Carcinogenicity Prediction of Aromatic Compounds Based on Molecular Descriptors

Results from Applying Self-Organising Data Mining Technologies

Data Source

provided by Dr. Emilio Benfenati, Istituto di Ricerche Farmacologiche "Mario Negri" generated within the COMET project, ENV4-CT97-0508, funded by the European Comission.

Authors

Frank Lemke, Dipl.-Ing., frank@knowledgeminer.net
Johann-Adolf Müller, Prof. Dr.rer.oec.habil., jamueller@knowledgeminer.net

1. INTRODUCTION

"Man is exposed to many chemicals of natural and synthetic origin. An urgent question concerns their potential negative effects on human health. To identify chemicals inducing toxicity and to limit the incidence of human cancers and other diseases, rodent bioassays are the principal methods used today. However, this approach is not altogether problem-free, on several accounts: (1) the cost of the assay (>1 million U.S. dollars per chemical); (2) the time needed for the tests (3-5 years); (3) ethical considerations and public pressure to reduce or eliminate the use of animals in research and testing; (4) difficulties in the extrapolation to man.

We were interested in the prediction of carcinogenicity, but cancer is not a single disease. Several mechanisms involved in the various processes leading to the different tumors. This makes the task of assessing the computational prediction particularly challenging. Dedicated expert systems have been employed for computerized prediction of carcinogenicity. However, these have limitations. These expert systems work mainly on the assumption that toxicity is linked to the presence of toxic residues, either defined by human experts or found by the expert system. In some cases, the

expert systems also use some simple physicochemical parameters. ...

Another widespread approach for predicting toxicity relies on molecular descriptors, which refer to global properties or characteristics of the molecule. In recent years a huge increase in the number of studies of theoretical molecular descriptors has appeared in the literature, including their use in toxicity prediction." [Gini et al., 1999]

Using the above data set, we generated several models using self-organising data mining technologies: Clustering based on Analog Complexing pattern recognition, GMDH Neural Networks, Fuzzy Rule Induction and Nets of Active Neurons. These technologies are described in the book: Müller/Lemke, "Self-Organising Data Mining", Libri, 2000, ISBN 3-89811-861-4, for reference.

All models are obtained from using the entire data set: 104 aromatic compounds and 34 molecular descriptors. This paper is a summary of the initial report and is divided into the following sections:

2. Clustering

Analog Complexing can be considered a pattern recognition method for predicting, clustering and classification of fuzzy objects. Used for clustering, it identifies groups of similar objects. Picking a representative object from each generated cluster, a so-called nucleus - a most destinctive subset of objects - can be obtained. Such a nucleus can form the basis for further modeling activities like GMDH modeling or rule induction.

Using a pattern length of 104 and a similarity threshold of 95%, e.g., this clustering was obtained:

19 cluster found
C1: V1 (MW), -V7 (tenergy), V8 (volume), V9 (randic), V12 (chiv0),
V13 (chiv1), V14 (chiv2), V28 (Kappa1), V31 (kA1),
C2: V2 (homo),
C3: V3 (lumo),
C4: V4 (heat),
C5: V5 (dipole),
C6: V6 (polariz),
C7: V10 (balaban),

C8: V11 (wiener), V22 (MOM2), V23 (MOM3),
C9: V15 (chiv3), V16 (chiv4),
C10: V17 (flex), V29 (Kappa2), V30 (Kappa3), V32 (kA2), V33 (kA3),
C11: V18 (pH=2),
C12: V19 (pH=7.4),
C13: V20 (pH=10),
C14: V21 (MOM1),
C15: V24 (SIZE1),
C16: V25 (SIZE2),
C17: V26 (SIZE3),
C18: V27 (EllipsV),
C19: V34 (electrot),

3. GMDH Neural Networks

This special type of Neural Networks combines the best of both statistics and NNs as it inductively self-organisies networks of (self-organised) elementary functions systematically and fast. Also, the generated network function of self-selected relevant input factors is not hidden in the network, it is analytically available on the fly. The results can be analysed, gaining some insights into the investigated black box.

