evaluated population after $g$ generations (65.97 h). The outer bar shows the variance and the inner bar shows the 95% confidence interval for the Wang data set (series3). *tc* is the number of training cycles.

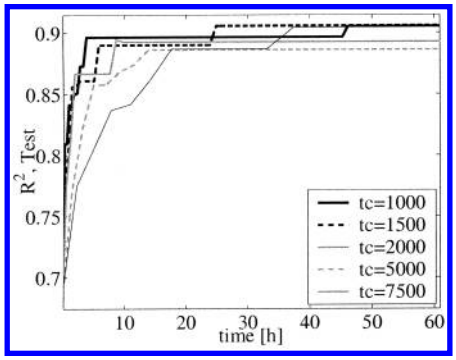



**Figure 3.** The evaluation of the Wang data set (series3) with *tc* training cycles shows that 1500 training cycles leads to the best $R^2_{ANN,Test}$ values.

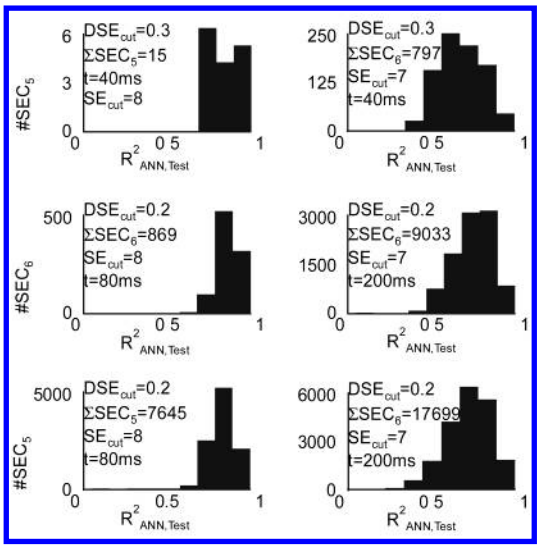



**Figure 4.** The SEC descriptor sets with $SE_{cut} = 8$ and $DSE_{cut} = 0.2$ already give a good correlation coefficient $R^2_{ANN,Test}$ for the series1 data set, which represents individuals near local maxima.

were taken randomly from all available SECs. This chapter shows that the cliques already contain relevant descriptor sets which can be used for the initialization of the population for the GA. Figure 2 shows the calculated correlation coefficient of the Huuskonen test set $R^2_{ANN,Test}$ (series1). For evaluating the clique based descriptor selection method the SECs of the size 5 and 6 were used. Only the diagrams with the $SE_{cut}$ value of 7 and 8 and the $DSE_{cut}$ of 0.2 and 0.3 are shown. The values $SE_{cut} = 7$ and $DSE_{cut} = 0.3$ are very

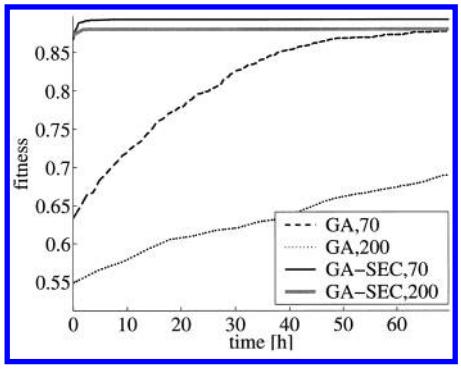



**Figure 5.** Results of different descriptor selection algorithms. The GA-SEC algorithm is faster than a standard GA and gives results with higher fitness values for the Huuskonen data sets of the series1 (GA,70 and GA-SEC,70) and series2 (GA,200 and GA-SEC,200).

restrictive and lead only to 15 cliques of size 5. Cliques of size 6 were not found. We chose $SE_{cut} = 8$ and $DSE_{cut} = 0.2$ for initializing the start population of the GA based on two arguments: First, there were 869 cliques of size 6 found, and no clique must be selected twice to initialize a population with 70 or 200 individuals. Second, the cliques of size 6 already have high $R^2_{ANN,Test}$ values, which represents individuals near local maxima. For the initialization, only the cliques of size 6 were used. The calculation of the cliques needed only 80 ms. For the initialization, 70 or 200 cliques were randomly chosen to produce a start population for the GA. Because the calculation of the $R^2_{ANN,Test}$ values for all cliques is very expensive ($\overline{t_{ANN}} = 15.12$ s), the overview of the SECs was only calculated for the series1 data set (Figure 4). For all calculations an IBM 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 and the JOELib[25] software are completely written in Java and uses SUN's JDK1.4.1_01-b01.

