Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Pharmaceutical and Biomedical Analysis 39 (2005) 1021–1030
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Prediction of gastro-intestinal absorption using multivariate adaptive regression splines

E. Deconinck [a], Q.S. Xu [a], R. Put [a], D. Coomans [b], D.L. Massart [a], Y. Vander Heyden [a,*]

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

Received 15 April 2005; received in revised form 30 May 2005; accepted 30 May 2005

Abstract

Multivariate adaptive regression splines (MARS) and a derived method two-step MARS (TMARS) were used for modelling the gastro-intestinal absorption of 140 drug-like molecules. The published absorption values for these molecules were used as response variable and calculated molecular descriptors as potential explanatory variables. Both methods were compared and their potential use in quantitative structure–activity relationship (QSAR) context evaluated.

The predictive abilities of the models were studied using different sequences of Monte Carlo cross validation (MCCV). It was shown that both types of models had good predictive abilities and that for the data used, MARS gave better results than TMARS. It could be concluded that both methods could be valuable for QSAR modelling.

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

1. Introduction

High throughput screening has become a very important issue in drug discovery. Since most new molecules, potentially useful, fail in a later phase of the drug development due to non-proper absorption, distribution, metabolisation, elimination and toxicity (ADME-Tox) properties, screening methods for these properties are necessary in the first stages of the drug development. In silico screening can be very useful, since it allows screening for ADME-Tox and other properties before the molecules are even synthesized. In silico methods try to build relationships between a dataset consisting of known values for the property of interest and some calculated theoretical and/or experimental parameters or descriptors. These kind of relationships are called quantitative structure–activity relationships (QSAR). This paper focuses on the relationships between theoretical descriptors and the gastro-intestinal absorption of drug molecules.

In the literature different QSAR-models can be found predicting the absorption of molecules, and built using linear modelling techniques like multiple linear regression (MLR) [1], principal components regression (PCR) [2], partial least squares (PLS) regression [2,3], and some more advanced non-linear techniques like artificial neural networks (ANN) [4] and classification and regression trees (CART) [5]. Two well known approaches used in screening are the Lipinski rule of five [6] and the linear free energy relationship (LFER) approach of Abraham et al. [7]. A disadvantage of these two methods is that they give a quite rough classification of the molecules, allowing the elimination of only a very limited set of molecules.

In this paper, it was tried to build models, that give a more accurate prediction of the absorption values of drug molecules. Therefore two techniques, multivariate adaptive regression splines (MARS) and two-step MARS (TMARS),

* 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 © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2005.05.034

were evaluated. The latter is in fact a combination of MLR and MARS [8]. The MARS technique was introduced by Friedman in 1991 [9] and successfully used in QSAR by Nguyen-Cong et al. [10] and Ren et al. [11,12] and in quantitative structure retention relationships (QSRR) by Put et al. [13]. TMARS was introduced and applied successfully in the prediction of retention in gas chromatography by Xu et al. [8]. It was proven that the combined method TMARS significantly improved the prediction abilities compared to the individual MLR and MARS models.

In a first step, absorption was modeled using MARS. The models were evaluated for their predictive abilities using Monte Carlo cross validation (MCCV) [14]. In a second step, a TMARS model was build, evaluated and compared to the MARS-models.

2. Theory

2.1. Multivariate adaptive regression splines (MARS)

MARS is a local modelling technique that divides the data space into several, possibly overlapping, regions and fits truncated spline functions in each of these regions. Truncated spline functions consist of two segments, i.e. left-sided Eq. (1) and right-sided Eq. (2) truncated functions, separated from each other by a so-called knot location [9].

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

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

where $b_q^-(x-t)$ and $b_q^+(x-t)$ are the spline functions describing, respectively, the regions right and left of the knot location t 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. A spline function is also called a base function. For each of the explanatory variables MARS selects the pair of splines and the knot location, that best describe the response variable. In a next step, the different base functions are combined in one multidimensional model, which describes the response as a function of the explanatory variables. The result is a complex non-linear model of the form:

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

where $\hat{y}$ is the predicted value for the response variable; a_0, the coefficient of the constant base function; M, the number of base functions and B_m and a_m the mth base function and its coefficient [9,15].

A MARS analysis generally consists of three steps. In the first step the variable for which the pair of spline functions gives the best description of the response is selected. After this selection new spline functions are added stepwise in order to eventually get a complex multivariate model – the global MARS-model – which almost perfectly describes the training set. The stepwise addition of spline functions is based on the improvement of the model. In each step the pair of splines, which gives the best improvement in the description of the training set, is added. The global MARS model usually shows overfitting. In a next step the global MARS-model is pruned using a sequence of generalised cross-validations (GCV) alternated with 10-fold cross-validation. During this procedure, the contribution of each base function to the descriptive abilities of the model is evaluated based on a lack-of-fit (LOF) criterion. The base functions contributing the least to the model are eliminated stepwise. This pruning process results in a sequence of models with different size. In the third and final step the optimal model is selected using a cross-validation technique [9,15].