We extended the given data base of the 34 descriptors x_i by adding their invers values $1/x_i$ (where applicable). So the resulting data base we used for GMDH modeling consisted of 65 variables. We generated several linear and nonlinear models, and the best linear and nonlinear model is below referenced GMDHM1 and GMDHM2 respectively. The best linear model was:

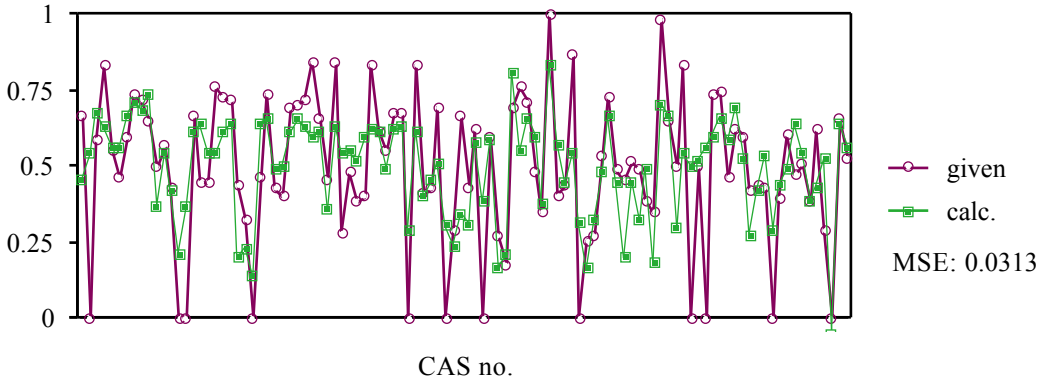
X74 = 1.7455 + 0.0006X6 -0.9133X66 -0.3383X18 -2.8758X47
+ 0.0449X33 + 0.0016X23 + 1.2038X60 -0.0333X50
-0.0335X8 -29.0323X41 + 3.4761X63

Prediction Error Sum Of Squares:	0.1344
Mean Absolute Percentage Error:	27.33 %
Approximation Error Variance:	0.5391

OUTPUT VARIABLE:
X74 - Carcinogenicity

RELEVANT INPUT VARIABLES:
X6 - heat
X66 - (kA2)^-1
X18 - chiv4
X47 - (chiv0)^-1

X33 - kA1
X23 - MOM1
X60 - (SIZE3)^-1
X50 - (chiv3)^-1
X8 - polariz
X41 - (polariz)^-1
X63 - (Kappa2)^-1



4. Fuzzy Rule Induction (FRI)

For fuzzy modeling, each descriptor x_i (i=1, 2, ..., 34) and the target variable y was fuzzified into 5 linguistic variables:

negative big	-	NB_<descriptor$_i$>
negative small	-	NS_<descriptor$_i$>
zero	-	ZO_<descriptor$_i$>
positive small	-	PS_<descriptor$_i$>
positive big	-	PB_<descriptor$_i$>

We used equidistant Lambda-type membership functions, $0 \le \mu^p(x) \le 1$, where $\mu^p(x)=0$ indicates no membership and $\mu^p(x)=1$ indicates full (exclusive) membership to the linguistic variable x^p (p=1, 2, ..., 5). Fuzzification was centered around the mean value $\bar{x}$ of descriptor x so that $\mu^{ZO}(\bar{x}) = 1$.

The resulting data base of 170 linguistic input variables and 5 linguistic target variables was used then to generate a separate rule for each linguistic target variable, i.e., 5 rules alltogether. A defuzzification model, finally, transforms the fuzzy results back into the initial data space. Fuzzification, rule induction and defuzzification was processed using KnowledgeMiner. We created two sets of fuzzy models: The first uses AND, OR operators [FM1], and the second model AND, OR, NOT operators [FM2].