**Optimization.** Figure 5 shows the results of the mean fitness values over eight experiments of a standard GA descriptor selection algorithm compared to the GA-SEC algorithm. The cross-correlation tables for the final descriptors used for the best logP and logS models are available in the Supporting Information.

Also a GA which uses SEC for generating a starting population (GA-SEC) and the GA-SEC algorithm using the Inverse Shannon Entropy (ISE) mutation operator are visualized (Figure 6). The GA-SEC with ISE finds better fitness values than the standard GA-SEC algorithm.

The GA-SEC optimization algorithm with the ISE mutation operator also finds higher fitness values than a standard GA-SEC algorithm for the Wang data set (Figure 7). Although we know that other logP descriptors/models helps us to build consensus analogue logP models,[48,49] we tried also to build our model without any other logP descriptors (ISE* and SEC* in Figure 7). The results are slightly inferior compared to our best model which uses the MOE[32] *logP-(o/w)* descriptor.

The differences between two mean values are considered as significant if their 95% confidence intervals do not intersect. The 95% confidence interval shows that the true mean value of the results lies with 95% probability in this interval under the assumption that the values are normally distributed. Figure 8 shows that the 95% confidence intervals

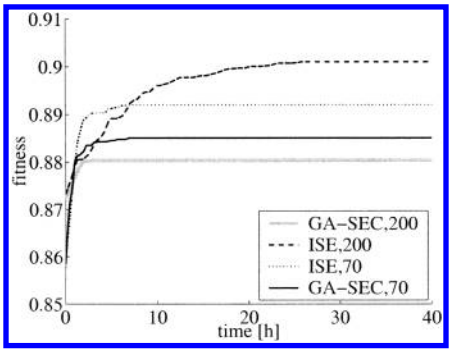



**Figure 6.** Results of comparing GA-SEC with and without the ISE mutation operator. The GA-SEC algorithms with the ISE mutation operator gives higher fitness values for the Huuskonen data sets of the series1 (GA-SEC,70 and ISE,70) and series2 (GA-SEC,200 and ISE,200).

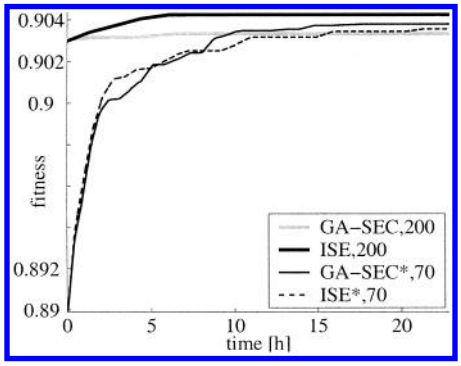



**Figure 7.** Results of comparing GA-SEC with and without the ISE mutation operator for the Wang data set (series3).

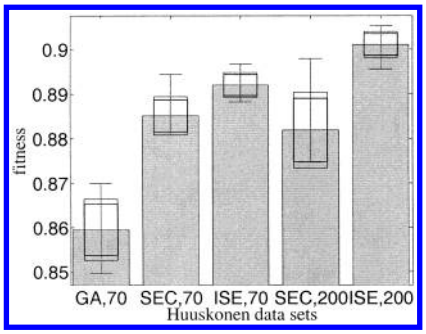



**Figure 8.** The outer lines show the minima and maxima of the highest fitness values of each experiment over eight experiments, the outer bar shows the variance, and the inner bar shows the 95% confidence interval for the Huuskonen data set (series1/2).

of the GA, GA-SEC, and GA-SEC with ISE algorithm do not intersect for the eight experiments of series1 (GA,70; SEC,70, and ISE,70) and series2 (SEC,200 and ISE,200). The GA-SEC with ISE of series2 (ISE,200) reaches the largest mean value. The standard deviation of the GA-SEC with ISE method is the smallest in comparison to both other methods. The best GA individual found has a correlation coefficient of $R^2_{ANN,Test} = 0.93$. It represents a neural net with nine input neurons (nine selected descriptors). The fitness value is 0.905.

