Available online at www.sciencedirect.com
ScienceDirect
Journal of Pharmaceutical and Biomedical Analysis 43 (2007) 119–130
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs

E. Deconinck[a], D. Coomans[b], Y. Vander Heyden[a,*]

[a] *Department of Analytical Chemistry and Pharmaceutical Technology, Pharmaceutical Institute, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Statistics & Intelligent Data Analysis Group, James Cook University, Townsville 4814, Australia*

Received 19 April 2006; received in revised form 9 June 2006; accepted 10 June 2006
Available online 21 July 2006

Abstract

In general, linear modelling techniques such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS), are used to model QSAR data. This type of data can be very complex and linear modelling techniques often model only a limited part of the information captured in the data. In this study, it was tried to combine linear techniques with the flexible non-linear technique multivariate adaptive regression splines (MARS). Models were built using an MLR model, combined with either a stepwise procedure or a genetic algorithm for variable selection, a PCR model or a PLS model as starting points for the MARS algorithm. The descriptive and predictive power of the models was evaluated in a QSAR context and compared to the performances of the individual linear models and the single MARS model.

In general, the combined methods resulted in significant improvements compared to the linear models and can be considered valuable techniques in modelling complex QSAR data. For the used data set the best model was obtained using a combination of PLS and MARS. This combination resulted in a model with a Pearson correlation coefficient of 0.90 and a cross-validation error, evaluated with 10-fold cross-validation of 9.9%, pointing at good descriptive and high predictive properties.

Keywords: QSAR; Drug absorption; *In silico* prediction; TMARS

1. Introduction

In silico screening of newly synthesised molecules for absorption, distribution, metabolisation, elimination and toxicity (ADME-Tox) properties has become a very important issue in drug development. This to prevent molecules, positively screened for their interaction with target molecules from being developed and failing in a later phase of the drug development due to non-proper ADME-Tox properties.

In silico screening has the advantage that it can take place before molecules are even synthesised. These methods try to build relationships between a data set consisting of known values for the property of interest and some calculated theoretical descriptors. Theoretical descriptors are derived from molecular representations of the molecules, thus no synthetisation nor experimental set-ups are required. These relationships are called quantitative structure–activity relationships (QSAR). This paper focuses on building QSARs for gastro-intestinal absorption of drugs.

Different QSAR-models for absorption of molecules can be found in the literature. For these models different modelling techniques were used, including linear techniques like multiple linear regression (MLR) [1–3], principal components regression (PCR) [4], partial least squares (PLS) [4–7], as well as non-linear chemometric techniques like artificial neural networks (ANN) [8], classification and regression trees (CART) [9,10] and multivariate adaptive regression splines (MARS) [11]. A possible problem with these models is that they are either linear or non-linear. Since the data space possibly consists of linear and non-linear regions, it can be assumed that a combination of linear with non-linear techniques can give significant improvements of the descriptive- and predictive properties of the models.

[*] Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

0731-7085/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2006.06.022

Previous work in quantitative structure–retention relationships (QSRR) [12] and QSAR [11] has shown that the combination of stepwise-MLR with MARS, called two-step MARS (TMARS), can substantially improve the description and prediction of a data set compared to the individual MLR-model. Xu et al. [12] showed that TMARS significantly improved the predictive abilities of the stepwise-MLR model for retention in gas chromatography and we [11] showed that TMARS can be a valuable technique in QSAR, since it is capable of significantly improving an MLR model for gastro-intestinal absorption.

In this paper it was investigated whether better models could be obtained when the stepwise linear regression step in the TMARS approach is replaced by other linear modelling techniques. Therefore linear models were built using MLR, PCR and PLS for the gastro-intestinal absorption of drugs. In a next step these models were used as starting point for MARS, resulting in different models combining a linear technique with the non-linear technique MARS. The descriptive and predictive properties of all models as well as the capability of MARS to improve these properties was evaluated.

2. Theory

2.1. Multiple linear regression (MLR)

MLR is the most widely known multiple linear modelling technique. Normally MLR cannot be used to model complex QSAR data, due to the fact that in most cases the number of descriptive variables exceeds the number of objects. Therefore variable selection is necessary prior to modelling. In this work two methods were used. The first is stepwise MLR. In this technique a forward selection procedure iterates with a backward elimination procedure. The forward selection procedure starts with the variable that has the highest correlation with the response variable, for example the gastro-intestinal absorption. If this variable results in a significant regression, evaluated with an overall *F*-test, the variable is retained and selection continues. In a next step the variable that gives the largest significant increase of the regression sum of squares, evaluated with a partial *F*-test, is added. After each step of the forward selection procedure, the backward elimination procedure is applied. In this procedure a partial *F*-test for the variables already in the model is performed. If a variable is found that does not longer contributes significantly to the regression it is removed from the model. The iteration is repeated until the model cannot be improved anymore by adding or removing variables [13].

The second variable selection method used consists of a genetic algorithm procedure that was followed by MLR with the selected variables. In a genetic algorithm procedure for variable selection, a population of strings is randomly created. Each string consists of a row-vector with a number of elements equal to the number of descriptive variables in the data set. Each string consists of zeros and ones, zero indicating that the corresponding variable is not selected and one indicating that it is. The fitness of each is equal to the value of the evaluation criterion that has to be optimised by the algorithm. In this work, the root mean squared error of cross-validation (RMSECV) of the MLR model build with the variables selected in the string was used as evaluation criterion. The randomly selected initial population of strings then evolves. The strings with the highest fitness are selected and undergo crossover and mutations. The procedure is repeated a number of times. The selection of variables with the lowest RMSECV value is then selected for MLR modelling [14,15].

2.2. Principal component regression (PCR)

PCR is in fact multiple linear regression using the scores on a number of principal components as descriptive variables. The principal components are latent variables. The two main advantages of PCR over MLR are the reduction of the number of variables to a maximal number below the number of objects and decorrelation of the variables. In a first step the significant principal components are selected. In a second step MLR is applied using the scores on the selected principal components as latent variables [13].

2.3. Partial least squares (PLS)

PLS is an alternative for PCR. The latent variables in PLS are also linear combinations of the descriptive variables in the data set, but instead of maximising the variance in the matrix with descriptive variables like in PCA, the covariance with the response variable is maximised. The scores on the PLS factors are used as input for multiple linear regression after selection of the optimal number of PLS-factors to be considered [13].

2.4. Multivariate adaptive regression splines (MARS)

MARS is a local modelling technique, dividing the dataspace in several, possibly overlapping regions and fitting truncated spline functions in each region. A truncated spline function consists of a left-sided, Eq. (1), and a right-sided, Eq. (2), segment, separated by a so-called knot location [11].

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases} (t-x)^q & \text{if } x < t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases} (x-t)^q & \text{if } x > t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $b_q^-(x-t)$ and $b_q^+(x-t)$ are the spline functions describing the regions right and left of the knot location *t*, respectively, and *q* the power to which the spline is raised. The subscript "+" indicates that the result of the function is 0 when the argument is not satisfied. For each of the descriptive variables in the data set MARS selects the pair of spline functions and the knot location that best describes the response variable. In the following step all the spline functions are combined in a complex non-linear model, describing the response as a function of the descriptive variables. The model has the form:

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{3}$$

where $\hat{y}$ is the predicted value for the response variable, a_0 the coefficient of the constant term, M the number of spline functions, and B_m and a_m the mth spline function and its coefficient [16,17], respectivily.