Fuzzy models using AND, OR, NOT operators (fuzzy model 2 [FM2])

IF NB_homo **&** NB_lumo **&** NS_MOM3 **OR** **NOT**_PB_balaban **&** NB_pH=10
& NS_kA3 **OR** NB_wiener **&** NB_SIZE2 **&** NS_Kappa2
OR NS_SIZE1 **&** NB_lumo **&** NS_chiv2
THEN NB_Carcinogenicity

Summarized Absolute Error: 7.91
Mean Absolute Percentage Error: 7.61 %
Approximation Error Variance: 0.5212

OUTPUT VARIABLE:
X182 - NB_Carcinogenicity

RELEVANT INPUT VARIABLES:
X7 - NB_homo
X12 - NB_lumo
X113 - NS_MOM3
X51 - NOT_PB_balaban
X97 - NB_pH=10
X163 - NS_kA3
X52 - NB_wiener
X122 - NB_SIZE2
X143 - NS_Kappa2
X118 - NS_SIZE1
X68 - NS_chiv2

IF NS_heat **&** ZO_Kappa1 **OR** **NOT**_PS_SIZE2 **&** NS_dipole **&** NB_polariz
& **NOT**_ZO_homo **&** **NOT**_ZO_chiv3
THEN NS_Carcinogenicity

Summarized Absolute Error: 9.88
Mean Absolute Percentage Error: 9.50 %
Approximation Error Variance: 0.6599

OUTPUT VARIABLE:

X183 - NS_Carcinogenicity

RELEVANT INPUT VARIABLES:
X18 - NS_heat
X139 - ZO_Kappa1
X125 - NOT_PS_SIZE2
X23 - NS_dipole
X27 - NB_polariz
X9 - NOT_ZO_homo
X74 - NOT_ZO_chiv3

IF **NOT**_NB_pH=10 **&** **NOT**_NS_homo **&** **NOT**_NB_chiv3 **&** **NOT**_ZO_kA2 **&** ZO_heat
& **NOT**_NS_MOM2 **&** **NOT**_PS_chiv4 **&** **NOT**_NB_polariz **&** **NOT**_ZO_kA3
OR PB_pH=2 **OR** PS_MOM1 **OR** **NOT**_NB_pH=10
& **NOT**_NB_tenergy **&** **NOT**_NB_chiv3 **&** **NOT**_NS_MOM2 **&** ZO_heat
& **NOT**_ZO_kA2 **&** **NOT**_NB_pH=2 **&** **NOT**_NB_polariz **&** **NOT**_ZO_kA3
& **NOT**_NS_homo **&** **NOT**_PS_chiv4
THEN ZO_Carcinogenicity

Summarized Absolute Error: 24.21
Mean Absolute Percentage Error: 23.28 %
Approximation Error Variance: 0.8242

OUTPUT VARIABLE:
X184 - ZO_Carcinogenicity

RELEVANT INPUT VARIABLES:
X97 - NOT_NB_pH=10
X8 - NOT_NS_homo
X72 - NOT_NB_chiv3
X159 - NOT_ZO_kA2
X19 - ZO_heat
X108 - NOT_NS_MOM2
X80 - NOT_PS_chiv4
X27 - NOT_NB_polariz
X164 - NOT_ZO_kA3
X91 - PB_pH=2
X105 - PS_MOM1
X32 - NOT_NB_tenergy
X87 - NOT_NB_pH=2

IF PB_heat **OR** PS_chiv1 **& NOT** _PB_EllipsV **OR** ZO_polariz
& ZO_chiv3 **OR** **NOT** _NB_SIZE1 **& NOT** _NB_pH=7.4
& NOT _NB_Kappa2 **& NOT** _PB_chiv2 **& NOT** _ZO_homo **& NOT** _NS_pH=10
& NOT _PB_polariz **&** NS_balaban
THEN PS_Carcinogenicity

Summarized Absolute Error: 15.20
Mean Absolute Percentage Error: 14.62 %
Approximation Error Variance: 0.4706