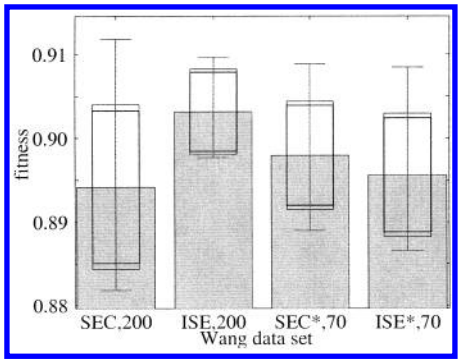



**Figure 9.** The outer lines show the minima and maxima of the highest fitness values of each experiment over eight experiments, the outer bar shows the variance, and the inner bar shows the 95% confidence interval for the Wang data set (series3).

Figure 9 shows that the standard deviation for the results of the GA-SEC algorithm with ISE mutation operator is smaller for the Wang data set (series3). And although the Wang data set (series3) contains 102 more descriptors than the Huuskonen data set (series2), the best ANN regression result was found with an analogous time need. The best GA individual found has a correlation coefficient of $R^2_{ANN,Test} = 0.92$. It represents a neural net with seven input neurons. The fitness value is 0.9097. For the best GA individual without using other logP values we obtained $R^2_{ANN,Test} = 0.90$ which represents a neural net with 31 input neurons and a fitness value is 0.842. In fact, the GA-SEC ANN model for logP yielded acceptable estimations for all 19 compounds[23,29,30,50−53] (Table 4) except *verapamil*, which is a well-known problem also for the models of Wang,[29,30] Rekker,[52] and Suzuki-Kudo.[53] When this compound is removed from the current list, the standard deviation shows that our logP ANN model achieves the second best result among all these methods.

## CONCLUSIONS

It was shown that the GA-SEC methods are fast and stable optimization methods for selecting descriptors. The moderate memory requirements allow the processing of large molecular data sets. Furthermore the GA-SEC feature selection algorithms are a transparent selection method with interpretable

**Table 1.** Comparison of the Prediction Power of the Models with Other Published Models Based on Huuskonen's Data Set[a]

| | | | training set | | | test set | | | validation set | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| model | type | *d* | *n* | $R^2$ | *s* | *n* | $R^2$ | *s* | *n* | $R^2$ | *s* |
| our model | ANN9-15-1 | 9 | 1016 | 0.94 | 0.52 | 253 | 0.93 | 0.54 | 21 | 0.82 | 0.79 |
| Gasteiger[21] | MLR (SPSS) | 40 | 797 | 0.79 | 0.93 | 496 | 0.82 | 0.79 | 21 | 0.56 | 1.20 |
| | ANN40-8-1 | 40 | 797 | 0.93 | 0.50 | 496 | 0.92 | 0.59 | 21 | 0.85 | 0.77 |
| Liu[22] | ANN7-2-1 | 7 | 1033 | 0.86 | 0.70 | 258 | 0.86 | 0.71 | 21 | 0.79 | 0.93 |
| Tetko[20] | MLR (SPSS) | 33 | 879 | 0.86 | 0.75 | 412 | 0.85 | 0.81 | 21 | 0.77 | 0.99 |
| | ANN33-4-1 | 33 | 879 | 0.94 | 0.47 | 412 | 0.91 | 0.60 | 21 | 0.90 | 0.64 |
| Huuskonen[18] | MLR (SPSS) | 30 | 884 | 0.89 | 0.67 | 413 | 0.88 | 0.71 | 21 | 0.83 | 0.88 |
| | ANN30-12-1 | 30 | 884 | 0.94 | 0.47 | 413 | 0.88 | 0.60 | 21 | 0.91 | 0.63 |

[a] $R^2$ is the quadratic correlation coefficient and *s* the empirical standard deviation.