In general, a MARS analysis consists of three steps. In a first step, the variable for which the selected pair of spline functions gives the best description of the response variable is selected as starting point for the model. After the selection of this first pair of splines, new spline functions are added stepwise. In each step of the stepwise addition procedure the pair of splines is added that gives the best improvement in the description of the training set. This procedure finally results in a complex multivariate model, the global MARS-model, which almost perfectly describes the training set and usually shows overfitting. In a second step of the analysis, the global MARS-model is pruned using a sequence of general cross-validations (GCV) alternated with 10-fold cross validation. During this procedure the contribution of each spline function to the descriptive abilities of the model is evaluated using a lack-of-fit (LOF) criterion. The least contributing functions are eliminated stepwise. This pruning process results in a sequence of models with different sizes. In the final step of the MARS analysis the optimal model is selected using a cross-validation technique. The theory concerning MARS is discussed in more detail in literature [11,16–18].

2.5. Two-step MARS (TMARS)

TMARS is in fact a combination of the stepwise MLR procedure described in Section 2.1 and MARS. In a first step the stepwise-MLR model is built. Starting from this linear model, MARS is applied in the second step. During this procedure it is evaluated if some of the descriptors of the linear model can be replaced by a pair of spline functions. If so the descriptors are removed from the model and replaced by a pair of spline functions. Starting from the obtained model, the stepwise addition procedure, described in Section 2.4 is applied, resulting in the so-called global TMARS model. After obtaining the global model, pruning and selection of the optimal model is carried out as in Section 2.4. The theory of TMARS is explained extensively in Refs. [11,12].

2.6. Theoretical molecular descriptors

A theoretical molecular descriptor is the final result of a logical and mathematical procedure, which converts the chemical information from a symbolic representation of the molecule in a useful numerical value [19]. The number of theoretical descriptors described in the literature is still growing. Different classification systems for these descriptors can be found. The most applied is that based on the molecular representation from which the descriptor is derived. This results in five classes, zero-, one-, two-, three- and four-dimensional descriptors, derived from a molecular formula, a substructure list, a topological, a geometrical and a stereoelectronic or lattice representation of the molecule, respectively. More information about molecular descriptors and their classification can be found in the work of Todeschini and Consonni [19].

3. Material and methods

3.1. Data

The used data set [11] consists of the percentages human intestinal absorption for 140 molecules. The names of the drugs and drug-like compounds together with their percentage intestinal absorption (%HIA) are listed in Table 1

. These data were selected because they consist of the absorption data for a high diversity of molecular structures and because they cover the whole range of the absorption scale (0–100%).

3.2. Three-dimensional structure optimisation

The three-dimensional structures of the molecules were drawn and optimized using the Hyperchem® 6.03 professional software (Hypercube, Gainesville, FL, USA). After the input of the molecule as a topological structure, geometry optimisation was obtained by the molecular mechanics force field method (MM+) using the Polak–Ribière conjugate gradient algorithm with a RMS gradient of 0.1 kcal/(Å mol) as stop criterion. The optimisation of the structure results in a data matrix consisting of the Cartesian coordinates of the atoms. This data matrix can then be used to calculate molecular descriptors [9,11].

3.3. Calculating molecular descriptors

Molecular descriptors were calculated using the Dragon® 4.0 professional software [20]. This program allows to calculate 48 constitutional descriptors, 119 topological descriptors, 47 walk and path counts, 33 connectivity indices, 47 information indices, 96 2D autocorrelations, 107 edge adjacency indices, 64 BCUT-descriptors, 21 topological charge indices, 44 eigenvalue-based indices, 41 randic molecular profiles, 74 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 197 GETAWAY descriptors, 121 functional group counts, 120 atom-centered fragments, 14 charge descriptors and 28 molecular properties. More information about the above descriptors can be found in the work of Todeschini et al. [19]. The software automatically eliminates descriptors resulting in constant values for a given data set. For descriptors with a correlation higher than 0.98, parameters are set as such that only one is retained in the data set. Furthermore the Hyperchem® 6.03 professional software was used to calculate solvent accessible surface area, molecular volume, octanol/water partition coefficient (log P), hydration energy, molar refractivity, molar polarisability and molar mass [9,11]. The Mc. Gowans volume, a descriptor used in the linear free energy relationship (LFER) of Abraham [19,21] was calculated manually. The final dataset consisted of the values for 761 different descriptors for 140 objects.

Table 1
The absorption data for the 140 molecules of the training set [11]

No.	Substance	%HIA
1	Acarbose	1.5
2	Acebutolol	89.75
3	Acetaminophen	85
4	Acetylsalicylic acid	100
5	Acrivastine	88
6	Acyclovir	25
7	Adefovir	12
8	Alprenolol	93.75
9	Aminopyrine	100
10	Amoxicillin	93.75
11	Amphotericin B	5
12	Amrinone	93
13	Antipyrine	100
14	Atenolol	51
15	Atropine	90
16	Azithromycin	36
17	Aztreonam	1
18	Benazepril	37
19	Benzylpenicillin	27.5
20	Betaxolol	90
21	Bornaprine	100
22	Bretyliumtosylate	23
23	Bromazepam	84
24	Bromocriptine	28
25	Bumetanide	100
26	Bupropion	87
27	Caffeine	100
28	Camazepam	99
29	Captopril	68
30	Cefatrezine	76
31	Ceftriaxone	1
32	Cefuroxime	5
33	Cefuroximeaxetil	36
34	Cephalexin	98.5
35	Chloramphenicol	90
36	Chlorothiazide	23.75
37	Cimetidine	82.5
38	Ciprofloxacin	84.5
39	Cisapride	100
40	Clonidine	96.25
41	Codein	95
42	Corticosterone	100
43	Cromolynsodium	0.5
44	Cymarin	47
45	Cyproterone acetate	100
46	Dexamethasone	98
47	Diazepam	99.25
48	Doxorubicin	5
49	Enalapril	66
50	Enalaprilat	17.5
51	Erythromycin	35
52	Ethambutol	77.5
53	Ethinylestradiol	100
54	Etoposide	50
55	Felbamate	92.5
56	Fenoterol	60
57	Fluconazole	96.25
58	Foscarnet	17
59	Fosinopril	36
60	Fosmidomycin	30
61	Furosemide	61
62	Gabapentin	50
63	Ganciclovir	3.6
64	Guanabenz	75
65	Guanoxan	50
66	Hydrochlorothiazide	72.75
67	Hydrocortisone	90.25
68	Imipramine	96.25
69	Indomethacin	100
70	Iothalamatesodium	1.9
71	Isoxicam	100
72	Isradipine	92.5
73	Labetalol	93.75
74	Lactulose	0.6
75	Lamotrigine	70
76	Levodopa	85
77	Lincomycin	27.5
78	Lisinopril	25
79	Loracarbef	100
80	Lormetazepam	100
81	Lovastatin	30.5
82	Mannitol	20
83	Meloxicam	90
84	Metaproterenol	44
85	Methotrexate	80
86	Methyldopa	41
87	Methylprednisolone	82
88	Metolazone	63
89	Metoprolol	95
90	Morphine	100
91	Nadolol	31
92	Nefazodone	100
93	Naloxone	91
94	Nordiazepam	99
95	Norfloxacin	35
96	Olsalazine	2.3
97	Ouabain	1.4
98	Oxatomide	100
99	Oxazepam	98.5
100	Oxprenolol	91.75
101	Phenoxymethylpenicillin	45
102	Phenytoin	90
103	Pindolol	91.75
104	Piroxicam	100
105	Practolol	98.75
106	Pravastatin	34
107	Prazosin	100
108	Prednisolone	98.9
109	Progesterone	93.25
110	Propranolol	92.5
111	Propiverine	84
112	Propylthiouracil	75
113	Quinidine	80.25
114	Raffinose	0.3
115	Ranitidine	52.75
116	Reproterol	60
117	Saccharin	88
118	Salicylic acid	100
119	Scopolamine	92.5
120	Sorivudine	82
121	Sotalol	96.25
122	Spironolactone	73
123	Sudoxicam	100
124	Sulfasalazine	38.75
125	Sulindac	90
126	Sulpiride	36
127	Sumatriptin	70
128	Terazosin	93.25
129	Terbutaline	66.5