OUTPUT VARIABLE:
X185 - PS_Carcinogenicity

RELEVANT INPUT VARIABLES:
X21 - PB_heat
X65 - PS_chiv1
X136 - NOT_PB_EllipsV
X29 - ZO_polariz
X74 - ZO_chiv3
X117 - NOT_NB_SIZE1
X92 - NOT_NB_pH=7.4
X142 - NOT_NB_Kappa2
X71 - NOT_PB_chiv2
X9 - NOT_ZO_homo
X98 - NOT_NS_pH=10
X31 - NOT_PB_polariz
X48 - NS_balaban

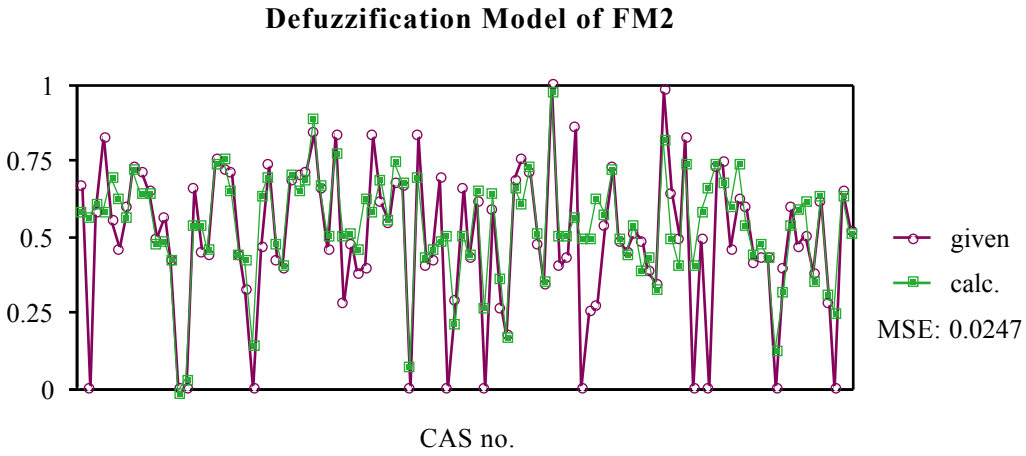
IF NB_homo **&** PB_polariz **OR** PB_balaban **&** PB_Kappa1
OR PB_flex **&** PB_volume **&** PS_balaban
THEN PB_Carcinogenicity

Summarized Absolute Error: 2.62
Mean Absolute Percentage Error: 2.52 %
Approximation Error Variance: 0.3080

OUTPUT VARIABLE:
X186 - PB_Carcinogenicity

RELEVANT INPUT VARIABLES:
X7 - NB_homo
X31 - PB_polariz
X51 - PB_balaban
X141 - PB_Kappa1
X86 - PB_flex

X41 - PB_volume
X50 - PS_balaban



5. Nets of Active Neurons

Since the fundamental McCulloch and Pitts work (1943) neurons are considered as binary, two or three equilibriums' states components of a Neural Network. The problem, however, is to organise a process that connects neurons into an effective ensemble or network known also as network topology optimisation. To solve this optimisation problem, the neurons have to have a Perceptrons-like structure. Using Perceptrons or other elements with a self-organisational behaviour as neurons, it is possible to create optimal Neural Networks. These elements are called Active Neurons.

Each neuron is an elementary system able to handle the same task as the complete network will do. This means, the transfer functions of active neurons are not fixed, they are self-organising systems themselves. One goal of nets of Active Neurons is improving accuracy. A second level of modelling is established using the models of the first level represented by their output values along with the initial input variables to train the network.