2.1.1. Building the global MARS-model

In the first step, the MARS-algorithm divides the data space into two subregions. This is done searching iteratively each of the descriptive variables as split variable and for each variable each available data point as knot location. These selections are done using the GCV-statistic:

$$\mathrm{GCV}(M) = \left(\frac{1}{n}\right) \frac{\sum_{m=1}^{M}(y_i - \hat{y}_i)^2}{(1 - C(M)/n)^2} \quad (4)$$

where n is the number of data objects; y_i, the response value for object I; $\hat{y}_i$, the predicted response value for object I; and $C(M)$, a penalty factor defined as

$$C(M) = M + \mathrm{d}M \quad (5)$$

where M is the number of non-constant base functions in the model and d a cost penalty factor for each base function optimisation. The GCV-statistic is first used to select the best knot location for each of the descriptive variables. Then the same statistic is used to select the most significant variable (and his previously selected knot location) for description of the training set. After selection of the variable the data space is divided into two subregions, defined by two splines (one at the left of the knot-location and one at the right). The procedure is now repeated for each of the subregions and then for the subregions of the subregions, and so on. This iterative procedure results in a two-by-two stepwise addition of spline base functions and is continued until a model is build with a predefined number of terms. This model is called the global MARS-model [9,15].

A spline base function can be either a single spline function or an interaction term consisting of the product of two or more spline functions. The level of interaction terms allowed is determined by the order q in MARS. If q equals 1, only single (linear) spline functions are allowed. If q equals 2 or 3, respectively, quadratic and cubic interaction terms can be added. When interaction terms are allowed the algorithm

checks, at the end of each iteration, whether the introduction of an interaction improves the model [9,15].

2.1.2. Pruning

The global MARS-model usually shows overfitting. Therefore a stepwise pruning procedure is applied, sequentially eliminating the least contributing base function(s). Usually the pruning process is based on GCV, but other cross-validation methods, like *n*-fold cross-validation can be used. GCV results in the GCV-statistic. In fact this statistic is the error sum of squares adjusted with a penalty factor for the complexity of the model. This is done to avoid the selection of too complex, overfitting models. In this work pruning was carried out by applying alternately GCV and 10-fold cross-validation. This alternating pruning process results in a series of smaller MARS-models [9,15].

2.1.3. Selection of the optimal model

The selection of the optimal model out of the series of models obtained from the pruning process is also based on cross-validation. Usually leave-one-out cross validation is used for this purpose, but in fact every cross validation technique can be used. The model with the lowest root mean square error of cross validation (RMSECV) is the most accurate model. [9,15]. The optimal model can be identified as the least complex one within one standard error of the most accurate model. The idea is here to choose the least complex model with a predictive error comparable to that of the most accurate one [16]. In this paper the selection of the most optimal model was carried out using Monte Carlo cross validation [14].

2.2. Two-step MARS (TMARS)

TMARS is in fact a combination of MLR and MARS. In a first step a linear model is build, which describes the response variable as a function of the explanatory variables, using a stepwise linear regression procedure. If the obtained linear model shows lack of fit, MARS is applied based on the linear model. During this procedure some of the descriptors used in the linear model are replaced by a pair of spline functions. In the next step the two-by-two stepwise addition procedure for building the global model is applied, resulting in a global TMARS model. After obtaining the global model, pruning and selection of the optimal model is carried out as described in Sections 2.1.2 and 2.1.3, respectively [8].

The TMARS procedure starts with building a linear model, which describes the response variable as a function of the explanatory variables. A model is obtained

$$\hat{y} = a_0 + \sum_{l=1}^{L} a_l x_l \tag{6}$$

in which a_0 is the intercept; L, the number of selected variables and a_l, the regression coefficient of variable x_l. The forward stepwise algorithm is used to select the best descriptors to be included in the model. This procedure starts with the variable, which shows the highest correlation with the response. If this variable results in a significant regression by F-test, the variable is retained and the stepwise procedure continues. At each step, the variable that gives the highest decrease in the error sum of squares is added to the model. The model building stops when none of the remaining variables causes a significant decrease in the sum of squares [8,17].

A test for lack of fit is carried out on the linear model [8,17]. If the multiple regression coefficient R is close to one and the F-ratio is not significant, the linear model can be considered as final. If not, the model building is continued applying the TMARS algorithm.

In a first step a forward stepwise procedure is used to determine whether some variables x_l in the linear model should be replaced by a pair of spline functions, resulting in a model

$$\hat{y} = c_0 + \sum_{k=1}^{K} c_k g_k(x) \tag{6}$$

where c_0 is the constant base function; K, the number of base functions derived from the linear model; c_k, the coefficient of the function $g_k(x)$, with $g_k(x)$ either one of the descriptors x_l or a pair of spline function $\lfloor \pm(x_l - x_{jl}) \rfloor$ [8].