**Table 2.** Predicted and Experimental Aqueous Solubility for 21 Compounds of the Validation Set[18−23] [a]

| no. | CAS no. | name | $\log S_{exp}$ | $\log S_{pred}$ | Gasteiger[21] | Liu[22] | Tetko[20] |
|---|---|---|---|---|---|---|---|
| 1 | 37680-73-2 | *2,2′,4,5,5′-PCB* | −7.89 | −7.66 | −7.85 | −7.55 | −7.57 |
| 2 | 94-09-7 | *benzocaine* | −2.32 | −2.05 | −2.19 | −1.45 | −1.63 |
| 3 | 50-78-2 | *aspirin* | −1.72 | −1.81 | −1.87 | −2.1 | −1.81 |
| 4 | 58-55-9 | *theophylline* | −1.39 | −1.21 | −1.27 | −0.73 | −0.69 |
| 5 | 60-80-0 | *antipyrine* | −0.56 | −1.74 | −1.31 | −1.41 | −0.89 |
| 6 | 1912-24-9 | *atrazine* | −3.85 | −2.82 | −3.83 | −1.51 | −3.70 |
| 7 | 50-06-6 | *phenobarbital* | −2.34 | −2.36 | −2.80 | −2.5 | −2.89 |
| 8 | 330-54-1 | *diuron* | −3.80 | −3.31 | −3.70 | −2.85 | −3.01 |
| 9 | 67-20-9 | *nitrofurantoin* | −3.47 | −2.82 | −2.52 | −2.89 | −3.09 |
| 10 | 57-41-0 | *phenytoin* | −3.99 | −2.90 | −3.18 | −3.09 | −3.52 |
| 11 | 439-14-5 | *azepam* | −3.76 | −4.14 | −4.81 | −4.08 | −4.37 |
| 12 | 58-22-0 | *testosterone* | −4.09 | −4.27 | −4.52 | −4.49 | −4.13 |
| 13 | 58-89-9 | *lindane* | −4.64 | −3.98 | −5.04 | −4.91 | −4.91 |
| 14 | 56-38-2 | *parathion* | −4.66 | −4.06 | −3.66 | −3.64 | −4.31 |
| 15 | 333-41-5 | *diazinon* | −3.64 | −4.18 | −2.66 | −3.56 | −3.43 |
| 16 | 77-09-8 | *phenolphthalein* | −2.90 | −4.64 | −4.62 | −4.16 | −4.31 |
| 17 | 121-75-5 | *malathion* | −3.37 | −2.96 | −2.79 | −2.52 | −3.73 |
| 18 | 2921-88-2 | *chlorpyriphos* | −5.49 | −6.41 | −4.79 | −4.5 | −5.31 |
| 19 | 363-24-6 | *prostaglandin_E2* | −2.47 | −3.98 | −3.07 | −3.8 | −3.52 |
| 20 | 50-29-3 | *p,p′-DDT* | −8.08 | −6.85 | −7.86 | −7.93 | −7.59 |
| 21 | 57-74-9 | *chlordane* | −6.86 | −6.47 | −7.66 | −7.32 | −7.23 |
| | | emp. SD *s* | | 0.79 | 0.77 | 0.93 | 0.64 |

[a] The values were predicted using our ANN9-15-1 model.

**Table 3.** Comparison of the Prediction Power of the Models We Present Here with Other Published Models Based on Wang's Data Set[a]

| | | | training set | | | test set | | |
|---|---|---|---|---|---|---|---|---|
| model | type | *d* | *n* | $R^2$ | *s* | *n* | $R^2$ | *s* |
| our model | ANN7-15-1 | 7 | 1853 | 0.96 | 0.41 | 138 | 0.92 | 0.44 |
| our model (*) | ANN31-15-1 | 31 | 1853 | 0.95 | 0.33 | 138 | 0.90 | 0.47 |
| Wang[29,30] | atom contrib | 90 | 1853 | 0.95 | 0.35 | 138 | 0.94 | 0.33 |

[a] $R^2$ is the quadratic correlation coefficient and *s* the empirical standard deviation. The ANN31-15-1 model contains no other logP values, which explains a higher number of descriptors to use to obtain analogous results.

descriptor values in contrast to the principal components obtained by the PCA feature extraction method. Using Shannon Entropy (SE) values allows the comparison of descriptors with different data ranges, and the number of selected descriptors is not restricted. The problem of the correlation between descriptors[41] can be addressed by using the presented Shannon Entropy Cliques (SEC). Neural networks with different feature sets and different net topologies may be optimized simultaneously by the genetic algorithm. The GA-SEC algorithm with ISE finds the highest value $R^2_{ANN,Test} = 0.93$ and leads to values with the lowest variance and the smallest 95% confidence interval in this series of GA-ANN-hybrid algorithms. Further experiments plan to use committees of models to allow descriptors with a high information content, where only a few descriptor values are missing (e.g. burden values for small molecules) using JOELib.[27] The GA-SEC method was able to select automatically nine relevant descriptors for the Huuskonen and seven descriptors for the Wang data set.