5.1 Using GMDH as Active Neuron

We prebuilt a NN topology of 65 input neurons, 4 hidden neurons (2 linear and 2 nonlinear GMDH algorithms) and one output neuron. The final network topology

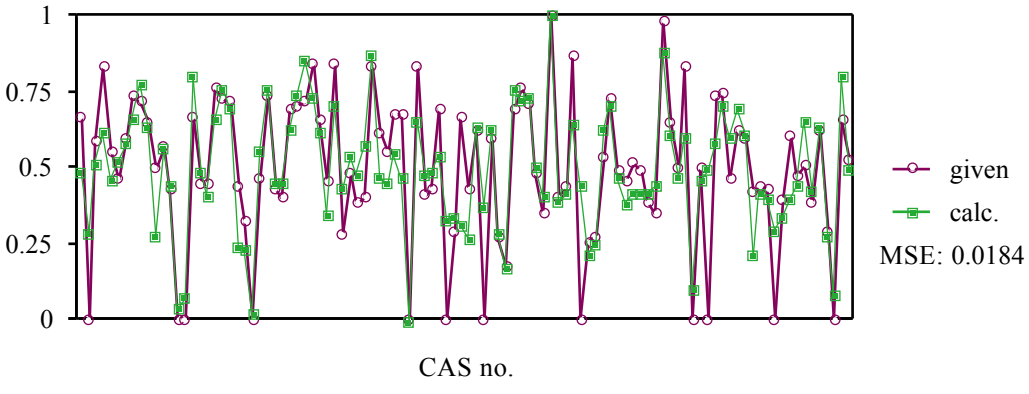
consists of 11 input neurons, one hidden neuron and the output neuron. Only one hidden neuron was selected, mainly because the GMDH algorithm we applied in 3. works using active neurons already (on the lowest level).

y=X79= $f(x_j)$, j - index of selected relevant input variables
(a very complex nonlinear function)

Prediction Error Sum Of Squares:	0.0778
Mean Absolute Percentage Error:	19.59 %
Approximation Error Variance:	0.3159

OUTPUT VARIABLE:
X79 - Carcinogenicity

RELEVANT INPUT VARIABLES:
X23 - MOM1
X71 - GMDHM2(Carcinogenicity)
X55 - (MOM1)^-1
X22 - pH=10
X54 - (pH=7.4)^-1
X49 - (chiv2)^-1
X40 - (heat)^-1
X53 - (pH=2)^-1
X18 - chiv4
X20 - pH=2
X48 - (chiv1)^-1
X17 - chiv3



5.2 Using FRI as Active Neuron

Here, the initial topology was 170 input neurons and 2 hidden neurons (FM1 and FM2). The resulting networks are much easier than corresponding GMDH based networks:

IF **NOT**_ZO_lumo **&** **NOT**_PB_SIZE1 **&** FM2(NB_Carcinogenicity)
THEN NB_Carcinogenicity

Summarized Absolute Error: 7.64
Mean Absolute Percentage Error: 7.34 %
Approximation Error Variance: 0.5165

OUTPUT VARIABLE:
X192 - NB_Carcinogenicity

RELEVANT INPUT VARIABLES:
X14 - NOT_ZO_lumo
X121 - NOT_PB_SIZE1
X187 - FM2(NB_Carcinogenicity)

IF **NOT**_NB_pH=2 **&** **NOT**_PS_polariz **&** FM1(NS_Carcinogenicity)
& **NOT**_PS_SIZE3 **&** **NOT**_NB_dipole **&** **NOT**_PS_chiv0 **&** **NOT**_NB_homo
& **NOT**_NS_tenergy
THEN NS_Carcinogenicity

Summarized Absolute Error: 7.28
Mean Absolute Percentage Error: 7.00 %
Approximation Error Variance: 0.2499

OUTPUT VARIABLE:
X193 - NS_Carcinogenicity

RELEVANT INPUT VARIABLES:
X87 - NOT_NB_pH=2
X30 - NOT_PS_polariz
X178 - FM1(NS_Carcinogenicity)
X130 - NOT_PS_SIZE3
X22 - NOT_NB_dipole
X60 - NOT_PS_chiv0
X7 - NOT_NB_homo
X33 - NOT_NS_tenergy

IF **NOT**_NB_tenergy **&** **NOT**_NB_pH=10 **&** FM2(ZO_Carcinogenicity)
& **NOT**_NB_pH=2 **&** **NOT**_ZO_kA2
THEN ZO_Carcinogenicity

Summarized Absolute Error: 22.77
Mean Absolute Percentage Error: 21.90 %
Approximation Error Variance: 0.7903