In a next step, pairs of spline functions are added to the model, following the same procedures as described in Section 2.1.1, resulting in a combined model:

$$\hat{y} = c_0 + \sum_{k=1}^{K} c_k g_k(x) + \sum_{m=1}^{M} a_m B_m(x) \tag{7}$$

where M is the number of MARS base functions; a_{m}, the coefficient of the mth MARS base function $B_m(x)$.

This model, called the global TMARS model, is pruned according to the procedure described in Section 2.1.2. Both $g_k(x)$ and $B_m(x)$-functions can be deleted in the pruning process. The pruning process results in a series of smaller models from which the optimal is selected using cross-validation (Section 2.1.3) [8,9].

2.3. 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 [18]. Several thousands of descriptors are already proposed in the literature and the number is still growing. Theoretical descriptors can be classified in different ways. The most applied classification is 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, respectively, from a molecular formula, a substructure list, a topological, a geometrical and a stereoelectronic or lattice representation. More information about molecular descriptors and their classification can be found in the work of Todeschini and Consonni [18].

3. Materials and methods

3.1. Data

The data consists of intestinal absorption values for a subset of 140 molecules extracted from a dataset collected by Zhao et al. [1]. For each of the molecules the name and the percentage intestinal absorption (%HIA) are listed in Table 1. These molecules were selected because they show a high diversity in molecular structure and cover the whole absorption range (0–100%) [5].

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, defining the structure. This data matrix can then be used to calculate molecular descriptors [5].

3.3. Calculating molecular descriptors

Molecular descriptors were calculated using the Dragon® 4.0 professional software [19]. 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 and Consonni [18]. The software automatically eliminates constant variables in a given dataset. For descriptors with a correlation higher than 0.98, parameters are set such that only one is retained in the dataset. Next to the Dragon® descriptors Hyperchem® was used to calculate some additional parameters, i.e. solvent accessible surface area, molecular volume, octanol/water partition coefficient (log P), hydration energy, molar refractivity, molar polarisability and molar mass [5].

3.4. Building MARS and TMARS models

The models were build using in-house algorithms written in Matlab 6.5 (The Mathworks, Matick, MA). Programming was done according to the original MARS algorithm proposed by Friedman [5]. The absorption values were used as

Table 1
The absorption data for the 140 molecules, extracted from Zhao et al. [1,5]

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
20	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

Table 1 (*Continued*)

No.	Substance	%HIA
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
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

response variables and the descriptors as explanatory variables.

4. Results and discussion

4.1. Building MARS-models

The model was build using the Briggsian logarithms of the percentages human intestinal absorption (%HIA) of all 140 molecules as response variable. The descriptors were used as descriptive variables. The global MARS-model is build and pruned. The order q of the MARS-model is set on 2, which means that both linear and second order splines can be used during model building. The maximum number of terms $M_{\max}$, the stop criterion in building the global MARS-model, was set to 100. Pruning was carried out using alternately 10-fold and general cross validation. A sequence of smaller MARS models was obtained. Selection of the optimal model was performed using Monte Carlo cross validation [14]. In MCCV, a given fraction of the dataset is used as test set. The process starts with a random selection of the test set, the remaining objects are used as training set. The obtained model is used to predict the test set and the error is calculated. In our work this process is repeated one hundred times, each time with a new randomly selected test set. The mean error is calculated. Eleven sequences of MCCV were carried out. The first sequence used only one object as test set and corresponds to a leave one out cross-validation. The other sequences used, respectively, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50% of the dataset as test set. For each size of test sets 100 repetitions are carried out. Fig. 1 shows the root mean square error of cross validation as a function of the complexity of the models. Each line stands for one sequence. The model showing a minimal RMSECV for all 11 sequences is selected as the optimal. Fig. 2 shows that the model with 31 base functions shows the minimal RMSECV. Based on the one standard error rule the model with 29 base functions was selected as the optimal. The different base functions and their coefficients are given in Table 2. The model consist of one constant function B_1 and 28 single lin-

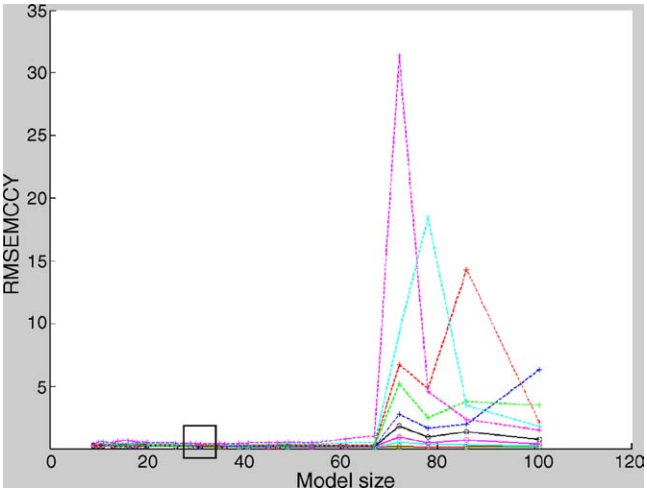

Fig. 1. RMSECV as a function of the MARS model size. Different lines represent different testset sizes.

ear spline functions. No second order splines were selected (Fig. 2).