Table 1 (*Continued*)

No.	Substance	%HIA
130	Testosterone	100
131	Theophylline	96
132	Timolol maleate	85.5
133	Tranexamicacid	55
134	Trimethoprim	97
135	Trovoflaxicin	88
136	Venlafaxine	92
137	Verapamil	95
138	Warfarin	98.5
139	Ximoprofen	100
140	Zidovudine	100

3.4. Building models

The MLR, PCR and PLS models were built using the algorithms of an in-house toolbox written for Matlab 6.5 (The Mathworks, Natick, MA). The genetic algorithm was programmed in Matlab 6.5 according to the algorithm described by Jouan-Rimbaud et al. [15]. The MARS algorithm was programmed according to the original MARS algorithm proposed by Friedman [16]. The log transformed absorption values were used as response variables and the calculated theoretical descriptors as explanatory variables. The complete set of descriptors was autoscaled.

4. Results and discussion

4.1. Multiple linear regression (MLR)

4.1.1. Stepwise MLR

The stepwise MLR algorithm was applied to the data set using the Briggsian logarithm of the %HIA values as response variables and the autoscaled calculated descriptors as descriptive variables. The obtained linear model consists of 13 terms, with one constant term and 12 terms based on different descriptors. The model is given by the following equation:

$$\begin{aligned}\hat{y} = {} & 1.89 - 1.81(n\mathrm{O}) - 1.19(\mathrm{TIE}) - 0.43(\mathrm{D/Dr05}) \\ & - 1.19(\mathrm{T(S..S)}) + 0.70(\mathrm{IC2}) - 2.04(\mathrm{Mor08m}) \\ & + 0.52(\mathrm{Mor16v}) + 1.13(\mathrm{HATS8v}) - 0.52(n\mathrm{N{=}N}) \\ & + 0.50(n\mathrm{N(CO)_2}) - 0.0814(n\mathrm{OH}) - 0.40(\mathrm{C\text{-}0.30}) \qquad (4)\end{aligned}$$

Table 2 shows the selected descriptors, their definition and class. Some of these descriptors can be related to the oxygen (nO, nOH, $n\mathrm{N(CO)_2}$), sulfur (T(S··S)) and nitrogen (nN = N, $n\mathrm{N(CO)_2}$) atoms present in the molecule. These atoms and the functional groups containing them are important as proton donors/acceptors in the molecule and so important in the calculation of the polar surface area (PSA). The PSA is a measure for the H-bonding capacity of a molecule and it has been found that processes involving passive diffusion depend primarily on these H-bonding properties [22]. Most of the other descriptors can be related to the two-dimensional (TIE, D/Dr05 and IC2) or three-dimensional (Mor08m, Mor16v and HATS8v) structure of the molecules. The descriptor C-030 is based on a code describing each carbon atom through its atom type, bonding types and neighbouring atom types in the molecule [19].

Table 2
Selected descriptors [11,19] in the stepwise MLR-model

Descriptor	Definition	Descriptor class
*n*O	Number of oxygen atoms	Constitutional descriptors
TIE	E-state topological parameter	Topological descriptors
D/Dr05	Distance/detour ring index of order 5	Topological descriptors
T(S··S)	Sum of topological distances between S··S	Topological descriptors
IC2	Information content index (neighborhood symmetry of 2-order)	Information indices
Mor08m	3D-MoRSE-signal 08/weighted by atomic masses	3D-MoRSE descriptors
Mor16v	3D-MoRSE-signal 16/weighted by atomic van der Waals volumes	3D-MoRSE descriptors
HATS8v	Leverage-weighted autocorrelation of lag 8/weighted by atomic van der Waals volumes	GETAWAY descriptors
*n*N=N	Number of N azo (aliphatic)	Functional group counts
$n\mathrm{N(CO)_2}$	Number of imides	Functional group counts
*n*OH	Number of total hydroxyl groups	Functional group counts
C-030	X–CH–X	Atom-centered fragments

The coefficient of determination R^2 equals 0.616. Fig. 1 shows the residual plot for the stepwise MLR model.

Based on the R^2-value it can be concluded that the predicted and observed log(%HIA) are not highly correlated. The residual plot shows that there is no random distribution of the residuals, which can imply that the model is showing under fitting. The root mean squared error of cross validation (RMSECV), evaluated with 10-fold cross-validation for this model is 0.331 or 19.6%, calculated on the mean value of the %HIA values.

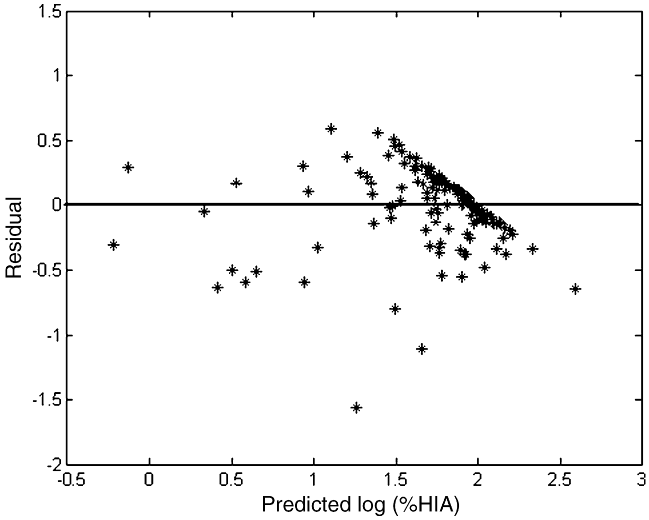

Fig. 1. Residual plot for the stepwise MLR model.

Despite the fact that the RMSECV-value is relatively low, it could be concluded that, based on the low correlation between real and predicted values, the absence of random distribution in the residual plot and the high residuals, predictions based on this model will not be reliable. The model is not suited for our purposes, it has no satisfying descriptive and predictive abilities and should be improved.

4.1.2. Variable selection with a genetic algorithm combined with MLR

The genetic algorithm was applied to the data set. The optimal input parameters were selected using a two-level full factorial design. For each of the indicated parameters two values, a high and a low, were selected. Each possible combination of these factors was used to build a model. The values for the input parameters, with which the best model was obtained, were set as default. Input parameters were set as follows: number of outer cycles*: 10; number of chromosomes in the population: 20; probability for a variable to undergo crossover*: 25%; probability of mutations*: 0.5%; number of generations*: 200; number of deletion groups for performing cross validation: 10; maximal RMSEP: 0.3; frequency for the backward stepwise elimination*: 100.