OUTPUT VARIABLE:
X194 - ZO_Carcinogenicity

RELEVANT INPUT VARIABLES:
X32 - NOT_NB_tenergy
X97 - NOT_NB_pH=10
X189 - FM2(ZO_Carcinogenicity)
X87 - NOT_NB_pH=2
X159 - NOT_ZO_kA2

IF **NOT**_PB_chiv2 **&** **NOT**_PS_MOM1 **&** **NOT**_NB_pH=7.4
& FM2(PS_Carcinogenicity)
THEN PS_Carcinogenicity

Summarized Absolute Error: 14.14
Mean Absolute Percentage Error: 13.60 %
Approximation Error Variance: 0.4072

OUTPUT VARIABLE:
X195 - PS_Carcinogenicity

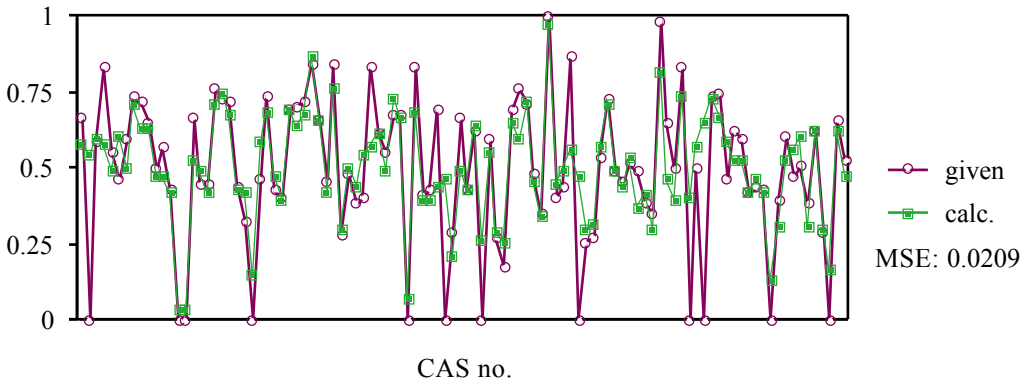
RELEVANT INPUT VARIABLES:
X71 - NOT_PB_chiv2
X105 - NOT_PS_MOM1
X92 - NOT_NB_pH=7.4
X190 - FM2(PS_Carcinogenicity)

IF **NOT**_PB_chiv3 **&** FM1(PB_Carcinogenicity)
THEN PB_Carcinogenicity

Summarized Absolute Error: 2.3566
Mean Absolute Percentage Error: 2.27 %
Approximation Error Variance: 0.2840

OUTPUT VARIABLE: X196 - PB_Carcinogenicity

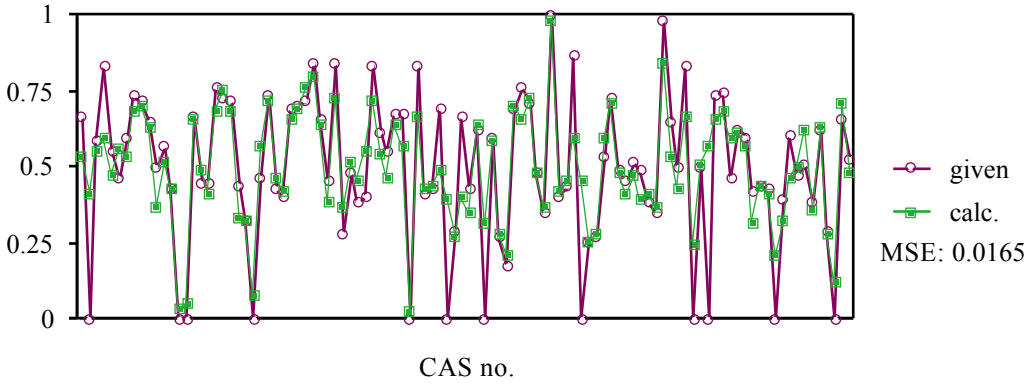
RELEVANT INPUT VARIABLES:
X76 - NOT_PB_chiv3
X181 - FM1(PB_Carcinogenicity)



6. Synthesis

All kinds of parametric, nonparametric, algebraic, binary/fuzzy logic models are only simplified reflections of reality. There are always several models with a sufficient degree of adequacy for a given data sample. Every model is a specific abstraction, a one-sided reflection of some important features of reality only. A synthesis of alternate model results gives a more thorough reflection.