Out of the 761 descriptors used to build the MARS-model, 20 different descriptors were selected. Fourteen of the selected base functions can be considered as seven pairs of spline functions: B_2 and B_3, B_5 and B_6, B_7 and B_8, B_9 and B_{10}, B_{11} and B_{12}, B_{20} and B_{21} and B_{24} and B_{25}. Take the example of pair B_2 and B_3:

$$(188 - T(\mathrm{O}\cdot\cdot\mathrm{O}))_+ = \begin{cases} 188 - T(\mathrm{O}\cdot\cdot\mathrm{O}), & \text{if } T(\mathrm{O}\cdot\cdot\mathrm{O}) > 188 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

$$(T(\mathrm{O}\cdot\cdot\mathrm{O}) - 188)_+ = \begin{cases} T(\mathrm{O}\cdot\cdot\mathrm{O}) - 188, & \text{if } T(\mathrm{O}\cdot\cdot\mathrm{O}) < 188 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

This means that when $T(\mathrm{O}\cdots\mathrm{O})$ is higher than 188, the second term (Table 2) in Eq. (3) equals $0.0021(188 - T(\mathrm{O}\cdot\cdot\mathrm{O}))$ and the third term is zero. When $T(\mathrm{O}\cdot\cdot\mathrm{O})$ is smaller than 188, the second term is zero and the third equals $0.0024(T(\mathrm{O}\cdot\cdot\mathrm{O}) - 188)$. The remaining terms are not paired. As example,

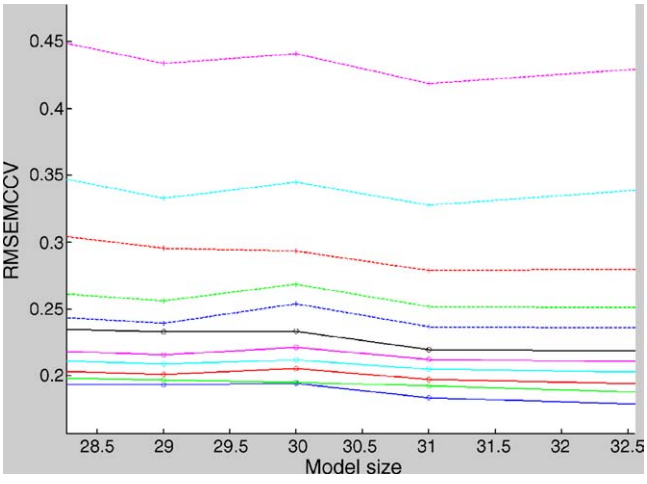

Fig. 2. Magnified section from Fig. 1, RMSECV as a function of the MARS model size. Different lines represent different testset sizes.

consider function B_4:

$$(2.1490 - \mathrm{H4p})_+ = \begin{cases} 2.1490 - \mathrm{H4p}, & \text{if H4p} > 2.1490 \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

what means that the fourth term in Eq. (3) equals $-0.3825(2.1490 - \mathrm{H4p})$, when H4p is higher than 2.1490 and zero when it is smaller than 2.1490.

The summation of the 29 base functions gives a response surface in multidimensional space consisting of small planes describing local regions of the dataspace.

4.2. The selected descriptors

Like previously mentioned 20 descriptors were selected in the model. Table 3 shows the different selected descriptors, their definition and their class. More information can be found in the work of Todeschini and Consonni [18]. Since these descriptors are calculated theoretical values, it is not evident to relate them physicochemically to the process of drug absorption. One exception is the selection of *n*-octanol/water partition coefficient (log *P*)-based descriptors (BLTF96 and ALOGP2). log *P* is one of the key properties of a molecule, often used to estimate whether that molecule can pass a biological membrane or not [6,20,21]. Two descriptors, $T(\mathrm{O}\cdots\mathrm{O})$ and *n*HDon are related to respectively the oxygen

Table 2
The different base functions (B_m) of the model and their coefficients (a_m)