The obtained linear model consists of 11 terms and is given by the following equation:

$$\begin{aligned}\hat{y} = {} & 2.68 - 2.07(n\text{O}) + 0.42(\text{Yindex}) - 0.73(\text{GATS2p}) \\ & + 0.62(\text{EEig11x}) + 0.06(\text{EEig01r}) + 0.38(\text{RDF065m}) \\ & - 0.64(\text{Mor05u}) - 1.70(\text{H3m}) + 0.30(\text{H6e}) \\ & + 0.24(\text{C-003}) \qquad (5)\end{aligned}$$

Table 3 shows the selected descriptors, their definition and class. The descriptor *n*O, reflects the number of oxygen atoms in the molecule. These atoms are important in the calculation of PSA, a property that can easily be related to the processes of gastrointestinal absorption. Most other descriptors can be related to the topological (Yindex, GATS2p, EEig11x and EEig01r) and geometrical (RDF065m, Mor05u, H3m and H6e) structure of the molecule. The descriptor C-003 is based on a code describing each carbon atom through its atom type, bonding types and neighbouring atom types in the molecule [19].

Table 3
Selected descriptors [19] in the genetic algorithm-MLR model

Descriptor	Definition	Descriptor class
*n*O	Number of oxygen atoms	Constitutional descriptors
Yindex	Balaban Yindex	Information indices
GATS2p	Geary autocorrelation-lag 2/weighted by atomic polarizabilities	2D autocorrelation
EEig11x	Eigenvalue 11 from edge adjacent matrix weighted by edge degrees	Edge adjacency indices
EEig01r	Eigenvalue 01 from edge adjacent matrix weighted by resonance integrals	Edge adjacency indices
RDF065m	Radial distribution function-6.5/weighted by atomic masses	RDF descriptors
Mor05u	3D-MoRSE-signal 05/unweighted	3D-MoRSE descriptors
H3m	H autocorrelation of lag 3/weighted by atomic masses	GETAWAY descriptors
H6e	H autocorrelation of lag 6/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
C-003	CHR3	Atom-centered fragments

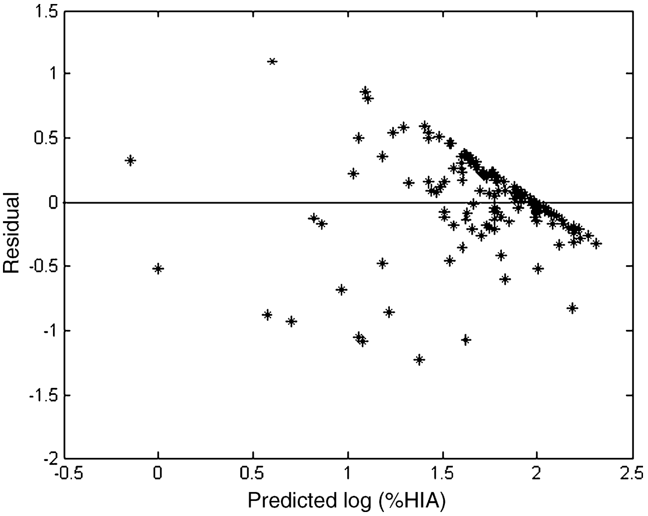

Fig. 2. Residual plot for the genetic algorithm-MLR model.

The coefficient of determination R^2 equals 0.576. Fig. 2 shows the residual plot for the genetic algorithms-MLR model.

The RMSECV-value for this model is 0.346 or 20.5%, evaluated with 10-fold cross-validation. Based on the R^2-value, the residual plot and the RMSECV similar conclusions could be made as for the stepwise-MLR model. Predicted and observed values are not highly correlated, there is no random distribution of the residuals and therefore the predictions based on this model will not be reliable.

4.2. Principal component regression (PCR)

PCA is applied to the data set. The first eight principal components are selected for the PCR-model. They represent 93.6% of the total variance in the matrix of the descriptive variables. The first principal component (PC1) represents 86.0% of the variance. The selection of the optimal number of principal components for the model is based on leave-one-out cross-validation. The model with the lowest RMSECV was selected as optimal. Based on the loadings of the different descriptors on PC1, it is possible to have an idea which descriptors have a higher impact on the model. Table 4 shows the 10 descriptors with the highest loadings on PC1, their definition and class.

All descriptors can be related to either the two-dimensional (BEHv1, BEHe1, BEHp1, BIC4, ATS1v and PJI2) or three-dimensional (ISH, Mecc, FDI and Mor04m) structure of the molecules [19].

Fig. 3 shows the residual plot for the PCR-model. The coefficient of determination R^2 equals 0.293 and the RMSECV-value is 0.449 or 26.6%. The residual plot of the PCR model shows a

Table 4
The 10 descriptors [19] with highest loading on PC1 and PLS factors

Descriptor	Definition	Descriptor class
ISH	Standardized information content on the leverage equality	GETAWAY descriptors
MEcc	Molecular eccentricity	Geometrical descriptors
BEHv1	Highest eigenvalue *n*. 1 of burden matrix/weighted by atomic van der Waals volumes	BCUT descriptors
BEHe1	Highest eigenvalue *n*. 1 of burden matrix/weighted by atomic Sanderson electronegativities	BCUT descriptors
BEHp1[a]	Highest eigenvalue *n*. 1 of burden matrix/weighted by atomic polarizabilities	BCUT descriptors
BELe6[b]	Lowest eigenvalue *n*. 6 of burden matrix/weighted by atomic Sanderson electronegativities	BCUT descriptors
FDI	Folding degree index	Geometrical descriptors
BIC4	Bond information content (neighborhood symmetry of 4-order)	Information indices
ATS1v	Broto–Moreau autocorrelation of a topological structure-lag 1/weighted by atomic van der Waals volumes	2D autocorrelations
PJI2	2D Petitjean shape index	Topological descriptors
Mor04m	3D-MoRSE-signal 04/weighted by atomic masses	3D-MoRSE descriptors

[a] No high loading on the PLS factors.
[b] No high loading on PC1.

trend in a smaller log(%HIA) interval, but larger residuals than the ones of the MLR models. Based on the R^2-, the RMSECV value and the residual plot, it can be concluded that the PCR model performs worse than the MLR models.

4.3. Partial least squares (PLS)

PLS was applied to the data set. The optimal PLS model, selected using a cross-validation sequence, consists of seven PLS latent variables. By observing the loadings of the different descriptors on the selected PLS factors, the variables with a higher impact on the model can be determined. Table 4 shows the 10 descriptors with highest loading on the PLS factors, their definition and class.