Averaging the output values of both Active Neuron models shows this improvements:



7. Summary

We generated several parametric and nonparametric models using self-organising data mining technologies to describe and analyse experimental carcinogenicity values from molecular descriptors. In particular, the following results are obtained.

1. A clustering in the sample space using all 34 descriptors and decreasing similarity thresholds results in increasingly distinctive clusterisations. By choosing one descriptor representatively from each cluster, a most important set of descriptors - called nucleus - is obtained. Such a nucleus can form the basis of a second modeling run.

2. Generated GMDH models confirm the results obtained from the Backpropagation NN: Nonlinear relation and comparable input-output behavior. Additionally, GMDH provides an anlytical model composed of a subset of relevant descriptors. This extracted information can be used for further analysis (table 7.1). Since this parametric models are appropriated for moderately noisy data sets, the noise level/ uncertainty of the target variable is an important factor here. From this perspective, model accuracy based on some closeness-of-fit measure can be of limited importance only, because GMDH takes noise into account to not overfit the data. So excluding badly described (or a priori uncertain) objects (fig. 7.1) from the data set may not only improve model accuracy, but its descriptive power too.

3. Fuzzy models seem most promising from both predictive and descriptive power. The models are easy to understand and to analyse, and their predictive behavior do quasi not differ from that of BP or GMDH NNs.

4. Nets of Active Neurons and a synthesis of model results confirmed being powerful tools to increase model accuracy beyond that of a single model. A mixed model approach using GMDH, FRI, and other NN models in the hidden layer might be interesting to test.

5. Concluding from all reported models, a set of 10 descriptors used in more than 80% of the models can be identified (table 7.1). Compared to the 13 descriptors used in BP NN modeling as reported in Gini [Gini et al., 1999], only 4 descriptors (homo, lumo, polariz, chiv3) are included in this selected set here too, while 1 descriptor (MW) was never used in a model.

6. In table 7.2 and fig. 7.1 the minimum, maximum and mean absolute errors of 8 different models having a MSE < 0.025 are shown. Looking at the max. absolute errors, a quite clear distinction between good and badly described compounds is possible when using an abs. error of 0.3 as a threshold (fig. 7.1.a). A very similar picture is shown if using the mean absolute error of the 8 models and a threshold of 0.18 (fig. 7.1.b).

Table 7.1: Descriptors used in reported models

	MW	homo	lumo	heat	dipole	polariz	tenergy	volume	randic	balaban	wiener	chiv0	
GMDHM1				1		1						1	GMDHM1
GMDHM2		1	1	1	1	1					1	1	GMDHM2
FM1		1	1	1		1		1		1	1		FM1
FM2		1	1	1	1	1	1	1		1	1		FM2
GMDH NAN		1	1	1	1	1					1	1	GMDH NAN
FRI NAN		1	1	1	1	1	1	1		1	1	1	FRI NAN
% used	0.0%	83.3%	83.3%	100.0%	66.7%	100.0%	33.3%	50.0%	0.0%	50.0%	83.3%	66.7%	

	chiv1	chiv2	chiv3	chiv4	flex	pH=2	pH=7.4	pH=10	MOM1	MOM2	MOM3		
GMDHM1			1	1					1				GMDHM1
GMDHM2	1	1	1	1	1				1				GMDHM2
FM1		1	1			1		1	1		1		FM1
FM2	1	1	1	1	1	1	1	1	1	1	1		FM2
GMDH NAN	1	1	1	1	1	1	1	1	1				GMDH NAN
FRI NAN	1	1	1	1	1	1	1	1	1	1	1		FRI NAN
% used	66.7%	83.3%	100.0%	83.3%	66.7%	66.7%	50.0%	66.7%	100.0%	33.3%	50.0%		