B_m	Definition	a_m
B_1	1	1.9917
B_2	$(188\text{-}T(\mathrm{O}\cdot\cdot\mathrm{O}))_+$	0.0021
B_3	$(T(\mathrm{O}\cdot\cdot\mathrm{O}) - 188)_+$	0.0024
B_4	$(2.1490 - \mathrm{H4p})_+$	−0.3825
B_5	$(-1.29 - \mathrm{BLTF96})_+$	0.0722
B_6	$(\mathrm{BLTF96} + 1.29)_+$	−0.1450
B_7	$(-0.6980 - \mathrm{Mor14v})_+$	−2.6274
B_8	$(\mathrm{Mor14v} + 0.6980)_+$	−0.2297
B_9	$(8 - n\mathrm{HDon})_+$	−0.0470
B_{10}	$(n\mathrm{HDon} - 8)_+$	−0.4967
B_{11}	$(35.602 - \mathrm{RDF075}m)_+$	0.0292
B_{12}	$(\mathrm{RDF075}m - 35.602)_+$	0.0653
B_{13}	$(-0.3450 - \mathrm{Mor18}m)_+$	1.2407
B_{14}	$(\mathrm{GATS4}m - 0.9600)_+$	0.8989
B_{15}	$(0.0650\text{-HATS4p})_+$	−1184274
B_{16}	$(\mathrm{MAXDN} - 5.6270)_+$	−33.0517
B_{17}	$(\mathrm{R1e} - 2.2010)_+$	−3.1528
B_{18}	$(0.1530 - \mathrm{Mor17}m)_+$	−0.4062
B_{19}	$(\mathrm{Mor19v} + 0.2210)_+$	0.3513
B_{20}	$(0.1690 - \mathrm{Mor22}m)_+$	0.3389
B_{21}	$(\mathrm{Mor22}m - 0.1690)_+$	−1.0480
B_{22}	$(\mathrm{R3u} - 1.6460)_+$	−0.4457
B_{23}	$(\mathrm{ALOGP2}\text{-}0.2880)_+$	−0.0123
B_{24}	$(0.7290 - \mathrm{GATS1e})_+$	−0.6363
B_{25}	$(\mathrm{GATS1e} - 0.7290)_+$	−0.7825
B_{26}	$(1.4250 - \mathrm{AAC})_+$	1.6485
B_{27}	$(0.2660 - \mathrm{E1}m)_+$	3.7735
B_{28}	$(1 - n\mathrm{Ct})_+$	−0.1747
B_{29}	$(0.0930 - \mathrm{Mor18}m)_+$	−0.9773

Table 3
Selected descriptors in the MARS-model [18]

Descriptor	Definition	Descriptor class
T(O · ·O)	Sum of topological distances between oxygen atoms	Topological descriptors
H4p	H autocorrelation of lag 4/weighted by atomic polarizabilities	GETAWAY descriptors
BLTF96	Verhaar model of Fish base-line toxicity from MLOGP (mmol/l)	Molecular properties
Mor14v	3D-MoRSE-signal 14/weighted by atomic van der Waals volumes	3D-MoRSE descriptors
*n*HDon	Number of donor atoms for H-bonds (with N and O)	Functional group counts
RDF075*m*	Radial distribution function-7.5/weighted by atomic masses	RDF descriptors
Mor18*m*	3D-MoRSE signal 18/weighted by atomic masses	3D-MoRSE descriptors
GATS4*m*	Geary autocorrelation-lag 4/weighted by atomic masses	2D autocorrelations
HATS4p	Leverage-weighted autocorrelation of lag 4/weighted by atomic polarizabilities	GETAWAY descriptors
MAXDN	Maximal electrotopological negative variation	Topological descriptors
R1e	R-autocorrelation of lag1/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
Mor17*m*	3D-MoRSE signal 17/weighted by atomic masses	3D-MoRSE descriptors
Mor19v	3D-MoRSE-signal 19/weighted by atomic van der Waals volumes	3D-MoRSE descriptors
Mor22*m*	3D-MoRSE signal 228/weighted by atomic masses	3D-MoRSE descriptors
R3u	R autocorrelation of lag 3/unweighted	GETAWAY descriptors
ALOGP2	Squared Ghose-Crippen octanol–water partition coefficient (log P_2)	Molecular properties
GATS1e	Geary autocorrelation-lag 1/weighted by atomic Sanderson electronegativities	2D autocorrelations
AAC	Mean information index on atomic composition	Information indices
E1*m*	1st component accessibility directional WHIM index/weighted by atomic masses	WHIM descriptors
*n*Ct	Number of total tertiary C(sp3)	Functional group counts

atoms and the donor atoms for H-bonding in the molecule. These descriptors describe properties related to 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]. The other selected descriptors can be related to the two-dimensional (GATS4*m*, MAXDN, GATS1e, AAC and *n*Ct) or three-dimensional (H4p, Mor14v, RDF075*m*, Mor18*m*, HATS4p, R1e, Mor17*m*, Mor19v, Mor22*m*, R3u and E1*m*) structure of the molecule [18].