**Table 4.** Predicted and Experimental Partition Coefficients for 19 Selected Compounds[23,29,30,50−53] [a]

| no. | name | $\log P_{exp}$ | $\log P_{pred,r92}$ | $\log P_{pred,r90}$ | Wang[29,30] | Moriguchi[50,51] | Rekker[52] |
|---|---|---|---|---|---|---|---|
| 1 | *atropin* | 1.83 | 2.37 | 2.29 | 2.29 | 2.21 | 1.88 |
| 2 | *chloramphenicol* | 1.14 | 1.32 | 0.66 | 1.46 | 1.23 | 0.32 |
| 3 | *chlorothiazide* | −0.24 | −0.05 | −0.25 | −0.58 | −0.36 | −0.68 |
| 4 | *chlorpromazine* | 5.19 | 5.19 | 5.18 | 4.91 | 3.77 | 5.10 |
| 5 | *cimetidine* | 0.40 | 0.40 | 1.83 | 0.20 | 0.82 | 0.63 |
| 6 | *diazepam* | 2.99 | 2.76 | 2.87 | 2.98 | 3.36 | 1.88 |
| 7 | *diltiazem* | 2.99 | 3.14 | 2.28 | 3.14 | 2.67 | 4.53 |
| 8 | *diphenhydramine* | 3.27 | 3.69 | 3.18 | 3.74 | 3.26 | 3.41 |
| 9 | *flufenamic acid* | 5.25 | 4.03 | 4.81 | 4.45 | 3.86 | 5.81 |
| 10 | *haloperidol* | 4.30 | 4.17 | 3.95 | 4.35 | 4.01 | 3.57 |
| 11 | *imipramine* | 4.80 | 4.19 | 4.24 | 4.26 | 3.88 | 4.43 |
| 12 | *lidocaine* | 2.26 | 2.13 | 2.38 | 2.47 | 2.52 | 2.30 |
| 13 | *phenobarbital* | 1.47 | 1.61 | 2.02 | 1.77 | 0.78 | 1.23 |
| 14 | *phenytoin* | 2.47 | 1.76 | 2.58 | 2.23 | 1.80 | 2.76 |
| 15 | *procainamide* | 0.88 | 1.02 | 1.45 | 1.27 | 1.72 | 1.11 |
| 16 | *propranolol* | 2.98 | 2.91 | 3.05 | 2.98 | 2.53 | 3.46 |
| 17 | *tetracaine* | 3.73 | 3.19 | 2.64 | 2.73 | 2.64 | 3.55 |
| 18 | *trimethoprim* | 0.91 | 1.04 | 1.70 | 0.72 | 1.26 | −0.07 |
| 19 | *verapamil* | 3.79 | 6.20 | 6.78 | 5.29 | 3.23 | 6.15 |
| | emp. SD *s* | | 0.72 | 0.90 | 0.52 | 0.54 | 1.19 |
| | *s* without mol. 19 | | 0.46 | 0.58 | 0.44 | 0.71 | 0.63 |

[a] The values were predicted using our ANN7-15-1 and ANN31-15-1 models.

## ACKNOWLEDGMENT

This work was realized within the scope of the SOL-project (Search and Optimization of Lead structures with neural networks and evolutionary algorithms) which was supported by the German Federal Ministry of Education and Research, bmb+f under contract no. 311681. We thank Dr. Claude Ostermann for data preparation, Gregor Wernet for the implementation of the global topological charge descriptor, and Kosmas Knödler and Jan Poland for their Matlab support. We also thank Hoffmann-La Roche and J. Huuskonen for providing us with the Huuskonen data set.

**Note Added after ASAP Posting.** This article was released ASAP on 5/2/2003. Two legend entries (GA,200 and GA,70) in Figure 5 were interchanged. The correct version was posted on 5/9/2003.