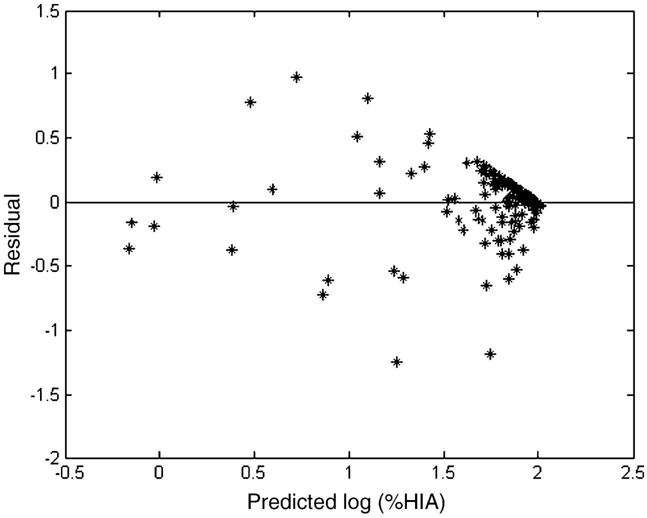

Fig. 3. Residual plot for the PCR-model.

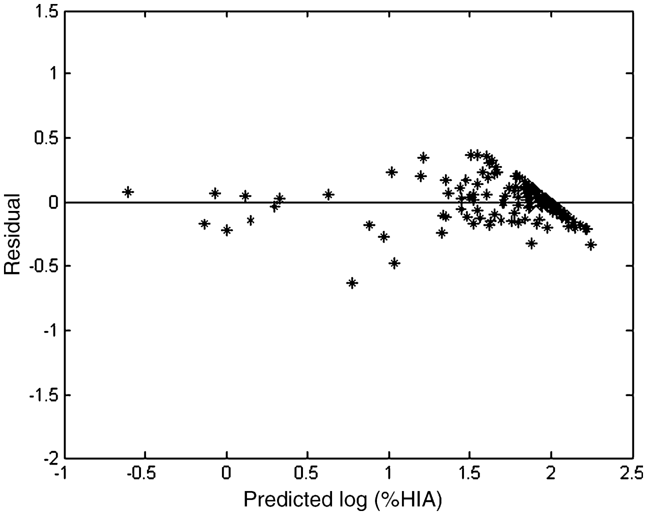

Fig. 4. Residual plot for the PLS-model.

Nine of the ten descriptors also belong to those with the highest loadings on the first principal component in the PCR analyses. Only one descriptor, BELe6, is different. This descriptor belongs to the class of the BCUT-descriptors and is related to the two-dimensional structure of the molecule [19].

Fig. 4 shows the residual plot for the PLS model. The coefficient of determination equals 0.879 and the RMSECV value is 0.185 or 11%. These values show that the PLS model fits the calibration data well and also has good predictive abilities. The residual plot of the PLS-model shows considerable smaller residuals compared to those for the MLR- and the PCR models. The trend in the higher part of the absorption range is observed in the log(%HIA) interval about 1.75–2.25. This is a range of 0.5 units and is almost half the range over which the trend is observed in the MLR-models. This is an indication for an improvement compared to the previous discussed models.

The dataset used contains a majority of the molecules in the higher parts of the absorption range. Several molecules have a %HIA equal to 100% and for an ideal model their residuals should be situated on a vertical line at log(%HIA) equal to 2. Therefore the smaller the interval over which a trend is observed in the above models, the better their predictive properties It could thus be concluded that, even if the residual plot still does not show a random distribution of the residuals, PLS gives a better model, with higher descriptive and predictive abilities compared to the MLR and PCR-models. This can be explained by the fact that PLS takes into account the co-variance of the descriptive variables with the response variable. This means that PLS is somehow capable of dealing with non-linearities in the data.

Table 5
The different base functions (B_m) of the MARS model and their coefficients (a_m) [11]

B_m	Definition	a_m
B_1	1	1.99
B_2	(188-T(O··O))$_+$	0.00210
B_3	(T(O··O)-188)$_+$	0.00240
B_4	(2.15-H4p)$_+$	−0.382
B_5	(−1.29-BLTF96)$_+$	0.0722
B_6	(BLTF96 + 1.29)$_+$	−0.145
B_7	(−0.698-Mor14v)$_+$	−2.63
B_8	(Mor14v + 0.698)$_+$	−0.230
B_9	(8,00-*n*HDon)$_+$	−0.0470
B_{10}	(*n*HDon-8,00)$_+$	−0.497
B_{11}	(35.6-RDF075m)$_+$	0.0292
B_{12}	(RDF075m-35.6)$_+$	0.0653
B_{13}	(−0.345-Mor18m)$_+$	1.24
B_{14}	(GATS4m-0.960)$_+$	0.899
B_{15}	(0.0650-HATS4p)$_+$	−118,4
B_{16}	(MAXDN-5.63)$_+$	−33.0
B_{17}	(R1e-2.20)$_+$	−3.15
B_{18}	(0.153-Mor17m)$_+$	−0.406
B_{19}	(Mor19v+0.221)$_+$	0.351
B_{20}	(0.169-Mor22m)$_+$	0.339
B_{21}	(Mor22m-0.169)$_+$	−1.05
B_{22}	(R3u-1.65)$_+$	−0.446
B_{23}	(ALOGP2-0.288)$_+$	−0.0123
B_{24}	(0.729-GATS1e)$_+$	−0.636
B_{25}	(GATS1e-0.729)$_+$	−0.782
B_{26}	(1.42-AAC)$_+$	1.65
B_{27}	(0.266-E1m)$_+$	3.77
B_{28}	(1,00-*n*Ct)$_+$	−0.175
B_{29}	(0.0930-Mor18m)$_+$	−0.977

4.4. Multivariate adaptive regression splines (MARS)

The MARS model used for comparison with previous models is the same as obtained in previous work [11]. Table 5 shows the different base functions of the MARS-model and Fig. 5 shows the residual plot for this model.

The obtained R^2 value is 0.933 [11] and the RMSECV, evaluated with 10-fold cross validation is 0.203 or 12.0%. It can be concluded that the MARS model gives similar results as the PLS-model, the R^2 and RMSECV values are very similar. When the residual plot of the MARS model is compared to the ones of the previous models, a high improvement can be seen towards the MLR- and PCR models. Compared to the residual plot of the PLS model, the residuals are slightly larger and the range over which the trend is observed is slightly broader. It can be concluded that the MARS model fits the data better, shows less under fitting, than the MLR and the PCR models described above. The MARS model has similar properties as the PLS model.

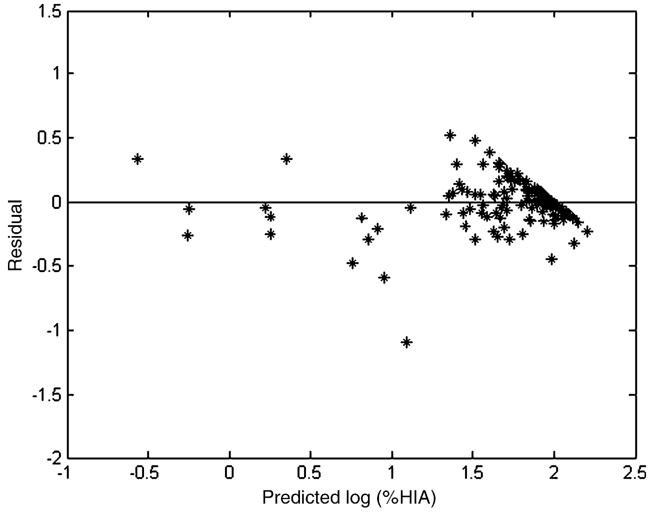

Fig. 5. Residual plot for the MARS-model.