4.3. Predictive power

To evaluate the predictive properties of the model obtained in Section 4.1 each molecule was predicted twice. Once as part of the training set and once as part of the different test sets used in the different sequences of the Monte Carlo cross validation. The model describes the training set quite well since the root mean squared error for the training set is 0.1183 or 7.0% and the R value is 0.9330. The mean value for the Root Mean Squared Error of Cross Validation (RMSECV) obtained for the different sequences of the Monte Carlo cross validation is 0.2594. This value can be seen as a mean error of 15.4% of the real absorption value (%HIA). Fig. 3 shows the residual plot for the selected MARS-model. In the major part of the absorption range an acceptable distribution of the residuals is obtained. In the highest part of the range, where most objects are situated, a trend is seen. This indicates that the model is not ideal. Since the range over which this trend is shown is small, one can conclude that predictions based on the MARS-model give a quite accurate indication about the absorption values of the predicted molecules.

4.4. Building the TMARS-model

The model was build using the Briggsian logarithms of the %HIA-values for the 140 molecules as response variable. The descriptors were used as descriptive variables. First the linear model was build and evaluated with a lack-of-fit test as described in Section 2.2. Since this test was significant MARS was applied. The global TMARS-model was build and pruned. The order q of the TMARS-model was set on 2 and the maximum number of terms M_{max} 100. Pruning was carried out using alternately 10-fold and general cross validation. A sequence of smaller TMARS models was obtained. The selection of the optimal model was carried out using Monte Carlo cross validation like before.

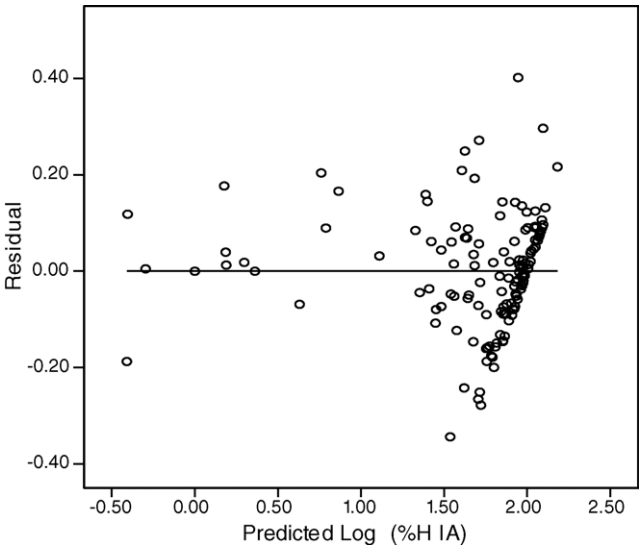

Fig. 3. Residual plot for MARS.

Table 4
Selected descriptors in the MLR-model [18]

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 sulfur atoms	Topological descriptors
IC2	Information content index (neighborhood symmetry of 2-order)	Information indices
Mor08*m*	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-centred fragments

The obtained linear model consists of 13 terms, with one constant function and 12 functions based on different descriptors. The linear model is given by following equation:

$$\begin{aligned}\hat{y} = {} & 1.2576 - 0.1008(n\mathrm{O}) - 0.0006(\mathrm{TIE}) \\ & - 0.0018\,(D/\mathrm{Dr05}) - 0.0497(T(\mathrm{S}\cdot\cdot\mathrm{S})) \\ & + 0.1760(\mathrm{IC2}) - 0.1239(\mathrm{Mor08}m) + 0.2937(\mathrm{Mor16v}) \\ & + 0.2496(\mathrm{HATS8v}) - 0.5254(n\mathrm{N} = N) \\ & + 0.5056(n\mathrm{N(CO)_2}) - 0.0814\,(n\mathrm{OH}) \\ & - 0.3987(\mathrm{C\text{-}030}) \end{aligned} \quad (11)$$

The multiple correlation coefficient *R* equals 0.701, the standard error *S* 0.305 and the *F*-ratio for overall regression fit F 24.782. Table 4 shows the selected descriptors, their definition and class. Fig. 4 shows the residual plot for the linear model. Based on the *R* and the *F* values it can be concluded that the predicted and observed log(%HIA) are not highly linearly correlated. The residual plot shows that there is no random distribution of the residuals, which can imply that the model shows underfitting. The root mean square error of cross validation for this model is 27.01, evaluated with leave-one-out cross-validation, which confirms that the model is not suited for prediction. Due to these findings the second step, MARS, was applied. After obtaining the global TMARS-model, it was pruned using alternately 10-fold and general cross-validation. Out of the series of models the optimal was selected using Monte Carlo cross validation (see Fig. 5). The graph shows that the minimum for all sequences is found at model size 9. The optimal model consists of nine terms in which seven linear terms and two spline functions (pair of splines) are integrated. The final model can be written as follows:

$$\begin{aligned}\hat{y} = {} & 1.8814 - 0.0886(n\mathrm{O}) - 0.0499(T(\mathrm{S}\cdot\cdot\mathrm{S})) \\ & - 0.1256(\mathrm{Mor08}m) + 0.4351(\mathrm{Mor16v}) \\ & + 0.1207(\mathrm{HATS8v}) - 0.3988(\mathrm{C\text{-}030}) \\ & - 0.0010(337.554 - \mathrm{TIE})_+ \\ & - 0.0133((\mathrm{TIE} - 337.554)(2.14 - \mathrm{R1e}))_+ \end{aligned} \quad (12)$$