**Supporting Information Available:** The full descriptor statistics used for the Huuskonen and Wang data sets including Shannon Entropy (SE) values in decreasing order, the validation data sets in structured data file (SDF) format, and the cross-correlation matrices for the final descriptors used. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(2) 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.
(3) 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.
(4) 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.
(5) Luçić, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 121−132.
(6) 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.
(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(8) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Molecular Diversity* **1999**, *4*, 1−22.
(9) Brown, R. D.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.
(10) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL Keys as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(11) 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.
(12) 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.
(13) 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.
(14) 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.
(15) Banerjee, S.; Yalkowsky, S. H.; Valvani, S. C. Water Solubility and Octanol/Water Partition Coeffcients of Organics. Limitations of the Solubility-Partition Coeffcient Correlation. *Environ. Sci. Technol.* **1980**, *14*, 1227−1229.
(16) Zell, A. *Simulation neuronaler Netze*; Oldenbourg Verlag: München, 1997.
(17) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: D-69469 Weinheim, Germany, 2000.
(18) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(19) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(20) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(21) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(22) Liu, R.; So, S. S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.
(23) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanl/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aid. Mol. Des.* **2001**, *15*, 741−752.
(24) Weinenger, D. SMILES: a Chemical Language for Information Systems. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(25) Weinenger, D. SMILES 2: Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(26) MDL Information Systems. MDL CT file Formats.
(27) JOELib, http://joelib.sourceforge.net/.
(28) OELib, http://www.eyesopen.com/oelib/.
(29) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coeffcients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
(30) Wang, R.; Gao, Y.; Lai, L. Calculating partition coeffcient by atom-additive method. *Perspectives Drug Discovery Design* **2000**, *19*, 47−66.
(31) 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.
(32) MOE (Molecular Operating Environment), Chemical Computing Group Inc., 2002.
(33) Gálvez, J.; Garcia-Domenech, R.; Julián-Ortiz, V. D.; Soler, R. Topological Approach to Analgesia. *J. Chem. Inf. Comput. Sci.* **1994**, *14*, 1198−1203.
(34) Gasteiger, J.; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, 3181−3184.
(35) 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.
(36) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1245−1252.
(37) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, U.S.A., 1963.
(38) Agrafiotis, D. K. On the Use of Information Theory for Assessing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576−580.
(39) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.
(40) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon Entropy Analysis Identifies Molecular Property Descriptors that Predict Aqueous Solubility of Synthetic Compounds with High Accuracy in Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550−558.
(41) *Introduction to Algorithms;* Cormen, T. H., Leiserson, C. E., Rivest, R. L., Eds.; MIT-Press: 1998.
(42) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(43) Bomze, I.; Budinich, M.; Pardalos, P.; Pelillo, M. The maximum clique problem. In *Handbook of Combinatorial Optimization*; Du, D.-Z., Pardalos, P. M., Eds.; Kluwer Academic Publishers: Boston, MA, 1999; Vol. 4.

(44) Bron, C.; Kerbosch, J. Finding all cliques of an undirected graph. *Comm. ACM* **1973**, *16*, 575−577.
(45) Gardiner, E. J.; Holliday, J. D.; Willett, P.; Wilton, D. J.; Artymiuk, P. J. Selection of reagents for combinatorial synthesis using clique detection. *Quant. Struct.-Act. Relat.* **1998**, 17, 232−236.
(46) Java Neural Network Simulator, http://www-ra.informatik.uni-tuebingen.de/forschung/javanns/welcome.html.
(47) Stuttgart Neuronal Network Simulator, http://www-ra.informatik.uni-tuebingen.de/snns/.
(48) Alpaydm, E. Techniques for Combining Multiple Learners. *Proc. Eng. Intelligent Systems '98 Conference* **1998**, *2*, 6−12.
(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) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I. Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(51) Moriguchi, I.; Hirono, S.; Nakagome, I. Matsushita, Y. Comparison of Reliability of logP Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42*, 976−978.
(52) Rekker, R. F.; Laak, A. M. On the Reliability of Calculated Log P-values: Rekker, Hansch/Leo and Suzuki Approach. *Quant. Struct.-Act. Relat.* **1993**, *12*, 152−157.
(53) Suzuki, T.; Kudo, Y. Automatic log P estimation based on combined additive modeling methods *J. Comput. Aided Mol. Des.* **1990**, *4*, 155−198.

CI034006U