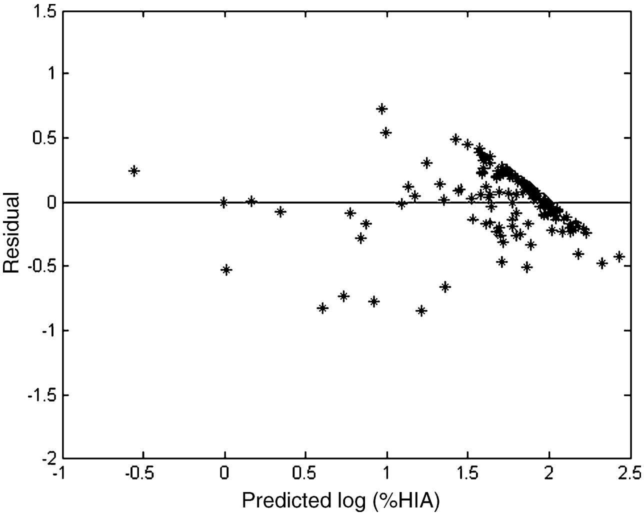

Fig. 6. Residual plot for the TMARS model.

4.5. Two-step MARS procedures

4.5.1. Stepwise MLR-MARS (TMARS)

The TMARS model for this data set was already built and evaluated in previous work [11]. The obtained model is given by following equation:

$$\begin{aligned}\hat{y} = {} & 1.88 - 0.0886(n\text{O}) - 0.0499(\text{T(S}\cdots\text{S)}) \\ & - 0.126(\text{Mor08m}) + 0.435(\text{Mor16v}) + 0.121(\text{HATS8v}) \\ & - 0.399(\text{C-030}) - 0.0010(337.6\text{-TIE})_+ \\ & - 0.0133((\text{TIE-337.6})(2.14\text{-R1e}))_+ \qquad (6)\end{aligned}$$

Fig. 6 shows the residual plot for the TMARS model. The obtained R^2 value is 0.720 [11] and the RMSECV, evaluated with 10-fold cross validation is 0.296 or 17.5%. Compared to the stepwise MLR model and the MARS model, the TMARS model performs slightly better than the MLR model, but worse than the MARS model. The residuals are smaller than in the stepwise MLR model. It can be concluded that the TMARS model fits the data better and shows less under fitting. The residual plots of both the MARS and the TMARS model are similar.

4.5.2. Genetic algorithm-MLR-MARS (GTMARS)

The MLR model obtained in Section 4.1.2 is used as starting point for the two-step MARS procedure. In a first step the

Table 6
Selected descriptors in the GTMARS model [19]

Descriptor	Definition	Descriptor class
*n*O	Number of oxygen atoms	Constitutional descriptors
GATS2p	Geary autocorrelation-lag 2/weighted by atomic polarizabilities	2D autocorrelation
EEig11x	Eigenvalue 11 from edge adjacent matrix weighted by edge degrees	Edge adjacency indices
RDF065m	Radial distribution function-6.5/weighted by atomic masses	RDF descriptors
H3m	H autocorrelation of lag 3/weighted by atomic masses	GETAWAY descriptors
C-003	CHR3	Atom-centered fragments
HATS4p	Leverage-weighted autocorrelation of lag 4/weighted by atomic polarizabilities	GETAWAY descriptors

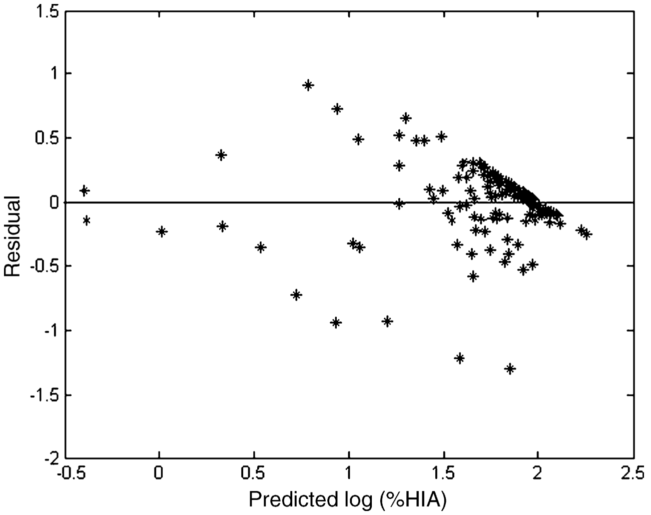

Fig. 7. Residual plot for the GTMARS model.

descriptors in the MLR model, which can be replaced by a pair of spline functions are selected. In the next step the stepwise addition procedure of pairs of spline functions is applied, resulting in the global GTMARS model. The global model is pruned using a sequence of general cross validations alternated with 10-fold cross validations. The optimal model is selected using Monte Carlo cross validation (MCCV) [11,13].

The selected model consists of nine terms, from which one is a constant, five are linear functions, one is a spline function of first order and two form a pair of spline functions of second order. The model is given by following equation:

$$\begin{aligned}\hat{y} = {} & 2.24 - 0.84(\text{GATS2p}) + 0.62(\text{EEig11x})\\ & + 0.73(\text{RDF065m}) - 1.88(\text{H3m}) + 0.27(\text{C-003})\\ & - 5.30(0.22\text{-}n\text{O}) + 73.01((0.22\text{-}n\text{O}) \cdot (\text{HATS4p-0.06}))_{+}\\ & + 352.37((0.22\text{-}n\text{O}) \cdot (0.06\text{-HATS4p}))_{+} \qquad (7)\end{aligned}$$

Table 6 shows the selected descriptors, their definition and class. Only one descriptor, HATS4p, is added by the TMARS algorithm. This descriptor belongs to the GETAWAY descriptors and can be related to the geometrical structure of the molecules [19]. All other descriptors were already selected in the original GA-MLR model.

Fig. 7 shows the residual plot of the GTMARS model. The R^2-value for this model is 0.549 and the RMSECV, evaluated with 10-fold cross-validation is 0.347 or 20.6%. Similar conclusions, as in previous section, can be made. The GTMARS model performs slightly better than the GA-MLR model, but worse than the MARS model. When residual plots are compared an improvement can be seen compared to the GA-MLR model. The residual plot of the GTMARS-model shows a similar trend, but has larger residuals compared to that of the MARS model.

4.5.3. *Principal Component Regression-MARS (PCR-MARS)*

The PCR-model obtained in Section 4.2 is used as starting point for the TMARS algorithm. Therefore the scores on the first eight principal components were added to the data matrix with descriptive variables. In a first step it was evaluated if some of the principal components in the PCR-model could be replaced by a pair of spline functions. In a next step pairs of splines were added stepwise, resulting in the global PCR-MARS model, followed by pruning and selection of the optimal model size using MCCV. The obtained model consists of five terms, one constant, two linear terms based on the scores on PC3 and PC5, respectively, and two form a pair of spline functions of second order:

$$\begin{aligned}\hat{y} = {} & 2.23 - 0.26(\text{PC3}) - 0.35(\text{PC5})\\ & - 38.05((0.17\text{-}n\text{O}) \cdot (\text{HATSv-0.48}))_{+}\\ & - 29.24((0.17\text{-}n\text{O}) \cdot (0.48\text{-HATSv}))_{+} \qquad (8)\end{aligned}$$

with *n*O, the number of oxygen atoms in the molecule and HATSv, the leverage-weighted total index/weighted by atomic van der Waals volumes [19]. HATSv belongs to the class of the GETAWAY-descriptors and can be related to the three-dimensional structure of the molecules.