This model shows an *R* value of 0.7200 and a RMSECV of 0.2927 evaluated with leave-one-out cross validation. Both

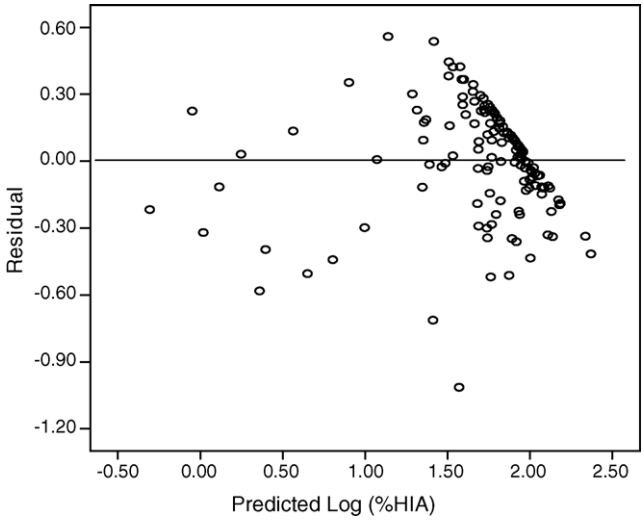

Fig. 4. Residual plot for the MLR model.

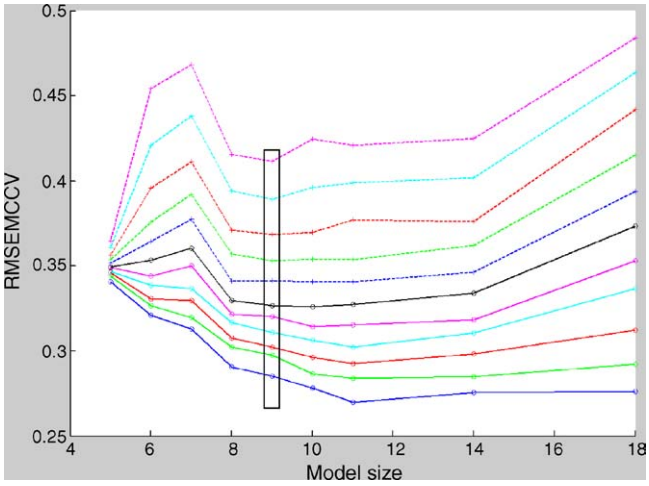

Fig. 5. RMSECV as a function of the TMARS model size. Different lines represent different testset sizes.

Table 5
Selected descriptors in the TMARS-model [18]

Descriptor	Definition	Descriptor class
*n*O	Number of oxygen atoms	Constitutional descriptors
T(S · ·S)	Sum of topological distances between sulfur atoms	Topological descriptors
Mor08*m*	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
C-030	X-CH-X	Atom-centred fragments
TIE	E-state topological parameter	Topological descriptors
R1e	R-autocorrelation of lag1/weighted by atomic Sanderson electronegativities	GETAWAY descriptors

values show that TMARS resulted in an improvement of the linear model.

4.5. The selected descriptors

Eight descriptors were selected in the final TMARS-model. Table 5 shows the different selected descriptors, their definition and their class. More information about these descriptors can be found in the work of Todeschini and Consonni [18]. All descriptors used are theoretical and can not easily be related to the process of membrane passage. In the selection two descriptors are found corresponding to oxygen and sulphur atoms (*n*O and *T*(S · ·S)). The properties described by these descriptors can again be related to the polar surface area (PSA) [22]. Most of the other descriptors can be related to the two-dimensional (TIE) or three-dimensional (Mor08*m*, Mor16v, HATS8v and R1e) structure of the molecule. 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 [18].

4.6. Predictive power of the TMARS-model

The predictive abilities of the model were evaluated similarly as the MARS-model in Section 4.3. As for the MARS-model each molecule was predicted twice. Once as part of the training set and once as part of the different test sets of the MCCV. The description of the training set by the model is acceptable with an *R* value of 0.7200 and a root mean squared error for the training set of 0.2772 or 16.44%. The mean value for the RMSECV evaluated with Monte Carlo cross validation is 0.3377 and can be considered as a mean error of 20.03% of the real absorption value. The errors obtained with the TMARS-model are larger than those obtained with the MARS-model. Fig. 6 shows the residual plot for the selected TMARS-model. Comparison of Figs. 6 and 3 shows that with TMARS a less good distribution of the residuals is obtained than for MARS. An analogue trend is found as for the MARS-model, but the range over which this trend is observed is larger than for MARS. Further, it can be observed that the residuals with TMARS are considerably larger. This led to the conclusion that predictions based on the TMARS-model, are less reliable than those based on the MARS-model.