	SIZE1	SIZE2	SIZE3	EllipsV	Kappa1	Kappa2	Kappa3	kA1	kA2	kA3	electrot		
GMDHM1			1			1		1	1			10	
GMDHM2										1	1	15	
FM1	1	1		1	1	1	1			1		20	
FM2	1	1		1	1	1			1	1		27	
GMDH NAN										1	1	18	
FRI NAN	1	1	1	1	1	1	1		1	1		30	
% used	50.0%	50.0%	33.3%	50.0%	50.0%	66.7%	33.3%	16.7%	50.0%	83.3%	33.3%		% used

Min. and Max. Absolute Errors from 8 Generated Models

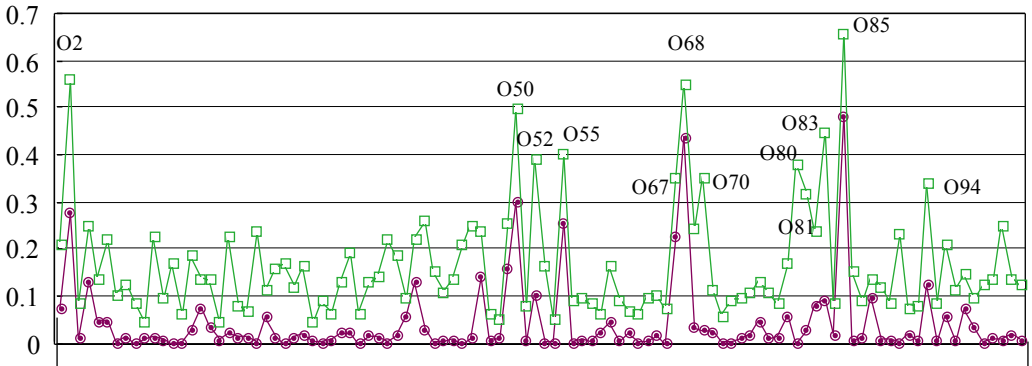


Fig. 7.1.a.

Averaged Absolute Error from 8 Generated Models

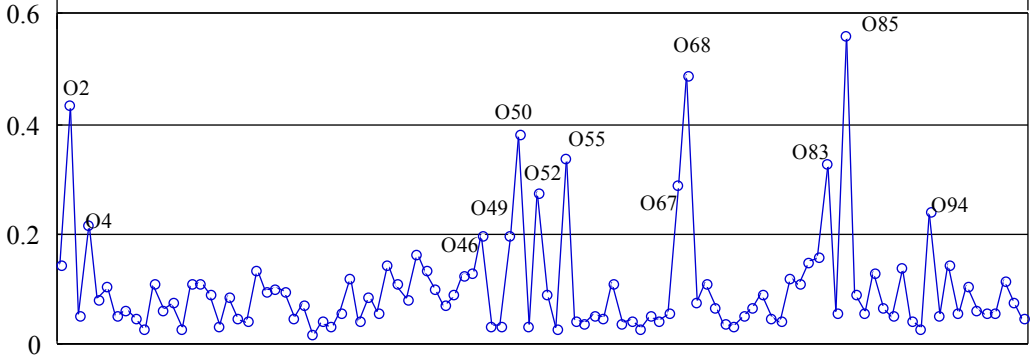


Fig. 7.1.b.

The obtained models and results may not be final. In fact, the knowledge and the conclusions gained from the reported set of models should be included into future modeling to improve reliability and understanding.

For questions on this report, self-organising data mining, or our data mining service,

please contact the authors at

frank@knowledgeminer.net
jamueller@knowledgeminer.net

Berlin, June 15, 2000

Frank Lemke *Johann-Adolf Müller*

Reference

Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M.: Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. In *Journal of Chemical Information and Computer Sciences* , 39(1999)6, pp. 1076-1080