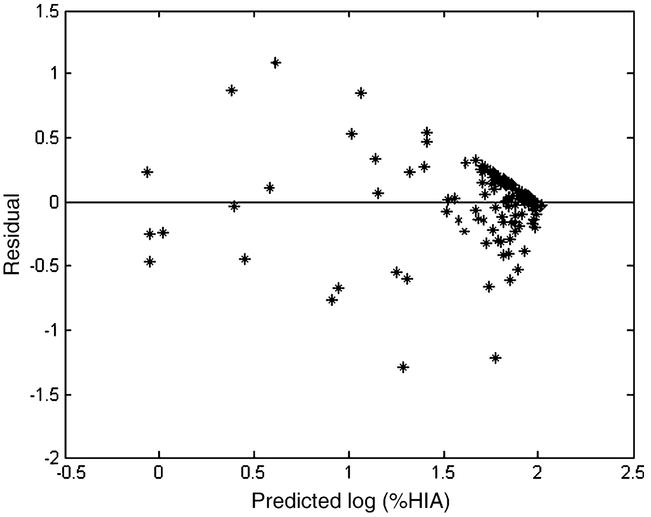

Fig. 8. Residual plot for the PCR-MARS-model.

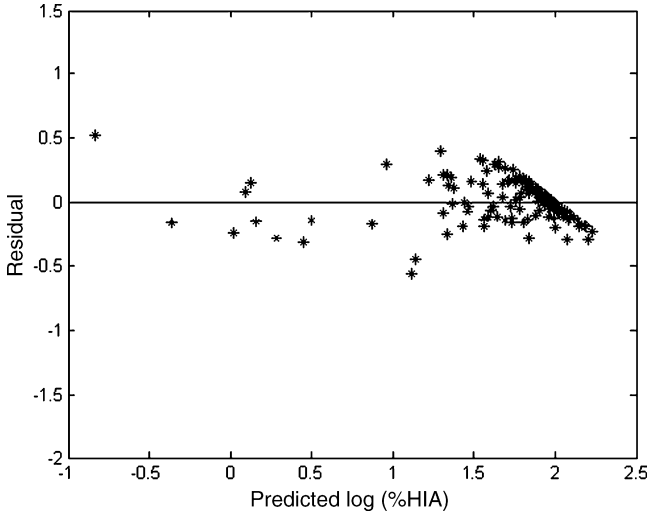

Fig. 9. Residual plot for the PLS-MARS model.

Fig. 8 shows the residual plot of the PCR-MARS model. The obtained R^2-value is 0.667 and the RMSECV is 0.307 or 18.2%. From these results the same conclusions can be made as in previous sections. The PCR-MARS model performs better than the PCR model, but worse than the MARS-model. When the residual plots are compared it can be seen that there is a high improvement compared to the PCR model. Comparing the residual plot to that of the MARS model shows that both residual plots show a similar trend, but residuals are smaller for the MARS model.

4.5.4. Partial Least Squares-MARS (PLS-MARS)

In analogy with PCR-MARS, the PLS model obtained in Section 4.3 was used as starting point for the application of the TMARS algorithm. The scores of the seven PLS factors selected for the PLS-model were added to the data set as descriptors. Again it was evaluated if some of the scores on the PLS-factors could be replaced by a pair of spline functions. In the following steps, pairs of splines were added stepwise resulting in the global PLS-MARS model, the global model was pruned and the optimal model size was selected using MCCV. The obtained model consists of nine terms, one constant, six linear terms based on the scores on respectively PLS-factors 1–6 (PLS1, PLS2, PLS3, PLS4, PLS5, PLS6), one spline function of first order based on the scores on PLS-factor 7 (PLS7) and a spline function of second order based on the scores on PLS7 and the values of the 3D-MoRSE descriptor, Mor15m (3D-MoRSE signal 15/weighted by atomic masses), a descriptor that can be related to the geometrical structure of the molecules [19]:

$$\begin{aligned}\hat{y} = &-2.52 + 0.60(\text{PLS1}) + 1.75(\text{PLS2}) + 1.27(\text{PLS3})\\ &+ 1.23(\text{PLS4}) + 0.81(\text{PLS5}) + 0.90(\text{PLS6})\\ &- 1.22(\text{PLS7} - 0.39)_+\\ &- 88.51((\text{PLS7} - 0.39)\cdot(0.82\text{-Mor15m}))_+ \end{aligned} \quad (9)$$

Fig. 9 shows the residual plot of the PLS-MARS model. The R^2-value is 0.903 and the RMSECV is 0.167 or 9.9%. From these results it can be seen that the predicted and observed log(%HIA) are highly correlated and that the prediction error evaluated with 10-fold cross validation is the lowest of all models presented. It can be concluded that the PLS-MARS model performs better than the individual PLS and the MARS models. The residual plots for respectivily the PLS-, the MARS- and the PLS-MARS models are very similar. A slight improvement compared to the residual plot of the MARS model can be observed. This model shows the smallest residuals of all models described in this paper. Still the trend in the higher part of the absorption range can still be observed.

4.6. Arcsine transformation

Since it is known that the classical transformation of percentage data is the arcsine transformation [13], this transformation was applied to the percentages HIA of the data set. The arcsine transformation is given as follows:

$$\text{Arcsin}(\%\text{HIA}) = \sin^{-1}(\%\text{HIA}/100) \quad (10)$$

All models described above were rebuilt using the arcsine transformed %HIA as response variable. In general, similar results were obtained as with the log transformed data. The RMSECV values were generally higher and the correlations lower than for the respective models build with the log transformed data. The best model obtained with the arcsine transformed data was again the PLS-MARS model. The selected model consists of nine terms, one constant, seven linear terms based on the scores on respectively PLS1, PLS2, PLS3, PLS4, PLS5, PLS6 and PLS7 and a spline function of second order based on the scores on PLS7 and the values of the 3D-MoRSE descriptor, Mor06v (3D-MoRSE signal 06/weighted by atomic van der Waals volumes), a descriptor related to the three-dimensional structure of the molecules [19]:

$$\begin{aligned}\hat{y} = &-3.06 + 0.59(\text{PLS1}) + 1.58(\text{PLS2}) + 1.17(\text{PLS3})\\ &+ 1.32(\text{PLS4}) + 0.67(\text{PLS5}) + 0.87(\text{PLS6}) - 0.62(\text{PLS7})\\ &- 3.77((\text{PLS7} - 0.30)\cdot(\text{Mor06v-0.45}))_+ \end{aligned} \quad (11)$$

Fig. 10 shows the residual plot for the PLS-MARS model for the arcsine transformed data. The obtained R^2-value is 0.633 and the RMSECV, evaluated with 10-fold cross validation is 0.312 or 20.9%. At first sight the residual plot seems to show a more random distribution of the residuals than the residual plot for the PLS-MARS model build with the log transformed data. A closer analysis shows that at the highest end of the absorption range a similar trend is seen as in the plot for the log transformed data. The higher concentration of objects in this region for the log transformed data is due to the fact that the log transformation maintains more the composition of the original data set. The data contains a high number of objects with high absorption values and a low number of objects with low absorption values. Fifty percent of the objects in the data set has an absorption value between 80 and 100% which explains the high concentration of objects in the high part of the absorption range in the residual plots for models build with the log transformed data. In the resid-