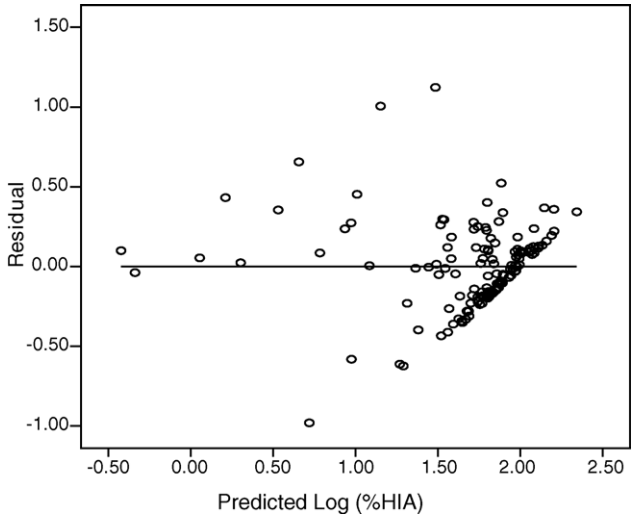

Fig. 6. Residual plot for TMARS.

5. Conclusions

Comparison of the MARS and the TMARS model shows that the MARS-model describes the dataset better and has a better predictive ability. The lower performance of the TMARS method can be explained by the fact that the TMARS model is based on a linear model. The obtained TMARS model shows high similarity with the linear model. Seven of the selected descriptors (*n*O, *T*(S · ·S), Mor08*m*, Mor16v, HATS8v, C-030 and TIE) correspond to descriptors from the linear model. Only one descriptor (R1e) corresponds to a descriptor selected in the MARS model. This leads to the conclusion that for our data the linear component in TMARS is too strongly represented, resulting in worse models and predictions than those based on the complete non-linear MARS technique.

From the results in this paper it can be concluded that for this dataset, MARS performs better than TMARS and that MARS can be a valuable tool in modelling the gastro-intestinal absorption of molecules.

It could also be shown that TMARS can dramatically improve an MLR model for gastro-intestinal absorption. Therefore it can also be supposed that TMARS is valuable in modelling gastro-intestinal absorption, if a dataset is avail-

able with a higher linear correlation between the %HIA and the descriptors, resulting in a better MLR-model.

As final conclusion it can be stated that both techniques can be valuable and deserve more attention in the field of QSAR.

Acknowledgment

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] S. Winiwarter, F. Ax, H. Lennernas, A. Hallberg, C. Pettersson, A. Karlen, J. Mol. Graph. Model. 21 (2003) 273–287.
[3] C.A. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, P. Artursson, J. Med. Chem. 46 (2003) 558–570.
[4] S. Agatonovic-Kustrin, R. Beresford, A. Pausi, M. Yusof, J. Pharm. Biomed. Anal. 25 (2001) 227–237.
[5] E. Deconinck, T. Hancock, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal., 2005, in press.
[6] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Deliv. Rev. 46 (2001) 3–26.
[7] M.H. Abraham, A. Ibrahim, A.M. Zissimos, Y.H. Zhao, J. Corner, D.P. Reynolds, Drug Discov. Today 7 (2002) 1056–1063.
[8] Q.S. Xu, D.L. Massart, Y.Z. Liang, K.T. Fang, J. Chromatogr. A 998 (2003) 155–167.
[9] J.H. Friedman, Ann. Stat. 19 (1991) 1–141.
[10] V. Nguyen-Cong, G. Van Dang, B.M. Rode, Eur. J. Med. Chem. 31 (1996) 797–803.
[11] S. Ren, H. Kim, J. Chem. Inf. Comput. Sci. 43 (2003) 2106–2110.
[12] S. Ren, J. Chem. Inf. Comput. Sci. 43 (2003) 1679–1687.
[13] R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1055 (2004) 11–19.
[14] Q.S. Xu, Y.Z. Liang, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.
[15] S. Sekulic, B.R. Kowalski, J. Chemometrics 6 (1992) 199–216.
[16] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Monterey, 1984.
[17] 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.
[18] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-UCH, Weinheim, 2000.
[19] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon® Professional version, Software version 5.0, Milano Chemometrics and QSAR Research Group, Copyright Talete srl© (2004) 1997–2004.
[20] S.K. Poole, C.F. Poole, J. Chromatogr. B 797 (2003) 3–19.
[21] S. Yang, J.F. Bumgarner, L.F.R. Kruk, M.G. Khaledi, J. Chromatogr. A 721 (1996) 323–335.
[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.