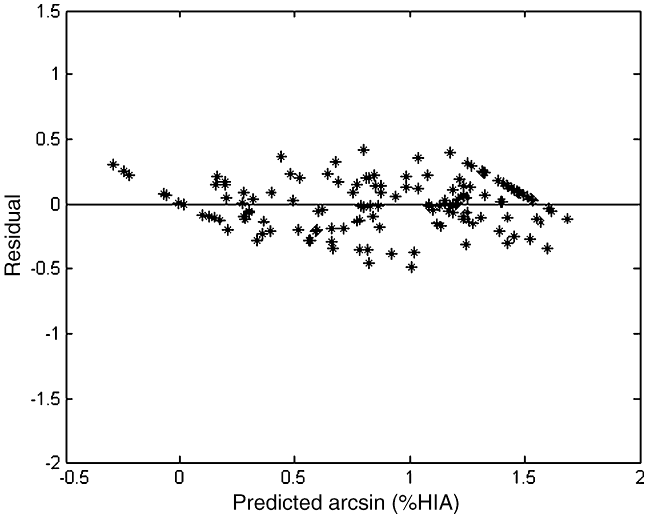

Fig. 10. Residual plot for the PLS-MARS model obtained with the arcsine transformed data.

ual plot of the model obtained with the arcsin transformed data, a trend can also be observed at the lowest end of the absorption range. Since an analogue trend is found in the residual plots of the models build with log and arcsine transformed data, since models build with log transformed data show higher linear correlations between predicted and observed values and given the better predictive properties of the latter models, it was decided to limit the discussion in this paper to the models obtained with log transformed data.

5. Conclusions

In general, it can be concluded that, for this data set, the combinations of linear modelling techniques with the non-linear technique MARS result in an improvement of the linear models. These improvements are clear for the descriptive properties of the models. In general the residual plots obtained with the combination techniques show smaller residuals compared to those from the linear techniques. Also the trend observed in the higher part of the absorption range is reduced to a more narrow interval. This points at a better distribution of the residuals, since for this dataset the ideal residual plot should contain a high number of points at the end of the absorption range, situated on a vertical line at log(%HIA) equal to 2. The smaller the interval over which the trend is seen at the end of the absorption range, the closer to the ideal situation and the better the model. In general, the combination techniques resulted in a better data fit with less under fitting. However, the RMSECV values of the combined models are only slightly better than those of the linear models, but because of the better data fit, predictions based on the combined models should be more robust and more reliable than those based on the linear models. The fact that the RMSECV values of the linear and combined models are similar, combined with the selection of the optimal models using eleven sequences of MCCV [11,23], which should protect the model from overfitting, indicates that the combination methods does not result in over fitted models. Still to be sure that the models are not overfitting, they should be tested with an external test set. The selection of a representative external test set was not possible here, due to the nature and the distribution of the data.

Table 7
Overview of the different models

Regression method	RMSECV	R^2	Number of terms
Stepwise MLR	0.331	0.616	13
GA-MLR	0.346	0.576	11
PCR	0.449	0.293	9
PLS	0.185	0.879	8
MARS	0.203	0.933	29
TMARS	0.296	0.720	9
GTMARS	0.347	0.549	9
PCR-MARS	0.307	0.667	5
PLS-MARS	0.167	0.903	9

Table 7 gives a summary of the model fit and the predictive properties for all presented models. In general, the models obtained with the combined methods perform very similar to the MARS-model. Only the PLS-MARS model performs better. The PLS-MARS model with a a cross-validation error of 9.9% and quite small residuals is the best model obtained for this data set.

To summarize it can be said that the observed improvement to a stepwise MLR model by including it in the two-step MARS procedure [11,12], can be generalized to other linear modelling techniques. In fact it was shown that the use of another variable selection technique for MLR, or the use of PCR and PLS as starting point for the two-step MARS procedure resulted in better models.

As already shown in previous work [11], it was confirmed that two-step MARS procedures with different linear techniques can be very valuable in QSAR.

Acknowledgements

The late Prof. D.L. Massart for providing the possibility to start the PhD study and for supporting the project. This research is financed with a specialization grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

[1] Y.H. Zhao, J. Le, M.H. Abraham, A. Hersey, P.J. Eddershaw, C.N. Luscombe, D. Boutina, G. Beck, B. Sherborne, I. Cooper, J.A. Platts, J. Pharm. Sci. 90 (2001) 749–784.
[2] O.A. Raevsky, K.J. Schaper, Eur. J. Med. Chem. 33 (1998) 799–807.
[3] J.A. Platts, M.H. Abraham, A. Hersey, D. Butina, Pharm. Res. 17 (2000) 1013–1018.
[4] S. Winiwarter, F. Ax, H. Lennernas, A. Hallberg, C. Pettersson, A. Karlen, J. Mol. Graph. Model 21 (2003) 273–287.
[5] C.A. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, P. Artursson, J. Med. Chem. 46 (2003) 558–570.
[6] T. Österberg, U. Norinder, J. Chem. Inf. Comput. Sci. 40 (2000) 1408–1411.
[7] U. Norinder, T. Österberg, P. Artursson, Eur. J. Pharm. Sci. 8 (1999) 49–56.
[8] S. Agatonovic-Kustrin, R. Beresford, A. Pausi, M. Yusof, J. Pharm. Biomed. Anal. 25 (2001) 227–237.

[9] E. Deconinck, T. Hancock, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 39 (2005) 91–103.

[10] J.P.F. Bai, A. Utis, G. Crippen, H. He, V. Fischer, R. Tullman, H. Yin, C. Hsu, L. Jiang, K. Hwang, J. Chem. Inf. Comput. Sci. 44 (2004) 2061–2069.

[11] E. Deconinck, Q.S. Xu, R. Put, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 39 (2005) 1021–1030.

[12] Q.S. Xu, D.L. Massart, Y.Z. Liang, K.T. Fang, J. Chromatogr. A 998 (2003) 155–167.

[13] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier Science, Amsterdam, 1997.

[14] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science, Amsterdam, 1997.

[15] D. Jouan-Rimbaud, R. Leardi, O.E. De Noord, D.L. Massart, Anal. Chem. 67 (1995) 4295–4301.

[16] J.H. Friedman, Ann. Stat. 19 (1991) 1–141.

[17] S. Sekulic, B.R. Kowalski, J. Chemometrics 6 (1992) 199–216.

[18] R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1055 (2004) 11–19.

[19] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley–UCH, Weinheim, 2000.

[20] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon® Professional version, software version 5.0, Milano Chemometrics and QSAR research group, copyright Talete srl© (1997-2004), 2004.

[21] M.H. Abraham, A. Ibrahim, A.M. Zissimos, Y.H. Zhao, J. Corner, D.P. Reynolds, Drug Discov. Today 7 (2002) 1056–1063.

[22] M.H. Abraham, H.S. Chadha, R.A.E. Leitao, R.C. Mitchell, W.J. Lambert, R. Kaliszan, A. Nasal, P. Haber, J. Chromatogr. A 766 (1997) 35–47.

[23] Q.S. Xu, Y.Z. Liang, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.